МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ДНІПРОВСЬКА ПОЛІТЕХНІКА»

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
Кафедра інформаційних технологій та комп'ютерної інженерії

А.В. Кожевников

МЕТОДИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ (BIG DATA)

Методичні рекомендації до виконання лабораторних робіт
для здобувачів ступеня бакалавра
освітньо-професійної програми «Інформаційні системи та технології»
зі спеціальності 126 Інформаційні системи та технології

Дніпро
НТУ «ДП»
2025

Кожевников А.В.

Методи та інформаційні технології обробки великих даних (Big Data) [Електронний ресурс] : методичні рекомендації до виконання лабораторних робіт для здобувачів ступеня бакалавра освітньо-професійної програми «Інформаційні системи та технології» зі спеціальності 126 Інформаційні системи та технології / А. В. Кожевников ; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро : НТУ «ДП», 2025. – 89 с.

Автор:
А.В. Кожевников, канд. техн. наук, доц.

Затверджено науково-методичною комісією зі спеціальності 126 Інформаційні системи та технології (протокол № 1 від 22.01.2025 р.) за поданням кафедри інформаційних технологій та комп’ютерної інженерії (протокол № 9 від 22.01.2025 р.).

Уміщено теоретичні відомості за темами лекційного курсу, варіанти лабораторних завдань з рекомендаціями до їх виконання, контрольні питання, список використаної та рекомендованої літератури.

Зміст методичних рекомендацій відповідає програмі обов’язкової навчальної дисципліни «Методи та інформаційні технології обробки великих даних (Big Data)» і адресовано студентам, які проходять підготовку за спеціальністю 126 «Інформаційні системи та технології», а також здобувачам інших спеціальностей, які вивчають цю дисципліну.

Відповідальний за випуск завідувач кафедри інформаційних технологій та комп’ютерної інженерії В. В. Гнатушенко, д-р техн. наук, проф.

1. ЗАГАЛЬНІ ПОЛОЖЕННЯ

Дисципліна “Методи та інформаційні технології обробки великих даних (Big Data)” – складова фахової підготовки бакалаврів спеціальності 126 “Інформаційні системи та технології”, що навчаються за освітньою програмою “Інформаційні системи та технології”. Мета дисципліни – формування у здобувачів вищої освіти компетентностей щодо методів та інформаційних технологій обробки даних, розмір яких перевищує можливості звичайних програмних аналітичних платформ та баз даних по аналізу, зберіганню, і управлінню інформацією, або великих даних (Big Data).

Методичні рекомендації призначені для закріплення теоретичних знань, набутих студентами в лекційному курсі, а також формування практичних навичок виконання лабораторних робіт щодо методів, програмного забезпечення та інформаційних технологій обробки великих даних.

У підсумку виконання лабораторних робіт студенти – майбутні фахівці повинні отримати результати навчання у відповідності з вимогами освітньої програми, які представлені в табл. 1.1.

Таблиця 1.1 Результати навчання, що отримуються у підсумку виконання лабораторних робіт з дисципліни “Методи та інформаційні технології обробки великих даних (Big Data)”.

ПР6.1-Ф20	Демонструвати знання методів та інформаційних технологій обробки великих даних
ПР12.1-Ф20	Використовувати сучасні методи розробки програмного забезпечення для обробки великих даних
ПР13.1-Ф20	Застосовувати методи регресійного аналізу для рішення задач прогнозування
ПР13.2-Ф20	Використовувати методи кластеризації та класифікації для обробки великих даних
ПР14.1-Ф20	Розробляти та використовувати спеціалізоване програмне забезпечення для обробки і візуалізації просторових та багатовимірних даних

Мета лабораторних робіт – поглибити і систематизувати набуті студентами на лекціях теоретичні знання з дисципліни “Методи та інформаційні технології обробки великих даних (Big Data)” та сформувати у майбутніх фахівців з інформаційних систем та технологій професійних компетентностей (знань, умінь і навичок) використання методів, програмного забезпечення та інформаційних технологій обробки великих даних.

Лабораторні роботи передбачають застосування загальновідомих математичних методів, які можуть бути використані для обробки і аналізу даних. Кожна лабораторна робота має назву, ціль, постановку задачі, контрольні питання і завдання, рекомендації щодо виконання роботи та вимоги до оформлення звіту, а також варіанти індивідуальних завдань.

Одержання результатів робіт передбачено розрахунковими методами з використанням комп’ютерної техніки, яка працює на платформі Windows та програм середовищ математичних розрахунків (MathCAD, MATLAB), електронних таблиць EXCEL, середовища програмування Python на базі оболонки Anaconda. Доступ до сервісів платформи Hadoop здійснюється за допомогою віртуальних машин Oracle VM VirtualBox та HDP Sandbox.

2. ПОСТАНОВКА ЗАДАЧ ЛАБОРАТОРНИХ РОБІТ, МЕТОДИЧНІ РЕКОМЕНДАЦІЇ ДЛЯ ЇХ ВИКОНАННЯ ТА ІНДИВІДУАЛЬНІ ЗАВДАННЯ

2.1 Лабораторні роботи № 1, 2
Імовірнісні розподіли дискретних і безперервних випадкових величин та точкові оцінки їх параметрів

Об'єкт – дискретні і безперервні випадкові величини. Предмет – імовірнісні розподіли дискретних і безперервних випадкових величин та точкові оцінки їх параметрів. Мета – побудова імовірнісних розподілів, визначення точкових оцінок їх параметрів.

Стислі теоретичні відомості

Якщо випадковій події A з деякої множини подій можна поставити у відповідність деяке числове значення, то говорять що задано випадкову величину $X=X(A)$, яку можна розглядати як функціонал події.

Коли множина значень $\{x_i\}_{i=1,\dots,m}$, які може приймати випадкова величина, є зліченною (тобто значення можна пронумерувати натуральними числами), величина зветься дискретною, у противному – неперервною. Тут m – кінцева або нескінченна кількість нетотожних значень, які може приймати випадкова величина.

Залежності, що встановлюють зв'язок між можливими значеннями випадкової величини і ймовірностями їхніх появ, називаються законами розподілу випадкової величини. До способів задання законів розподілу відносяться функції диференціальних і кумулятивних (накопичених) розподілів. Стосовно неперервних величин вони також носять назви відповідно до щільностей імовірності й функцій розподілу ймовірностей випадкової величини.

Функцією розподілу випадкової величини X називається залежність

$$F(x) = P(X \le x), \tag{2.1.1}$$

де $P(X \le x)$ – імовірність події, яка складається в тому, що випадкова величина X є не більшою значення x.

Зворотна до неї функція $x = F^{-1}(P)$ зветься інверсним кумулятивним розподілом випадкової величини.

Імовірність влучення неперервної випадкової величини X в проміжок $(a, b]$ визначається співвідношенням

$$P(a < X \le b) = F(b) - F(a).. \tag{2.1.2}$$

Диференціальний розподіл для неперервної випадкової величини $p(x)$ можна визначити як

$$p(x) = \frac{dF(x)}{dx}. \qquad (2.1.3)$$

Відповідно вираз для кумулятивного розподілу через диференціальний подається у вигляді

$$F(x) = \int_{-\infty}^{x} p(x)dx. \qquad (2.1.4)$$

Диференціальні й кумулятивні розподіли неперервної дискретної випадкової величини прийнято подавати за допомогою звичайних графіків.

Диференціальний розподіл дискретної випадкові величини характеризують множиною ймовірностей $\{p_i\}_{i=1,\dots,m}$ прийняття випадковою величиною X відповідно значень $\{x_i\}_{i=1,\dots,m}$.

Для дискретної випадкової величини кумулятивну функцію розподілу можна визначити як

$$F_i = \sum_{x_k \le x_i} p_k. \qquad (2.1.5)$$

Графічні представлення диференціального й кумулятивного розподілів дискретної випадкової величини носять назви відповідно полігона й накопиченого полігона ймовірностей.

Розглянуті вище закони розподілу випадкових величин, які визначаються на основі аналізу загальних закономірностей виникнення описуваних ними випадкових подій, носять назви теоретичних розподілів. Закони розподілу випадкових величин, які визначаються на основі досліду або експериментальним шляхом, носять назви емпіричних розподілів.

При проведенні досліду, метою якого є побудова емпіричного розподілу, із множини однорідних об’єктів, кожний з яких характеризується випадковою величиною (або сукупністю випадкових величин), здійснюється випадковий відбір об’єктів. Іншим варіантом досліду є проведення випробувань із випадковими наслідками, що характеризуються випадковою величиною. Уся множина об’єктів, з якої робиться вибірка, називається генеральною сукупністю, множина відібраних об’єктів – вибірковою сукупністю або вибіркою, а їхня кількість – об’ємом вибірки n.

Числа об’єктів $\{n_i\}_{i=1,\dots,N}$, для яких у вибірці дискретна випадкова величина X приймає відповідно значення $\{n_i\}_{i=1,\dots,m}$, називають частотами появи цих значень, а величини $w_i = \frac{n_i}{n}$ – відносними частотами. Відносні частоти $\{w_i\}_{i=1,\dots,N}$ є наближеннями відповідних імовірностей $\{p_i\}_{i=1,\dots,m}$, причому $\lim\limits_{n\to\infty} w_i = p_i$. Для дискретної випадкової величини полігон і накопичений полігон

відносних частот є емпіричними аналогами відповідно полігона й накопиченого полігона теоретичних розподілів.

Нехай у вибірці неперервна випадкова величина X приймає значення, що лежать на сегменті $[a, b]$. Зазначений діапазон розбитий на m однакових класових інтервалів величиною $\Delta = \frac{b-a}{m}$, а відносні частоти влучення в інтервали $\{\Delta x_i\}_{i=1,\dots,m}$, де $\Delta x_i = (x_i, x_i + \Delta]$, а $x_i = a + (i-1)\Delta$, становлять $\{\Delta w_i\}_{i=1,\dots,m}$. Стовпчасті діаграми з основами стовпців $\{\Delta x_i\}_{i=1,\dots,m}$ та їхніми висотами $\left\{\frac{\Delta w_i}{\Delta}\right\}_{i=1,\dots,m}$ і $\left\{\sum_{j=1}^{i} \Delta w_j\right\}_{i=1,\dots,m}$ носять відповідно назви гістограми й накопиченої гістограми розподілу випадкової величини. Ці гістограми є емпіричними аналогами графіків диференціального і кумулятивного теоретичного розподілів неперервної випадкової величини.

Визначені на основі безпосередньо даних вибірки або емпіричних розподілів параметри називаються статистичними параметрами або оцінками параметрів

Постановка задачі

Дано

Закони розподілу дискретної та безперервної випадкових величин та їх параметри.

Потрібно

Побудувати теоретичні та емпіричні функції ймовірнісних розподілів.

Визначити значення основних параметрів розподілу, а також їх емпіричні оцінки.

Рекомендації щодо виконання роботи

1. **Лабораторна робота 1.** Відповідно до номеру варіанта та згідно даних табл. 2.1.1 для заданого закону розподілу дискретної випадкової величини, використовуючи середовище математичних розрахунків, яке має статистичні функції (MathCAD, MATLAB, Excel тощо):

1.1. Визначити межі зміни випадкової величини зі співвідношень $i_{min} = F^{-1}(0.001)$, $i_{max} = F^{-1}(0.999)$.

1.2. Згенерувати вибірку об’ємом 10^3 реалізацій випадкової величини на сегменті $[i_{min}, i_{max}]$ й визначити кількість реалізацій, у яких випадкова величина приймає кожне ціле значення з цього сегмента.

1.3. Побудувати в єдиному полі полігони теоретичного диференціального розподілу випадкової величини з використанням заданої формули функції розподілу й убудованої функції середовища математичних розрахунків, а також полігон емпіричного розподілу, отриманого на основі вибірки.

1.4. Побудувати в єдиному полі накопичений полігон теоретичного кумулятивного розподілу випадкової величини з використанням убудованої функції середовища математичних розрахунків, а також накопичений полігон емпіричного розподілу, отриманого на основі вибірки.

1.5. Визначити значення параметрів: математичного очікування, дисперсії, медіани та асиметрії випадкової величини з використанням убудованих функцій середовища математичних розрахунків.

1.6. Визначити емпіричні оцінки параметрів: середнього значення та статистичної дисперсії – за даними вибірки та за емпіричним розподілом; медіани, асиметрії та ексцесу – тільки за даними вибірки об'ємом $n=10^3$.

1.7. Порівняти значення математичного очікування і дисперсії випадкової величини, а також їх емпіричних оцінок – середнього значення та статистичної дисперсії для вибірок об'ємом n: 50, 10^2, $5\cdot10^2$, 10^3, $5\cdot10^3$. Результати представити у вигляді графіка залежності оцінки від об'єму вибірки з логарифмічним масштабом по осі аргументу.

2. **Лабораторна робота 2.** Відповідно до номеру варіанта та згідно даних табл. 2.1.2 для заданого закону розподілу неперервної випадкової величини, використовуючи середовище математичних розрахунків, яке має статистичні функції (MathCAD, MATLAB, Excel тощо):

2.1. Визначити межі зміни випадкової величини зі співвідношень $x_{min}= F^{-1}(0.001)$, $x_{max}= F^{-1}(0.999)$.

2.2. Визначити межі двадцяти однакових класових інтервалів, що покривають сегмент $[x_{min}, x_{max}]$. Згенерувати вибірку об'ємом 10^3 реалізацій випадкової величини й визначити кількість реалізацій, що попадають у кожний класовий інтервал.

2.3. Побудувати в єдиному полі графіки теоретичного диференціального розподілу випадкової величини з використанням заданої формули функції розподілу й убудованої функції середовища математичних розрахунків, а також гістограму емпіричного розподілу, отриманого на основі вибірки.

2.4. Побудувати в єдиному полі графік теоретичного кумулятивного розподілу випадкової величини з використанням убудованої функції середовища математичних розрахунків, а також накопичену гістограму емпіричного розподілу, отриманого на основі вибірки.

2.5. Визначити значення параметрів: математичного очікування, дисперсії, медіани та асиметрії випадкової величини з використанням убудованих функцій середовища математичних розрахунків.

2.6. Визначити емпіричні оцінки параметрів: середнього значення та статистичної дисперсії – за даними вибірки та за емпіричним розподілом; медіани, асиметрії та ексцесу – тільки за даними вибірки об'ємом $n=10^3$.

2.7. Порівняти значення математичного очікування і дисперсії випадкової величини, а також їх емпіричних оцінок – середнього значення та статистичної дисперсії для вибірок об'ємом n: 50, 10^2, $5\cdot10^2$, 10^3, $5\cdot10^3$. Результати представити у вигляді графіка залежності оцінки від об'єму вибірки з логарифмічним масштабом по осі аргументу.

Вимоги до звіту

Звіт роботі повинен містити:

1. Назву дисципліни та лабораторної роботи.
2. Прізвище, ім'я та по батькові студента, шифр групи.
3. Об'єкт, предмет і мету лабораторної роботи.

4. Аналітичний вираз для диференціального розподілу випадкової величини.

5. Код програми, що реалізує поставлені завдання.

6. Полігони, графіки, гістограми теоретичних і емпіричних диференціальних розподілів випадкової величини.

7. Накопичені полігони, графіки, гістограми теоретичних і емпіричних кумулятивних розподілів випадкової величини.

8. Результати розрахунків значень параметрів випадкової величини та їх точкових оцінок.

9. Графіки залежності середнього значення та статистичної дисперсії випадкової величини від об'єму вибірки.

10 Висновки.

Контрольні питання і завдання

1. Дайте визначення і сформулюйте основні властивості ймовірності випадкової події.

2. Дайте визначення випадкових величин. Наведіть приклади дискретних та безперервних випадкових величин.

3. Наведіть співвідношення між диференціальними і кумулятивними теоретичними розподілами дискретних і неперервних випадкових величин.

4. Сформулюйте основні властивості теоретичних розподілів дискретних і неперервних випадкових величин.

5. Дайте визначення інверсного кумулятивного розподілу випадкової величини.

6. Назвіть найпоширеніші закони розподілу дискретних і неперервних випадкових величин.

7. Поясніть різницю між теоретичними й емпіричними розподілами випадкових величин.

8. Наведіть співвідношення між диференціальними і кумулятивними емпіричними розподілами дискретних і неперервних випадкових величин.

9. Побудуйте полігон і накопичений полігони для теоретичних розподілів випадкової величини, що характеризує результат кидання монети, а також їхні емпіричні аналоги для п'яти випадкових реалізацій.

10 Неперервна випадкова величина може приймати значення, що лежать на сегменті [0, 4]. Внаслідок досліду отримані значення величини {1,15, 2,38, 1,89, 3,33, 0,06}. Побудуйте гістограми й накопичені гістограми для емпіричних розподілів випадкової величини при кількості m класових інтервалів 2 та 4.

Таблиця 2.1.1. Характеристики і параметри дискретних імовірнісних розподілів

№ варіанта	Назва розподілу і його параметри [1]	Формула диференціального розподілу [2]	Формула математичного очікування	Формула дисперсії	Покажчики функцій розподілу середовища MathCAD [3]
1, 6, 11, 16, 21, 26, 31	Рівномірний $i_{\min} < i_{\max} \in \mathbf{N}$; $i_{\min}$ =№ mod 3; $i_{\max}$ =№ mod 5 +10	$\frac{1}{i_{\max} - i_{\min} + 1}$; $i_{\min} \le i \le i_{\max}$	$\frac{i_{\max} + i_{\min}}{2}$	$\frac{(i_{\max} - i_{\min} + 1)^2 - 1}{12}$	$dunif(i, i_{\min} - 0.5, i_{\max} + 0.5)$ $punif(i, i_{\min} - 0.5, i_{\max} + 0.5)$ $qunif(P, i_{\min} - 0.5, i_{\max} + 0.5)$ $round(runif(K, i_{\min} - 0.5, i_{\max} + 0.5))$
2, 7, 12, 17, 22, 27, 32	Пуассона $\lambda \in \mathbf{R}$; λ =№ mod 10 +1	$\frac{\lambda^i}{i!}\exp(-\lambda)$; $i \ge 0$	λ	λ	$dpois(i, \lambda)$ $ppois(i, \lambda)$ $qpois(P, \lambda)$ $rpois(K, \lambda)$
3, 8, 13, 18, 23, 28	Біноміальний $n \in \mathbf{N}$; n =№ + 5; $p \in \mathbf{R}$; $0 < p < 1$; $p = 1/$№	$C_n^i p^i (1-p)^{n-i}$; $i \ge 0$	np	$np(1-p)$	$dbinom(i, n, p)$ $pbinom(i, n, p)$ $qbinom(P, n, p)$ $rbinom(K, n, p)$
4, 9, 14, 19, 24, 29	Геометричний $p \in \mathbf{R}$; $0 < p < 1$; $p = 2/$№	$p(1-p)^i$; $i > 0$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$dgeom(i, p)$ $pgeom(i, p)$ $qgeom(P, p)$ $rgeom(K, p)$
5, 10, 15, 20, 25, 30	Гіпергеометричний $n,\ a,\ b \in \mathbf{N}$; $0 \le n \le a+b$; a=№ div 2; b=№ div 3; n = (a+b) div 2	$\frac{C_a^i C_b^{n-i}}{C_{a+b}^n}$ $\max(0, n-b) \le i \le \min(n, a)$ $\max(0, n-a) \le n-i \le \min(n, b)$	$\frac{na}{a+b}$	$\frac{nab(a+b-n)}{(a+b-1)(a+b)^2}$	$dhypergeom(i, a, b, n)$ $phypergeom(i, a, b, n)$ $qhypergeom(P, a, b, n)$ $rhypergeom(K, a, b, n)$

[1] Тут і далі при визначенні параметрів варіанта завдання з номером № використовуються операції цілочисельного ділення, що позначаються символами div та mod і визначаються як: a div b = $[a/b]$; a mod b = a – $[a/b]\cdot b$, де квадратні дужки символізують функцію виділення цілої частини, [2] За межами зазначеної області визначення щільність імовірності покладається рівної нулю, [3] Функціям заданого типу відповідають префікси: *d* – диференціального розподілу, *p* –

кумулятивного розподілу, q – інверсного кумулятивного розподілу, r – генератора випадкових чисел із заданим розподілом.

Таблиця 2.1.2. Характеристики і параметри неперервних імовірнісних розподілів

№ варіанта	Назва розподілу і його параметри [1)]	Формула диференціального розподілу [2)]	Формула математичного очікування	Формула дисперсії	Покажчики функцій розподілів середовища MathCAD [3)]
1, 8, 15, 22, 29	Нормальний $\mu,\sigma\in\mathbf{R},\ \sigma>0;$ μ=№ mod 3; σ=№ mod 3+1	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2	$dnorm(x,\mu,\sigma)$ $pnorm(x,\mu,\sigma)$ $qnorm(P,\mu,\sigma)$ $rnorm(K,\mu,\sigma)$
2, 9, 16, 23, 30	Пірсона (χ^2) $n\in\mathbf{N};\ n$=№	$\frac{\left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{2\Gamma\left(\frac{n}{2}\right)}\exp\left(-\frac{x}{2}\right);\ x\ge 0$	n	$2n$	$dchisq(x,n)$ $pchisq(x,n)$ $qchisq(P,n)$ $rchisq(K,n)$
3, 10, 17, 24, 31	Гамма $s\in\mathbf{R};\ s>0;\ s$=№	$\frac{x^{s-1}}{\Gamma(s)}\exp(-x);\ x\ge 0$	s	s	$dgamma(x,s)$ $pgamma(x,s)$ $qgamma(P,s)$ $rgamma(K,s)$
4, 11, 18, 25, 32	Вейбулла $s\in\mathbf{R};\ s>0;\ s$=1+1/№	$sx^{s-1}\exp\left(-x^s\right);\ x\ge 0$	$\Gamma\left(1+\frac{1}{s}\right)$	$\Gamma\left(1+\frac{2}{s}\right)-$ $-\Gamma^2\left(1+\frac{1}{s}\right)$	$dweibull(x,s)$ $pweibull(x,s)$ $qweibull(P,s)$ $rweibull(K,s)$
5, 12, 19, 26	Експонентний $s\in\mathbf{R};\ s>0;\ s$=№/10	$s\exp(-sx);\ x\ge 0$	$\frac{1}{s}$	$\frac{1}{s^2}$	$dexp(x,s)$ $pexp(x,s)$ $qexp(P,s)$ $rexp(K,s)$

6, 13, 20, 27	Рівномірний $x_{\min} < x_{\max} \in \mathbf{R}$; $x_{\min}$ =№ mod 3; $x_{\max}$ =№ mod 5 +10	$\frac{1}{x_{\max} - x_{\min}}$; $x_{\min} \le x \le x_{\max}$	$\frac{i_{\max} + i_{\min}}{2}$	$\frac{(i_{\max} - i_{\min})^2}{12}$	*dunif* $(x, x_{\min}, x_{\max})$ *punif* $(x, x_{\min}, x_{\max})$ *qunif* $(P, x_{\min}, x_{\max})$ *runif* $(K, x_{\min}, x_{\max})$
7, 14, 21, 28	Логарифмічно нормальний $\mu, \sigma \in \mathbf{R}$, $\sigma > 0$; μ =№ mod 2; σ =1	$\frac{1}{\sqrt{2\pi}x\sigma}\exp\left\{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}\right\}$; $x \ge 0$	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$	$\left(\exp\left(\sigma^2\right)-1\right)$ $\times\exp\left(2\mu+\sigma^2\right)$	*dlnorm*(x, μ, σ) *plnorm*(x, μ, σ) *qlnorm*(P, μ, σ) *rlnorm*(K, μ, σ)

[1] Тут $C_n^m = \frac{n!}{m!(n-m)!}$ – число сполучень із *n* по *m* елементів без повторення; $\Gamma(x) = \int_{-\infty}^{+\infty} t^{x-1}\exp(-t)dt$ – гамма-функція;

1. Покажчики відповідних функцій середовища MathCAD мають вигляд: *combin*(*n*,*m*); $\Gamma(x)$.

2.2 Лабораторна робота № 3
Спеціальні функції математичної статистики, інтервальні оцінки параметрів випадкових величин та тести перевірки статистичних гіпотез

Об'єкт – дискретні і безперервні випадкові величини. Предмет – інтервальні оцінки параметрів випадкових величин, тести перевірки статистичних гіпотез. Мета – визначення інтервальних оцінок параметрів випадкових величин та перевірка відповідності емпіричного ймовірнісного розподілу теоретичному.

Стислі теоретичні відомості

Найпоширеніші функції математичної статистики, які використовуються для перевірки статистичних гіпотез та їх покажчики в середовищі MathCAD наведені в табл.1.

Таблиця 2.2.1. Основні інверсні кумулятивні функції математичної статистики

Назва розподілу	Символічна позначка функції	Покажчик функції в середовищі MathCAD
Нормальний з параметрами (0,1)	$\Phi_0^{-1}\left(\frac{\beta}{2}\right)$	$qnorm\left(\frac{1}{2}+\frac{\beta}{2},0,1\right)$
	$\Phi^{-1}(\beta)$	$\frac{1}{\sqrt{2}}\cdot qnorm\left(\frac{1+\beta}{2},0,1\right)$
Стьюдента з n ступенями волі	$\Phi S^{-1}(\beta,n)$	$qt\left(\frac{1+\beta}{2},n\right)$
	$\Phi S^{-1}(\alpha,n)$	$qt\left(\frac{1-\alpha}{2}+0.5,n\right)$
Фішера з n_1, n_2 ступенями волі	$\Phi F^{-1}(\alpha,n_1,n_2)$	$qF\left(1-\alpha,n_1,n_2\right)$
χ^2 з q ступенями волі	$\Phi C^{-1}(\alpha,q)$	$qchisq\left(1-\alpha,q\right)$
	$\Phi C^{-1}(\beta,q)$	$qchisq\left(\beta,q\right)$

Тут α – рівень значимості (ймовірність помилки I-го роду), $\beta=1-\alpha$ – довірча ймовірність. Слід мати на увазі, що областю визначення функції $\Phi_0^{-1}\left(\frac{\beta}{2}\right)$ є $[0;0,5]$.

Серед оцінок параметрів розрізняють точкові оцінки, які характеризують власно значення параметра та інтервальні, які характеризують величину похибки та надійність точкових оцінок. До інтервальних оцінок відносяться довірчий інтервал та довірча ймовірність. Довірчий інтервал характеризує окіл точкової оцінки параметру ймовірність потрапляння у яку істинного значення параметру дорівнює довірчій ймовірності.

Розглянемо побудування довірчого інтервалу для математичного очікування. Центральна гранична теорема стверджує, що коли випадкова величина представляє собою суму великої кількості незалежних доданків, при чому внесок кожного з доданків у суму невеликий, то ймовірнісний розподіл значення суми наближається до нормального. Якщо ж доданки також мають однаковий довільний ймовірнісний розподіл з математичним очікуванням a та дисперсією σ^2, то ймовірнісний розподіл їх середнього значення наближається до нормального з параметрами $\left(a, \frac{\sigma^2}{n}\right)$. Довірча ймовірність β представляє собою ймовірність потрапляння математичного очікування в ε окіл його точкової оцінки $\tilde{a}$ – середнього значення становить

$$\beta = P(|\tilde{a} - a| < \varepsilon) = 2\Phi_0\left(\sqrt{n}\frac{\varepsilon}{\sigma}\right) = \Phi\left(\sqrt{\frac{n}{2}}\frac{\varepsilon}{\sigma}\right). \quad (2.2.1)$$

Тоді довірчий інтервал для математичного очікування може бути визначений як $(\tilde{a} - \varepsilon, \tilde{a} + \varepsilon)$ де його величина дорівнює

$$\varepsilon = \sqrt{\frac{2}{n}}\sigma\Phi^{-1}(\beta). \quad (2.2.2)$$

Формула (2.2.2) є справедливою, якщо відома дисперсія генеральної сукупності σ^2 для якої на основі вибірки визначається оцінка математичного очікування або об'єм цієї вибірки великий. В інших випадках коли оцінка дисперсії здійснюється за вибіркою невеликого об'єму необхідно використовувати співвідношення

$$\varepsilon = \frac{\tilde{\sigma}}{\sqrt{n}}\Phi\,\mathrm{S}^{-1}(\beta). \quad (2.2.3)$$

Довірчий інтервал для дисперсії нормально розподіленої випадкової величини, який будується за даними вибірки зі статистичною дисперсією $\tilde{\sigma}^2$, визначається як

$$\left(\frac{n\tilde{\sigma}^2}{\Phi\,\mathrm{C}^{-1}\left(\frac{1+\beta}{2}\right)}, \frac{n\tilde{\sigma}^2}{\Phi\,\mathrm{C}^{-1}\left(\frac{1-\beta}{2}\right)}\right). \quad (2.2.4)$$

Методи випадкових іспитів (статистичних іспитів, методи Монте-Карло) чисельного моделювання відтворюваних випадкових величин, подій або

процесів полягають в використанні випадкових (псевдовипадкових) чисел для багатократного відтворення (розіграшу) випадкових реалізацій відповідних величин, подій або процесів з урахуванням їх імовірнісних характеристик. Результатами моделювання, як правило, виступають або середнє за множиною реалізацій значення досліджуваної випадкової величини, або оцінка імовірності виникнення випадкової події. Остання визначається як відносна частота виникнення випадкової події, або відношення кількості реалізацій, в яких трапилась ця подія, до загальної кількості реалізацій. Так, наприклад, якщо $\{A_i\}_{i=1,\dots,N}$ – значення які приймала випадкова величина в i – й реалізації при їх загальній кількості N,оцінка ймовірності її потрапляння в інтервал значень $[a,b]$ в середовищі математичних розрахунків MathCAD може бути обчислена як

$$P := \frac{1}{N} \cdot \sum_{i=0}^{N-1} \text{if}(a \le A_i \wedge A_i \le b, 1, 0). \qquad (2.2.5)$$

Тести перевірки статистичних гіпотез дозволяють на основі даних вибірки (або вибірок) перевірити певні припущення (гіпотези) щодо властивостей генеральних сукупностей з яких отримані ці вибірки.

Процедура перевірки статистичних гіпотез за звичай передбачає наступні кроки:

- формулювання нульової та альтернативної гіпотез;
- обчислення на основі даних вибірки (або вибірок) контрольної величини тесту;
- задання рівня значимості тесту α (ймовірності помилки I-го роду, яка полягає в відхиленні правильної нульової гіпотези) і обчисленні на його основі критичного значення тесту;
- прийнятті або відхиленні нульової гіпотези на користь альтернативної шляхом порівняння контрольного значення тесту з критичним.

Критерій χ^2 є найбільш застосовуваним при перевірках відповідності емпіричних імовірнісних розподілів теоретичним. Нульова гіпотеза критерію формулюється як відповідність емпіричного ймовірнісного розподілу теоретичному, а альтернативна – як невідповідність.

Контрольна величина критерія визначається як

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)}{np_i}, \qquad (2.2.6)$$

де n – об’єм вибірки, m – кількість значень (класів) які може приймати дискретна випадкова величина, або кількість класових інтервалів в які може потрапляти неперервна випадкова величина, $\{n_i\}_{i=1,\dots,m}$, $\{p_i\}_{i=1,\dots,m}$ – відповідно частоти та ймовірності потрапляння величини в i-й клас (класовий інтервал).

Останні або безпосередньо визначають диференціальний теоретичний розподіл для дискретної величини, або для безперервної можуть бути обчислені виходячи з функції щільності імовірності $p(x)$

$$p_i = \int_{x_i}^{x_i+\Delta} p(x)dx, \tag{2.2.7}$$

де $\{x_i\}_{i=1,\ldots,m}$ – координати лівих меж класових інтервалів, Δ – величини класових інтервалів.

Критичне значення тесту розраховується за допомогою інверсної кумулятивної функції χ^2 з q ступенями волі

$$\chi^2{}_{кр}(\alpha) = \Phi\mathrm{C}^{-1}(\alpha, q). \tag{2.2.8}$$

Тут $q = m - r - 1$, де r – число параметрів теоретичного розподілу, які апріорно невідомі, а обраховуються за даними вибірки. Якщо $r = 0$ гіпотеза зветься простою, інакше складною.

При виконанні умови

$$\chi^2 \geq \chi^2{}_{кр}(\alpha) \tag{2.2.9}$$

нульова гіпотеза про відповідність емпіричного ймовірнісного розподілу теоретичному відхиляється, інакше вважається що вона не протирічить даним вибірки.

Постановка задачі

Дано

Закони розподілу дискретної та безперервної випадкових величин та їх параметри.

Потрібно

Визначити інтервальні оцінки математичного очікування та дисперсії випадкової величини.

Перевірити відповідність емпіричного розподілу середніх значень однаково розподілених вибірок нормальному розподілу.

Рекомендації щодо виконання роботи

Відповідно до номеру варіанта та згідно даних лабораторної роботи 1 або 2 за вказівкою викладача для заданого закону розподілу випадкової величини, використовуючи середовище математичних розрахунків, яке має статистичні функції (MathCAD, MATLAB, Excel тощо):

1. Визначити функції залежностей величини довірчого інтервалу математичного очікування від довірчої ймовірності з використанням інверсних кумулятивних функцій нормального розподілу та розподілу Стьюдента.

Побудувати сімейства графіків цих функцій для вибірок з одиничним значенням дисперсії та об'ємами n: 2, 5, 10.

2. Згенерувати матрицю A значень випадкової величини, який містить $N = 10^3$ вибірок об'єму $n = 100$ для ймовірнісного розподілу, який розглядався в лабораторній роботі 1.

3. Методом випадкових іспитів визначити оцінки довірчої ймовірності потрапляння математичного очікування випадкової величини в довірчі інтервали для величини довірчої ймовірності $\beta = 0{,}9 + 0{,}01 \cdot (№ \bmod 10)$, де № – номер варіанта. При визначенні величини довірчих інтервалів використовувати: для інтервалу на основі нормального розподілу – СКВ (дисперсію) генеральної сукупності (теоретичного розподілу випадкової величини), а для інтервалу на основі розподілу Стьюдента – оцінку СКВ (статистичну дисперсію) отриману за даними вибірки.

4. Повторити п. 2, 3 для ймовірнісного розподілу, який розглядався в лабораторній роботі 2. Подальші пункти роботи виконувати тільки для цього ймовірнісного розподілу.

5. Сформувати вибірку B – вектор середніх значень вибірок, які входять до вибірки A. Згідно центральної граничної теореми, вибірка B повинна підпорядковуватися нормальному закону розподілу незалежно від закону розподілу вибірки A.

6. Визначити середнє значення та статистичну дисперсію вибірки B. Перевірити їх відповідність математичному очікуванню та дисперсії теоретичного розподілу вибірки A.

7. Побудувати в спільному графічному полі гістограму диференціального емпіричного розподілу на основі вибірки B та відповідний до нього графік нормального розподілу з параметрами за п.6. Мінімальне та максимальне значення випадкової величини для емпіричного розподілу обчислити за правилом "трьох сигм", а кількість класових інтервалів розподілу – за формулою Стержесса.

8. Перевірити за допомогою критерія χ^2 відповідність отриманого емпіричного розподілу нормальному з параметрами за п.6. Рівень значимості прийняти $\alpha = 0{,}01 + 0{,}01 \cdot (№ \bmod 9)$.

Вимоги до звіту

Звіт роботі повинен містити:

1. Назву дисципліни та лабораторної роботи.
2. Прізвище, ім'я та по батькові студента, шифр групи, номер варіанта.
3. Об'єкт, предмет і мету лабораторної роботи.
4. Код програми, що реалізує поставлені завдання.
5. Сімейства графіків залежності величини довірчого інтервалу математичного очікування від довірчої ймовірності з використанням інверсних кумулятивних функцій нормального розподілу та розподілу Стьюдента.
6. Оцінки довірчої ймовірності потрапляння математичного очікування випадкової величини в довірчі інтервали, що побудовані з використанням

інверсних кумулятивних функцій нормального розподілу та розподілу Стьюдента.

Оцінку довірчої ймовірності потрапляння дисперсії випадкової величини в довірчий інтервал.

Гістограму диференціального емпіричного розподілу середніх значень однаково розподілених вибірок та відповідний до нього нормальний розподіл.

Результати перевірки за допомогою критерія χ^2 відповідності емпіричного розподілу значень однаково розподілених вибірок нормальному.

Контрольні питання і завдання

1. Назвіть основні точкові та інтервальні оцінки випадкових величин та поясніть, що вони характеризують.

2. Дайте визначення довірчої ймовірності.

3. Опишіть загальну процедуру перевірки статистичних гіпотез.

4. Дайте визначення помилок І-го та ІІ-го роду при перевірці статистичних гіпотез.

5. Поясніть різницю між одно- та двосторонніми гіпотезами.

6. Поясніть різницю між параметричними та непараметричними тестами.

7. Назвіть найпоширеніші функції математичної статистики, які використовуються для перевірки статистичних гіпотез.

8. Назвіть основні критерії перевірки статистичних гіпотез.

9. Сформулюйте нульові гіпотези критеріїв Стьюдента та Фішера.

10 Поясніть різницю між простими та складними гіпотезами перевірки відповідності емпіричних розподілів теоретичним. Дайте визначення кількості ступенів волі гіпотези.

11. Поясніть яким чином за даними вибірки визначається можливий закон розподілу генеральної сукупності, якщо він апріорно невідомий.

2.3 Лабораторна робота № 4
Кореляційний аналіз

Об’єкт – сукупності випадкових величини. Предмет – коефіцієнти кореляції випадкових величин. Мета – визначення показників кореляційного зв’язку випадкових величин та перевірка їх значимості.

Стислі теоретичні відомості

Розглянемо вибірку об’єму n, яка представлена варіантами, що характеризуються m кількісними ознаками $\left\{\xi_{ij}\right\}_{\substack{i=1,\dots,n \\ j=1,\dots,m}}$. Під вибірковим парним коефіцієнтом кореляції Пірсона ознак j та k мається на увазі величина

$$r_{jk}=\frac{\sum_{i=1}^{n}\left(\xi_{ij}-\overline{\xi}_j\right)\cdot\left(\xi_{ik}-\overline{\xi}_k\right)}{\sqrt{\sum_{i=1}^{n}\left(\xi_{ij}-\overline{\xi}_j\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(\xi_{ik}-\overline{\xi}_k\right)^2}}=\frac{\tilde{C}_{jk}}{\sqrt{\tilde{D}_j\cdot\tilde{D}_k}}.,\qquad(2.3.1)$$

Тут $\overline{\xi}_j$, $\overline{\xi}_k$ – середні вибіркові значення ознак j та k відповідно,

$\tilde{D}_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_{ij} - \overline{\xi}_j\right)^2$ та $\tilde{D}_k = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_{ik} - \overline{\xi}_k\right)^2$ – їх статистичні дисперсії, а

$\tilde{C}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_{ij} - \overline{\xi}_j\right)\cdot\left(\xi_{ik} - \overline{\xi}_k\right)$ – їх статистична коваріація. Парні коефіцієнти кореляції можуть приймати значення в діапазоні $[-1;1]$.

Взаємний зв'язок між всілякими парами ознак характеризується матрицею парних коефіцієнтів кореляції $R = \left\{r_{jk}\right\}_{\substack{j=1,\ldots,m \\ k=1,\ldots,m}}$ розміром $m \cdot m$, яка має наступні властивості:

1. Елементи головної діагоналі є одиничними $r_{jj} = 1$;

2. Матриця є симетричною $r_{jk} = r_{kj}$ або $R = R^T$.

Зв'язок певної ознаки i зо всіма іншими визначається за допомогою множинного коефіцієнту кореляції, котрий визначається як

$$r_i = \sqrt{1 - \frac{|R|}{|R_{ii}|}}, \tag{2.3.2}$$

де $|R|, |R_{ii}|$ – відповідно визначники матриці парних коефіцієнтів кореляції та її підматриці, що утворюється внаслідок викреслювання i-х рядка та стовпчика.

Множинні коефіцієнти кореляції можуть приймати значення в діапазоні $[0;1]$.

При перевірці значимості (відмінності від нуля) парного коефіцієнту кореляції r у якості контрольного використовується значення

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}. \tag{2.3.3}$$

Критичне значення тесту розраховується за допомогою інверсної кумулятивної функції Стьюдента з $n-2$ ступенями волі

$$t_{кр}(\alpha) = \Phi S^{-1}(\alpha, n-2). \tag{2.3.4}$$

При виконанні умови

$$|t| > t_{кр}(\alpha) \tag{2.3.5}$$

нульова гіпотеза про рівність нулю коефіцієнта кореляції відхиляється, інакше вважається що вона не протирічить даним вибірки.

При перевірці значимості множинного коефіцієнту кореляції r у якості контрольного використовується значення

$$F = \frac{r^2}{1-r^2} \cdot \frac{n-m}{m-1}. \tag{2.3.6}$$

Критичне значення тесту розраховується за допомогою інверсної кумулятивної функції Фішера з $(m-1, n-m)$ ступенями волі

$$F_{кр}(\alpha) = \Phi \mathrm{F}^{-1}(\alpha, m-1, n-m). \tag{2.3.7}$$

При виконанні умови

$$F > F_{кр}(\alpha) \tag{2.3.8}$$

нульова гіпотеза про рівність нулю коефіцієнта кореляції відхиляється, інакше вважається що вона не протирічить даним вибірки.

Як показник кореляційного зв'язку ознак, які задані у порядковій шкалі, зокрема використовується коефіцієнт рангової кореляції Спірмена, котрий визначається як

$$\rho = \frac{\frac{n^3-n}{6} - \sum_{i=1}^{n}(r_i - s_i)^2 - B_r - B_s}{\sqrt{\frac{n^3-n}{6} - 2B_r} \cdot \sqrt{\frac{n^3-n}{6} - 2B_s}}, \tag{2.3.9}$$

де n – об'єм вибірки, $\{r_i, s_i\}_{i=1,\dots,n}$ – ранги ознак r та s для кожної варіанти вибірки, $B_r = \frac{1}{12}\sum_{j=1}^{p} K_j\left(K_j^2 - 1\right)$ – поправка для ознаки r на зв'язані ранги (які мають однакове середнє значення), p – кількість груп зв'язаних рангів для ознаки r, $\{K_j\}_{j=1,\dots,p}$ – кількість зв'язаних рангів у відповідній групі. Коефіцієнти рангової кореляції Спірмена можуть приймати значення в діапазоні $[-1;1]$.

Постановка задачі

Дано

Багатовимірна вибірка, ознаки якої надаються у порядковій шкалі.

Потрібно

Сформувати: з вибірки даних, що відображають результати діагностики апендициту, підвибірку згідно номеру варіанту.

Визначити: матриці парних коефіцієнтів кореляції Пірсона та Спірмена та вектор множинних коефіцієнтів кореляції для підвибірок даних в звичайній порядковій та ранговій шкалах.

Перевірити: значимість мінімальних за абсолютними значеннями парного та множинного коефіцієнтів кореляції.

Рекомендації щодо виконання роботи

В табл.Д.1.1, що міститься у додатку 1 та у файлі *data.txt* представлені варіанти, що відображають результати діагностики апендициту. Кожна варіанта (рядок) відповідає одному хворому. В першому стовпчику вказаний точний діагноз, який відображає ступінь тяжкості захворювання, (1 – непідтверджений діагноз, 2 – катаральний апендицит, 3 – флегмозний апендицит, 4 – гангренозний апендицит), в стовпчиках з другого по дев'ятий – значення симптомів x1 – x8, які виражені в порядкових шкалах:
x1 – болі в правій підвздошній області 1 – незначні, 2 – виражені;
x2 – тривалість болей 1 – до 12 годин, 2 – 13 – 24 годин, 3 – 25 – 48 годин, 4 – понад 2 діб;
x3 – частота пульсу 1 – до 80 уд/хв, 2 – 80 – 100 уд/хв, 3 – понад 100 уд/хв;
x4 – лейкоцити крові 1 – до 8 тис., 2 – 8–14 тис., 3 – понад 14 тис.;
x5 – зміни язика 1 – не обкладений, 2 – обкладений,;
x6 – симптом Щоткіна – Блюмберга 1 – відсутній, 2 – виражений;
x7 – симптом Розвіга 1 – відсутній, 2 – виражений;
x8 – захисне м'язове напруження 1 – відсутнє, 2 – виражено.

Програма імпорту даних вибірки, виділення з неї підвибірки, яка визначається номером варіанту № та перетворення значень її ознак в рангове представлення містяться в файлі Rank.mcd. Фрагмент програми перетворення вибірки, в якій значення ознак представлені в звичайній порядковій шкалі в вибірку, в якій значення ознак представлені рангами шляхом лінійного масштабування з урахуванням об'єму вибірки, наведений в додатку 2.

Відповідно до номеру варіанта, використовуючи середовище математичних розрахунків MathCAD:

1. Імпортувати дані вибірки у середовище та сформувати підвибірку з варіант с діапазоном індексів: №С+№Г-1÷ №С+№Г +8, №С+№Г +23 ÷ №С+№Г +32, №С+№Г +49 ÷ №С+№Г +58, №С+№Г +74 ÷ №С+№Г +83. Тут №С – номер студента по журналу, №Г – номер групи.
2. Розрахувати матриці парних коефіцієнтів кореляції Пірсона для підвибірок даних в звичайній порядковій та ранговій шкалах.
3. Розрахувати вектор множинних коефіцієнтів кореляції для кореляційних матриць, отриманих з підвибірок даних в звичайній порядковій та ранговій шкалах.
4. Перевірити значимість мінімальних за абсолютними значеннями парного та множинного коефіцієнтів кореляції.
5. Розрахувати матриці парних коефіцієнтів кореляції Спірмена для підвибірок даних в звичайній порядковій та ранговій шкалах. Для

підвибірки даних в ранговій шкалі розрахунок провести без і з поправками на пов'язані ранги.

Вимоги до звіту

Звіт роботі повинен містити:

1. Назву дисципліни та лабораторної роботи.
2. Прізвище, ім'я та по батькові студента, шифр групи, номер варіанта.
3. Об'єкт, предмет і мету лабораторної роботи.
4. Код програми, що реалізує поставлені завдання.
5. Сформовані підвибірки результатів діагностики апендициту, які представлені у порядкових та рангових одиницях.
6. Матриці: парних коефіцієнтів кореляції Пірсона, парних коефіцієнтів кореляції Спірмена без та з поправкою на зв'язані ранги, вектор множинних коефіцієнтів кореляції.
7. Результати перевірки значимості коефіцієнтів кореляції.

Контрольні питання і завдання

1. Запишіть вираз для визначення парного коефіцієнту кореляції Пірсона.
2. Наведіть еліпси розсіювання для системи двох випадкових величин при від'ємному, додатному та нульовому коефіцієнту кореляції між ними.
3. Назвіть основні властивості матриці парних коефіцієнтів кореляції.
4. Поясніть які значення приймають ранги варіант порядкової вибірки та чому дорівнює їх середнє значення.
5. Поясніть у якому випадку у порядковій вибірці виникають зв'язані ранги.
6. Назвіть основні показники рангової кореляції.
7. Дайте визначення порядку часткового коефіцієнта кореляції.
8. Назвіть можливі межі зміни парних, часткових та множинних коефіцієнтів кореляції.
9. Визначте кількість парних та множинних коефіцієнтів кореляції, що нетотожні один одному і не дорівнюють одиниці та відповідають матриці парних коефіцієнтів кореляції розміром $m \cdot m$.
10. Запишіть вираз для визначення множинного коефіцієнту кореляції.
11. Назвіть критерії, які використовуються для перевірки значимості парних та множинних коефіцієнтів кореляції.

2.4 Лабораторна робота № 5
Регресійний аналіз

Об'єкт – сукупності випадкових величини. Предмет – моделі регресійних залежностей. Мета – побудування моделей регресійних залежностей та перевірка їх значимості.

Стислі теоретичні відомості

Завданнями регресійного аналізу є:

- побудова математичної моделі регресії у вигляді залежності середнього значення залежної змінної (змінної стану) від незалежних (факторних) змінних;

- оцінка параметрів моделі регресії і встановлення її відповідності вибірковим спостереженням (оцінка якості моделі регресії);
- визначення точкових та інтервальних прогнозів результативної ознаки.

Регресія – це одностороння стохастична залежність, що встановлює відповідність між випадковими змінними. Якщо досліджують стохастичну залежність змінної Y від X, то встановлюють регресію Y на X (в іншому випадку регресію X на Y). Така одностороння стохастична залежність виражається за допомогою функції регресії або просто регресії.

Щодо числа змінних розрізняють парну регресію – регресія між двома ознаками і багатовимірну (множинну) регресію – регресія між результативною ознакою і декількома факторними ознаками.

За типом з'єднання явищ розрізняють:

- безпосередню регресію, коли причина здійснює прямий вплив на наслідок, тобто результативний і факторний ознаки пов'язані безпосередньо між собою;
- непряму регресію, коли факторний ознака діє через якийсь третій або ряд інших факторних ознак на результативний ознака;
- нонсенс-регресія (помилкова або абсурдна регресія), яка виникає при формальному підході до досліджуваних явищ, в результаті чого приходять до встановлення помилкових і навіть безглуздих залежностей.

Предметом регресійного аналізу є дослідження залежності випадкової величини від сукупності випадкових і невипадкових величин. Регресійний аналіз дозволяє на основі вибіркових спостережень створити математичну модель залежності середнього значення результативної ознаки від факторних ознак.

У загальному вигляді залежність результативної ознаки Y від спільного і одночасного впливу факторних ознак $X1, X2, \ldots, Xk$ (k – кількість факторних ознак) має вигляд:

$$Y = \varphi(X1, X2, \ldots, Xk, a_0, a_1, \ldots, a_m) + e, \qquad (2.4.1)$$

де $\varphi(X1, X2, \ldots, Xk, a_0, a_1, \ldots, a_m)$, – функція регресії, яка виражає об'єктивну закономірну залежність результативної ознаки від спільного впливу факторних ознак, $a_0, a_1, \ldots, a_m$ – коефіцієнти регресії, e – випадкова величина, що виражає вплив неконтрольованих і неврахованих факторних ознак, а також помилок вимірювання.

З останнього виразу випливає, що:

$$e = Y - \varphi(X1, X2, \ldots, Xk, a_0, a_1, \ldots, a_m), \qquad (2.4.2)$$

тобто e – відхилення значень результативної ознаки від значень, обчислених за функцією регресії. Оцінкою функції регресії є рівняння регресії

$$Y(\vec{X}) = f(X1, X2, \ldots, Xk, a_0, a_1, \ldots, a_m). \qquad (2.4.3)$$

Вхідними даними для визначення величин коефіцієнтів регресії є вибірка, кожна варіанта якої представлена певним значенням результативної ознаки і відповідними їй значеннями факторних ознак $\{Y_i, X1_i, X2_i, \ldots, Xk_i\}_{i=1,\ldots,n}$. При визначенні найбільш застосовуваним є метод найменших квадратів, який мінімізує суму квадратів відхилення значень результативної ознаки за даними вибірки від значень, обчислених за функцією регресії

$$E(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} [Y_i - \varphi(X1_i, X2_i, \ldots, Xk_i, a_0, a_1, \ldots, a_m)]^2. \qquad (2.4.4)$$

Система рівнянь для визначення величин коефіцієнтів регресії є системою з $m+1$ рівняння з $m+1$ невідомим, носить назву системи нормальних рівнянь і має вигляд:

$$\begin{cases} \dfrac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_0} = 0 \\ \dfrac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_1} = 0 \\ \vdots \\ \dfrac{\partial E(a_0, a_1, \ldots, a_m)}{\partial a_m} = 0 \end{cases}. \qquad (2.4.5)$$

Рівняння парної лінійної регресії має вигляд:

$$Y(X) = a_1 X + a_0. \qquad (2.4.6)$$

Відповідна система нормальних рівнянь

$$\begin{cases} \dfrac{\partial E(a_0, a_1)}{\partial a_0} = 2\sum_{i=1}^{n} (a_1 X_i + a_0 - Y_i) = 0 \\ \dfrac{\partial E(a_0, a_1)}{\partial a_1} = 2\sum_{i=1}^{n} (a_1 X_i + a_0 - Y_i) X_i = 0 \end{cases}, \text{ або} \qquad (2.4.7)$$

$$\begin{cases} a_1 \sum_{i=1}^{n} X_i + n a_0 = \sum_{i=1}^{n} Y_i \\ a_1 \sum_{i=1}^{n} X_i^2 + a_0 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i X_i \end{cases}. \qquad (2.4.8)$$

Система має просте розв’язання у вигляді:

$$a_1 = \frac{\tilde{C}_{XY}}{\tilde{D}_X}; \quad a_0 = \overline{Y} - \frac{\tilde{C}_{XY}}{\tilde{D}_X}\overline{X}. \tag{2.4.9}$$

Тут $\tilde{D}_X$ – статистична дисперсія змінної X, $\overline{X}$, $\overline{Y}$ – середні значення змінних X та Y, $\tilde{C}_{XY}$ – їх статистична коваріація.

Значення коефіцієнтів регресії також можуть бути отримані за допомогою функцій *slope, intercep*t, *regress* середовища MathCAD.

Рівняння парної квадратичної регресії має вигляд:

$$Y(X) = a_2 X^2 + a_1 X + a_0. \tag{2.4.10}$$

Відповідна система нормальних рівнянь

$$\begin{cases} \dfrac{\partial E(a_0, a_1, a_2)}{\partial a_0} = 2\sum\limits_{i=1}^{n}\left(a_2 X_i^2 + a_1 X_i + a_0 - Y_i\right) = 0 \\ \dfrac{\partial E(a_0, a_1, a_2)}{\partial a_1} = 2\sum\limits_{i=1}^{n}\left(a_2 X_i^2 + a_1 X_i + a_0 - Y_i\right)X_i = 0 \\ \dfrac{\partial E(a_0, a_1, a_2)}{\partial a_2} = 2\sum\limits_{i=1}^{n}\left(a_2 X_i^2 + a_1 X_i + a_0 - Y_i\right)X_i^2 = 0 \end{cases} \text{, або} \tag{2.4.11}$$

$$\begin{cases} a_2\sum\limits_{i=1}^{n} X_i^2 + a_1\sum\limits_{i=1}^{n} X_i + na_0 = \sum\limits_{i=1}^{n} Y_i \\ a_2\sum\limits_{i=1}^{n} X_i^3 + a_1\sum\limits_{i=1}^{n} X_i^2 + a_0\sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} Y_i X_i \\ a_2\sum\limits_{i=1}^{n} X_i^4 + a_1\sum\limits_{i=1}^{n} X_i^3 + a_0\sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} Y_i X_i^2 \end{cases}. \tag{2.4.12}$$

Рівняння парної гіперболічної регресії має вигляд:

$$Y(X) = \frac{a_1}{X} + a_0. \tag{2.4.13}$$

Відповідна система нормальних рівнянь

$$\begin{cases} \dfrac{\partial E(a_0, a_1)}{\partial a_0} = 2\sum\limits_{i=1}^{n}\left(\dfrac{a_1}{X_i} + a_0 - Y_i\right) = 0 \\ \dfrac{\partial E(a_0, a_1)}{\partial a_1} = 2\sum\limits_{i=1}^{n}\dfrac{1}{X_i}\left(\dfrac{a_1}{X_i} + a_0 - Y_i\right) = 0 \end{cases} \text{, або} \tag{2.4.14}$$

$$\begin{cases} a_1 \sum_{i=1}^{n} \frac{1}{X_i} + n a_0 = \sum_{i=1}^{n} Y_i \\ a_1 \sum_{i=1}^{n} \frac{1}{X_i^2} + a_0 \sum_{i=1}^{n} \frac{1}{X_i} = \sum_{i=1}^{n} \frac{Y_i}{X_i} \end{cases} \tag{2.4.15}$$

Рівняння парної логарифмічної регресії має вигляд:

$$Y(X) = a_1 \ln(X) + a_0. \tag{2.4.16}$$

Відповідна система нормальних рівнянь

$$\begin{cases} \frac{\partial E(a_0, a_1)}{\partial a_0} = 2\sum_{i=1}^{n} \left(a_1 \ln(X_i) + a_0 - Y_i \right) = 0 \\ \frac{\partial E(a_0, a_1)}{\partial a_1} = 2\sum_{i=1}^{n} \left(a_1 \ln(X_i) + a_0 - Y_i \right) \ln(X_i) = 0 \end{cases}, \text{ або} \tag{2.4.17}$$

$$\begin{cases} a_1 \sum_{i=1}^{n} \ln(X_i) + n a_0 = \sum_{i=1}^{n} Y_i \\ a_1 \sum_{i=1}^{n} \ln^2(X_i) + a_0 \sum_{i=1}^{n} \ln(X_i) = \sum_{i=1}^{n} Y_i \ln(X_i) \end{cases} \tag{2.4.18}$$

Критерієм оцінки якості отриманої моделі регресії є оцінка статистичної значущості рівняння регресії в цілому і окремих його параметрів. Оцінка статистичної значущості рівняння регресії в цілому проводиться за допомогою F-критерію Фішера, а оцінка статистичної значущості її параметрів, в разі значущості рівняння регресії в цілому, проводиться за t-розподілом Стьюдента.

Для перевірки нульової гіпотези про незначущості рівняння регресії обчислюють контрольне значення величини:

$$F = \left(\frac{Q_{reg}}{Q_{res}} \right) \cdot \left(\frac{k_2}{k_1} \right), \tag{2.4.19}$$

яка підпорядковується статистиці Фішера з $k_1 = 1,\ k_2 = n - 2$ ступенями волі.

$$Q_{tot} = \sum_{i=1}^{n} (\bar{Y} - Y_i)^2, \tag{2.4.20}$$

$$Q_{res} = \sum_{i=1}^{n} \left[Y(X_i) - Y_i \right]^2 = \sum_{i=1}^{n} \left(a_0 + a_1 X_i - Y_i \right)^2, \tag{2.4.21}$$

$$Q_{reg} = Q_{tot} - Q_{res}. \tag{2.4.22}$$

Тут $Q_{tot}, Q_{reg}, Q_{res}$ – відповідно повна, регресійна і залишкова суми квадратів відхилення результативної ознаки, $\bar{Y}$ – його середнє значення. Регресійна сума квадратів відхилень викликається розсіюванням значень результативної ознаки під впливом факторної ознаки включеної в модель, а залишкова характеризується впливом неврахованих факторних ознак. Чим менше залишкове розсіювання, тим менше вплив неврахованих факторних ознак і тим краще математична модель регресії відповідає даним спостережень, так як зміни результативної ознаки в цьому випадку в основному пояснюється впливом розглянутих факторних ознак. величина

$$R^2 = \frac{Q_{reg}}{Q_{tot}}, \tag{2.4.23}$$

носить назву коефіцієнту детермінації. Він являє собою частину повного розсіювання значень результативної ознаки під впливом розглянутих факторних ознак. Чим більше коефіцієнт детермінації, тим краще обрана модель регресії відповідає даним спостережень. Якщо коефіцієнт детермінації дорівнює одиниці, то всі дані спостережень розташовані на лінії регресії.

У разі, якщо контрольне значення перевищує критичне, яке визначається статистикою Фішера $F \geq F_{kp}(\alpha, k_1, k_2)$, рівняння регресії вважається значущим.

Для оцінки статистичної значущості (відмінності від нуля) коефіцієнтів моделі регресії обчислюються значення контрольних величин

$$t = \left| \frac{a}{\tilde{s}_a} \right|, \tag{2.4.24}$$

яка підпорядковується статистиці Стьюдента з $k = n - 2$ ступенями волі. Тут $a,\ \tilde{s}_a$ – відповідно оцінки значення коефіцієнта і його стандартної похибки. Останні розраховуються за формулами

$$\tilde{s}_{a_0} = \tilde{\sigma}_{res} \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2}}, \tag{2.4.25}$$

$$\tilde{s}_{a_1} = \frac{\tilde{\sigma}_{res}}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}, \tag{2.4.26}$$

$$\tilde{\sigma}_{res} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - Y(X_i))^2}{n-2}}, \tag{2.4.27}$$

де $X_i, \overline{X}$ – відповідно i – е та середнє значення факторної ознаки, $D_{res} = \tilde{\sigma}^2_{res}$ – залишкова дисперсія.

У разі, якщо контрольне значення перевищує критичне, яке визначається статистикою Стьюдента $t \geq t_{kp}(\alpha, k)$ відповідний коефіцієнт регресії вважається значущим. Величина довірчого інтервалу для значущого коефіцієнта визначається як

$$\Delta_a = t_{kp}(\alpha, k) \cdot \tilde{s}_a \qquad (2.4.28)$$

При оцінці якості моделі регресії можливі наступні випадки:

- рівняння регресії на основі перевірки за F-критерієм Фішера в цілому статистично значимо і його параметри статистично значущі. Така модель може бути використана для прийняття рішень і прогнозування;
- рівняння регресії за критерієм Фішера статистично значуще, але хоча б один його параметр статистично незначущий. У цьому випадку модель придатна для прийняття деяких рішень, але не для прогнозування;
- рівняння регресії за критерієм Фішера статистично незначуще. У цьому випадку модель регресії вважається ненадійною і непридатною для використання в практиці.

Постановка задачі

Дано

Поле кореляції випадкових значень незалежної (факторної змінної) і залежної (результативної змінної).

Потрібно

Сформувати: з поля кореляції вектори значень незалежної і відповідних значень залежної змінної згідно номеру варіанту.

Визначити: оцінки коефіцієнтів лінійної та нелінійної регресій, значення повної, регресійної і залишкової суми квадратів відхилення результативної ознаки та коефіцієнти детермінації для них.

Побудувати в єдиному графічному полі лінійну, нелінійну регресійні залежності та крапки, що відповідають даним вибірки.

Перевірити значущість рівняння лінійної регресії та її коефіцієнтів.

Рекомендації щодо виконання роботи

В додатку 3 представлені дані вибірок, що містять по 30-ть варіант та підпорядковуються різним нелінійним залежностям. Для кожної групи потоку передбачені окремі вибірки. В файлі *Variant.mcd* (додаток 4) міститься програма, яка дозволяє за номером студента по журналу обрати номери 10-ти варіант вибірок для подальшого опрацювання.

Відповідно до номеру варіанта та групи, використовуючи середовище математичних розрахунків MathCAD:

1. Визначити дані для подальшого опрацювання та занести їх в середовище MathCAD. Студенти з номерами за журналом групи № обробляють дані квадратичних залежностей, якщо залишок від ділення номеру на 3 дорівнює 0,

гіперболічних, якщо залишок дорівнює 1 та логарифмічних залежностей, якщо залишок дорівнює 2. Згідно номеру студента за журналом, за допомогою програми, що міститься в файлі *Variant.mcd,* з відповідної вибірки обираються 10 варіант для подальшого опрацювання. Так, наприклад, студент з номером по журналу 5 буде опрацьовувати дані для логарифмічної регресії з номерами варіант 3, 5, 8, 11, 14, 17, 20, 23, 26, 29.

2. Розрахувати оцінки коефіцієнтів парних лінійної та нелінійної регресій. Розрахунок здійснюється шляхом розв'язання системи лінійних нормальних рівнянь відносно невідомих оцінок коефіцієнтів. Система лінійних рівнянь у середовищі MathCAD може бути розв'язана в один з наступних способів: використання розв'язуючого блока *Given-Find*, функцією *lsolve*, матричним методом – шляхом множення зворотної матриці коефіцієнтів системи на вектор її вільних членів або методом Крамера.

3. Знайти значення повної, регресійної і залишкової суми квадратів відхилення результативної ознаки та коефіцієнти детермінації для лінійної та нелінійної регресій.

4. Побудувати: в єдиному графічному полі лінійну, нелінійну регресійні залежності та крапки, що відповідають даним вибірки.

5. Перевірити: значущість рівняння лінійної регресії та її коефіцієнтів.

Вимоги до звіту

Звіт роботі повинен містити:

1. Назву дисципліни та лабораторної роботи.

2. Прізвище, ім'я та по батькові студента, шифр групи, номер варіанта.

3. Об'єкт, предмет і мету лабораторної роботи.

4. Код програми, що реалізує поставлені завдання.

5. Результати розрахунків оцінок коефіцієнтів лінійної та нелінійної регресій, значення повної, регресійної і залишкової суми квадратів відхилення результативної ознаки та коефіцієнтів детермінації для них.

6. Графічне подання лінійної, нелінійної регресійних залежностей та вхідних даних.

7. Результати перевірки значущості рівняння лінійної регресії та її коефіцієнтів.

Контрольні питання і завдання

1. Назвіть завдання регресійного аналізу.

2. Поясніть яка величина мінімізується при застосуванні методу найменших квадратів у регресійному аналізі.

3. Поясніть що є вхідними даними для визначення величин коефіцієнтів регресії.

4. Поясніть для чого використовується і що являє собою система нормальних рівнянь.

5. Поясніть на якому положенні математичного аналізу ґрунтуються системи нормальних рівнянь.

6. Запишіть загальне рівняння парної кубічної регресії та відповідну до нього систему нормальних рівнянь.

7. Поясніть що характеризують регресійна і залишкова суми квадратів відхилення результативної ознаки.

8. Запишіть вирази для повної та залишкової сум квадратів відхилення результативної ознаки

9. Дайте визначення коефіцієнту детермінації, поясніть що він характеризує та вкажіть його можливі межі зміни.

10. Назвіть критерії, які використовуються для перевірки значущості рівняння лінійної регресії та її коефіцієнтів.

11. Сформулюйте нульову гіпотезу тесту з перевірки статистичної значущості коефіцієнта моделі регресії.

12. Запишіть вирази для контрольних величин тестів для перевірки значущості рівняння лінійної регресії та її коефіцієнтів.

2.5 Лабораторна робота № 6
Кластерний аналіз

Об'єкт – вибірка результатів діагностики апендициту. Предмет – методи кластерного аналізу. Мета – кластеризація вибірки за значеннями факторних ознак (симптомів хвороби) для встановлення необхідності оперативного втручання.

Стислі теоретичні відомості

Класичними методами класифікації без навчання є методи кластерного аналізу (таксономії). За їх допомогою вирішують проблему такого розбиття (кластеризації) множини об'єктів, за якого всі об'єкти, що належать до одного класу, були б більш подібними один до одного, ніж до об'єктів інших класів.

Розглянемо математичну постановку завдання кластерного аналізу. Припустимо, що існує множина об'єктів, що підлягають класифікації $\mathbf{I} = \left\{ \vec{I}_j \right\}_{j=1,\ldots,n}$. Кожний об'єкт характеризується множиною ознак $\left\{ C_i \right\}_{i=1,\ldots,p}$. Ознаки можуть бути як невипадковими так і випадковими величинами та вимірюваними в будь-яких шкалах. Як правило, перед процедурою кластерного аналізу застосовується стандартизація даних – значень *i*-ї ознаки для *j*-го об'єкту ξ_{ji} згідно співвідношення

$$x_{ji} = \frac{\xi_{ji} - \overline{\xi_i}}{\sqrt{D_i}}. \qquad (2.5.1)$$

Тут $\overline{\xi_i}$, D_i – відповідно середнє значення та статистична дисперсія *i*-ї ознаки.

Це дозволяє перейти до нормованих та безрозмірних величин. Стандартизовані результати виміру *i*-ї ознаки для *j*-го об'єкту задаються матрицею спостережень

$\left\{x_{ji}\right\}_{\substack{j=1,\dots,n \\ i=1,\dots,p}}$, а сукупність всіх ознак для кожного з об'єктів – множиною векторів $\mathbf{X}=\left\{\overrightarrow{X}_j\right\}_{j=1,\dots,n}$. Нехай m – ціле число, $m<n$, яке може бути заздалегідь невідоме. Завдання кластерного аналізу полягає в тому щоб на основі даних про величини ознак об'єктів розбити множину об'єктів на m кластерів $\left\{\pi_k\right\}_{k=1,\dots,m}$, при виконанні умов

$$\begin{cases} \pi_k \neq \varnothing \ \forall k \\ \pi_{k1} \bigcap \pi_{k2} = \varnothing \ \forall k1 \neq k2 \\ \bigcup_{k=1}^{m} \pi_k = \mathbf{I} \end{cases} \tag{2.5.2}$$

При цьому всі об'єкти, що належать до одного кластеру, повинні бути як можна більш схожі між собою, а ті що належать до різних кластерів – більш різні між собою.

Розв'язанням завдання кластерного аналізу є розбиття множини об'єктів на кластери, яке оптимізує цільову функцію, що враховує внутрішньокластерні та міжкластерні міри збіжності. Прикладом такої функції може виступати сума по всіх кластерах внутрішньокластерних сум квадратів відхилень ознак об'єктів від середніх за кластером. Якщо j-й об'єкт характеризується скалярною ознакою x_j, ця функція буде мати вигляд

$$W = \sum_{k=1}^{m} \sum_{I_j \in \pi_k} \left(x_j - \frac{1}{n_k} \sum_{I_i \in \pi_k} x_i \right)^2 . \tag{2.5.3}$$

Тут n_k – число об'єктів в кластері π_k, $n = \sum_{k=1}^{m} n_k$.

Для формування кластерів застосовують міри відмінності серед яких найбільше поширення отримали міри типу "відстань". При їх застосуванні об'єкти вважають тим більш подібними один до одного, чим меншою є відстань між ними. Як міру відстані (метрику) можна використовувати будь-яку функцію $\rho\left(\overrightarrow{X}_r, \overrightarrow{X}_v\right)$, що визначена на множині $\mathbf{X}=\left\{\overrightarrow{X}_j\right\}_{j=1,\dots,n}$ і задовольняє таким вимогам:

– $\rho\left(\overrightarrow{X}_r, \overrightarrow{X}_v\right) \geq 0 \ \forall r, v \in \mathbf{X}$;

– $\rho\left(\overrightarrow{X}_r, \overrightarrow{X}_v\right) = 0 \Leftrightarrow \overrightarrow{X}_r = \overrightarrow{X}_v$;

– $\rho\left(\vec{X}_r, \vec{X}_v\right) = \rho\left(\vec{X}_v, \vec{X}_r\right)\ \forall r, v \in \mathbf{X}$ (рефлексивність);

$\rho\left(\vec{X}_r, \vec{X}_v\right) \le \rho\left(\vec{X}_r, \vec{X}_w\right) + \rho\left(\vec{X}_w, \vec{X}_v\right)\ \forall r, v, w \in \mathbf{X}$ (правило трикутника).

Відстані між всілякими парами множини з n об'єктів зручно описувати матрицею відстаней $\mathrm{P} = \left\{\rho_{rv}\right\}_{\substack{r=1,\ldots,n \\ v=1,\ldots,n}}$ розміром $n \times n$, яка має наступні властивості:

1. Елементи головної діагоналі є нульовими $\rho_{rr} = 0$;

2. Матриця є симетричною $\rho_{rv} = \rho_{vr}$ або $\mathrm{P} = \mathrm{P}^{\mathbf{T}}$.

Вибір міри відстані істотно впливає на результат класифікації. У якості міри для кількісних ознак найчастіше використовують евклідову відстань

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sqrt{\sum_{i=1}^{p}\left(x_{ri} - x_{vi}\right)^2} \qquad (2.5.4)$$

Приклад. Нехай вектора ознак двох об'єктів $\vec{X}_r = (1,\ 0,\ -1,\ 2);\ \vec{X}_v = (3,\ 3,\ -2,\ -1)$. Евклідова відстань меж ними становитиме

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sqrt{(1-3)^2 + (0-3)^2 + (-1+2)^2 + (2+1)^2} = \sqrt{4+9+1+9} = \sqrt{23}.$$

Поряд з евклідовою використовуються такі міри відстані: зважена евклідова

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sqrt{\sum_{i=1}^{p} w_i \cdot \left(x_{ri} - x_{vi}\right)^2}\ , \qquad (2.5.5)$$

де $0 \le w_i \le 1$ – вага i-ї ознаки, $\sum_{i=1}^{p} w_i = 1$;

манхеттенська відстань (city block, l_1-норма),

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sum_{i=1}^{p}\left|x_{ri} - x_{vi}\right|\ , \qquad (2.5.6)$$

l_p-норма,

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sqrt[p]{\sum_{i=1}^{p} \left|x_{ri} - x_{vi}\right|^p} \ , \tag{2.5.7}$$

супремум-норма,

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sup_i \left\{\left|x_{ri} - x_{vi}\right|\right\} \ , \tag{2.5.8}$$

Відстань Махаланобіса

$$\rho\left(\vec{X}_r, \vec{X}_v\right) = \sqrt{\left(\vec{X}_r - \vec{X}_v\right)^{\mathbf{T}} \cdot S^{-1} \cdot \left(\vec{X}_r - \vec{X}_v\right)} \ , \tag{2.5.9}$$

де S – коваріаційна матриця.

Іноді замість міри відстані застосовується міра подібності яка задовольняє вимогам:

- $0 \le S\left(\vec{X}_r, \vec{X}_v\right) \le 1 \ \forall r, v \in \mathbf{X}$;
- $S\left(\vec{X}_r, \vec{X}_v\right) = 1 \Leftrightarrow \vec{X}_r = \vec{X}_v$;

$S\left(\vec{X}_r, \vec{X}_v\right) = S\left(\vec{X}_v, \vec{X}_r\right) \ \forall r, v \in \mathbf{X}$ (рефлексивність).

Міри подібності між всілякими парами множини з n об'єктів зручно описувати матрицею подібностей $\mathbf{S} = \left\{S_{rv}\right\}_{\substack{r=1,\ldots,n \\ v=1,\ldots,n}}$ розміром $n \times n$, яка має наступні властивості:

1. Елементи головної діагоналі є одиничними $S_{rr} = 0$;

2. Матриця є симетричною $S_{rv} = S_{vr}$ або $\mathbf{S} = \mathbf{S}^{\mathbf{T}}$.

Для порядкових ознак використовуються коефіцієнти рангової кореляції Спірмена й Кендалла. Для номінальних та дихотомічних ознак застосовуються коефіцієнт спряженості Чупрова, коефіцієнти асоціації та колігації Юла, коефіцієнт спряженості Бравайса. Розглянуті показники можна перетворити у відстані, віднімаючи обчислені значення від одиниці.

Стислий огляд методів кластерного аналізу. Залежно від кількості вихідних спостережень виділяють задачі кластеризації невеликих за обсягом (до декількох десятків об'єктів) масивів спостережень і задачі кластеризації великих масивів. Такий поділ зумовлений різницею методів, які доцільно використовувати при кластеризації відповідних даних.

З погляду апріорної інформації про кількість кластерів вирізняють такі типи задач

- із заданою кількістю класів;

- з невідомою кількістю класів, яку треба оцінити;
- з невідомою кількістю класів, яку не потрібно оцінювати (таку задачу зазвичай формулюють як побудову ієрархічного дерева, або дендрограми вхідної сукупності.

Найпоширенішими методами кластерного аналізу є

- метод повного перебирання;
- ієрархічні агломеративні та дивізимні методи (ближнього зв'язку, середнього зв'язку Кінга, Уорда, далекого зв'язку);
- ієрархічні дивізимні методи;
- ітеративні методи групування (метод K-середніх (K-means) Мак-Куїна);
- алгоритми типу розрізування графа (кореляційних плеяд Терентьєва, вроцлавська таксономія);
- метод пошуку локальних згущень;
- факторні методи.

Окремо слід виділити методи нечіткої кластеризації в результаті застосування яких об'єкти можуть з певними ймовірностями потрапляти до декількох кластерів (метод нечітких C-середніх (Fuzzy C-means)).

Метод повного перебирання. Полягає у повному перебиранні всіх можливих варіантів розбиття множини об'єктів на кластери, визначення значення цільової функції для кожного з варіантів і вибір оптимального серед них. Метод застосовується при невеликій кількості об'єктів, що кластеризуються. Число варіантів розбиття n об'єктів на m непорожніх підмножин (число варіантів кластеризації) визначається числом Стірлінга II роду і дорівнює

$$N = \frac{1}{m!}\sum_{j=0}^{m} C_m^j (-1)^j (m-j)^n \qquad (2.5.10)$$

Приклад. Число варіантів кластеризації n=4 об'єктів на m=2 кластери становить

$$N = \frac{1}{2!}\left(C_2^0 \cdot (-1)^0 \cdot (2-0)^4 + C_2^1 \cdot (-1)^1 \cdot (2-1)^4 + C_2^2 \cdot (-1)^2 \cdot (2-2)^4\right) = \frac{1}{2}(16-2) = 7$$

Можливий склад кластерів: (123),(4); (124),(3); (134),(2); (234),(1); (12),(34); (13),(24); (14),(23).

Ієрархічні агломеративні методи. Призначені переважно для побудови ієрархічних дерев відносно невеликих за обсягом сукупностей. Іноді їх використовують також для задач із заданою кількістю класів, або з невідомою кількістю класів, яку треба оцінити. У цьому випадку реалізацію ієрархічного алгоритму продовжують до досягнення кількості класів, яка дорівнює

заздалегідь заданому числу, або до досягнення екстремуму одного з критеріїв якості розбиття.

Перевагами ієрархічних методів є можливість більш повного і тонкого аналізу структури досліджуваної сукупності порівняно з іншими методами, а також наочність подання результатів кластеризації. Їх основними недоліками є громіздкість обчислювальної процедури, яка пов'язана з перерахунком усієї матриці відстаней на кожному кроці.

Метод ближнього зв'язку є найпростішим для розуміння з ієрархічних агломеративних методів кластерного аналізу. Процес в цьому випадку починають з пошуку та об'єднання двох найближчих один до одного об'єктів у матриці відстаней.

На наступному етапі знаходять два наступні найближчі об'єкти й так само до повного вичерпання матриці відстаней. Як правило, робота алгоритму закінчується, коли всі спостереження об'єднані в один кластер. Для відокремлення кластерів після закінчення кластеризації задають пороговий рівень відстані, на якому можна виділити більше, ніж один кластер.

У методі ближнього зв'язку два об'єкти потрапляють до одного й того самого кластера в тому випадку, коли існує ланцюжок близьких один до одного об'єктів, які їх з'єднують. Іноді це призводить до необґрунтованого зарахування об'єктів до одного й того самого кластера (ланцюжковий ефект). У процесі кластеризації можна явно простежити утворення таких ланцюжків. Для запобігання цьому ефекту можна задавати обмеження на максимальну відстань між елементами одного кластера. Кластери, одержувані за методом ближнього зв'язку, не обов'язково бувають опуклими. Залежно від обставин, це можна розглядати і як перевагу, і як недолік методу.

Результати ієрархічних методів кластерного аналізу стають більш наочними, якщо їх подати у вигляді вертикальних або горизонтальних дендрограм. Приклад вертикальної дендрограми наведено на рис. 1. Тут по горизонталі відкладаються номери об'єктів, що кластеризуються, а по вертикалі – значення відстаней. Вертикальні лінії дендрограми відповідають об'єктам та кластерам, а горизонтальними відображається процес об'єднання об'єктів та кластерів нижнього рівня у кластери верхнього рівня. Об'єкти на горизонтальній осі прийнято розташовувати таким чином, щоб дендрограми не мали самоперетинів. За допомогою дендрограми легко визначити склад кластерів для будь-якого їх числа m такого, що $1<m<n$. Для цього треба знайти рівень відстані, якому відповідає горизонтальна пряма, така що кількість її перетинів з вертикальними лініями дендрограми – кластерами дорівнює m. До складу певного отриманого кластеру входять всі об'єкти, що розташовуються у гілках дендрограми нижче точки перетину, яка визначає кластер.

Приклад. Для заданої матриці відстаней між $n=6$ об'єктами потрібно провести кластеризацію ієрархічним агломеративним методом ближнього зв'язку, побудувати дендрограму, визначити кількість та склад кластерів для різних значень відстані об'єднання.

Номер об’єкта 1 2 3 4 5 6

$$\begin{pmatrix} 0 & 0,5 & 0,1 & 0,5 & 0,5 & 0,3 \\ & 0 & 0,5 & 0,4 & 0,6 & 0,5 \\ & & 0 & 0,5 & 0,5 & 0,4 \\ & & & 0 & 0,2 & 0,5 \\ & & & & 0 & 0,5 \\ & & & & & 0 \end{pmatrix}$$

Найближчими один до одного у матриці відстаней є об’єкти 1 та 3. Вони об’єднуються у кластер на відстані 0,1. Два наступні найближчі об’єкти 4 та 5 об’єднуються у окремий кластер на відстані 0,2. На відстані 0,3 між об’єктами 1 та 6 утворюється кластер (1,3,6), на відстані 0,4 між об’єктами 2 та 4 утворюється кластер (2,4,5) і на відстані 0,5 відбувається повне об’єднання. Дендрограма, яка ілюструє описаний процес, представлена на рис. 2.5.1.

Метод середнього зв’язку Кінга подібний до методу ближнього зв’язку. Його відмінність полягає в тому, що об’єднані до одного кластера об’єкти надалі вважають одним об’єктом з усередненими за кластером параметрами. В іншому варіанті методу середнього зв’язку відстань між кластерами розраховують як середнє значення відстаней між усіма можливими парами представників цих кластерів. При використанні методу середнього зв’язку в процесі кластеризації також простежується формування ланцюжків об’єктів.

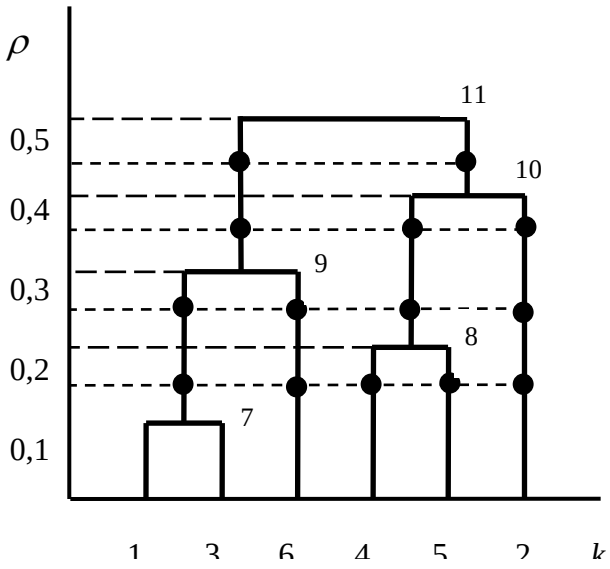

Рис. 2.5.1. Вертикальна дендрограма

Метод Уорда відрізняється від методу середнього зв’язку тим, що підставою для приєднання об’єкта до кластера є не близькість у значенні певної міри відстані, а мінімум дисперсії всередині кластера після поміщення до нього обраного об’єкта.

Формула Ланса і Уільямса дозволяє описати правила групування для будь-якого ієрархічного агломеративного метода

$$\rho(h,k) = A(i)\cdot\rho(h,i) + A(j)\cdot\rho(h,j) + B\cdot\rho(i,j) + C\cdot\left|\rho(h,i) - \rho(h,j)\right|. \quad (2.5.11)$$

Тут $\rho(h,k)$ – відстань між кластерами *h* та *k*, при чому кластер *k* є результатом об'єднання кластерів (або об'єктів *i* та *j*). За допомогою цієї формули можна обрахувати значення відстані $\rho(h,k)$ на якій відстані відбуватиметься чергове об'єднання.

Зокрема для методу ближнього зв'язку $A(i) = A(j) = 0{,}5,\ B = 0,\ C = -0{,}5.$

Для розглянутого вище прикладу утворення кластеру (1,3,6) відбувається на відстані

$$\rho(6,(1,3)) = A(1)\cdot\rho(6,1) + A(3)\cdot\rho(6,3) + C\cdot\left|\rho(6,1) - \rho(6,3)\right| =$$

$$0{,}5\cdot 0{,}3 + 0{,}5\cdot 0{,}4 - 0{,}5\cdot\left|0{,}3 - 0{,}4\right| = 0{,}3.$$

а кластеру (2,4,5) – на відстані

$$\rho(2,(4,5)) = A(4)\cdot\rho(2,4) + A(2)\cdot\rho(2,5) + C\cdot\left|\rho(2,4) - \rho(2,5)\right| =$$

$$0{,}5\cdot 0{,}4 + 0{,}5\cdot 0{,}6 - 0{,}5\cdot\left|0{,}4 - 0{,}6\right| = 0{,}4.$$

Метод K- means Мак-Куїна. Цей метод відноситься_до класу ітеративних методів групування. Вхідними даними для алгоритму є матриця спостережень $\mathbf{X} = \{x_{ji}\}_{\substack{j=1,\dots,n \\ i=1,\dots,p}}$ та кількість кластерів, що формується. *m.* Графічна схема алгоритму кластеризації за методом K- means Мак-Куїна представлена на рис. 2.5.2. Початкове розбиття об'єктів за кластерами є довільним і може бути випадковим. Координати центроідів кластеру π_k визначаються як

$$\bar{x}_{ki} = \frac{1}{n_k}\sum_{I_j\in\pi_k} x_{ji} \ . \quad (2.5.12)$$

Тут n_k – число об'єктів в кластері π_k. За звичай графічне представлення результатів кластеризації цим методом полягає в побудуванні залежностей величин ознак від номеру кластеру. Перевагами методу є простота, недоліками – невелика припустима кількість об'єктів, що кластеризуються та необхідність задання кількості кластерів. Критерієм припинення ітерацій є незмінність розподілу об'єктів за кластерами при здійсненні чергової ітерації.

Метод нечіткої кластеризації Fuzzy C-means. На відміну від чітких, алгоритми нечіткої кластеризації не відносить об'єкт однозначно до якого-небудь кластеру, а визначає для кожного кластеру ймовірність віднесення до нього відповідних об'єктів, формуючи, так звану, матрицю приналежності. Таким чином, на кожному кроці алгоритму кожен об'єкт одночасно відноситься до декількох кластерів.

Як і у попередніх алгоритмів вхідними даними для алгоритму нечіткої кластеризації Fuzzy C-means є матриця спостережень $\mathbf{X} = \{x_{ji}\}_{j=1,\dots,n}$ та кількість кластерів, що формується. *m*.

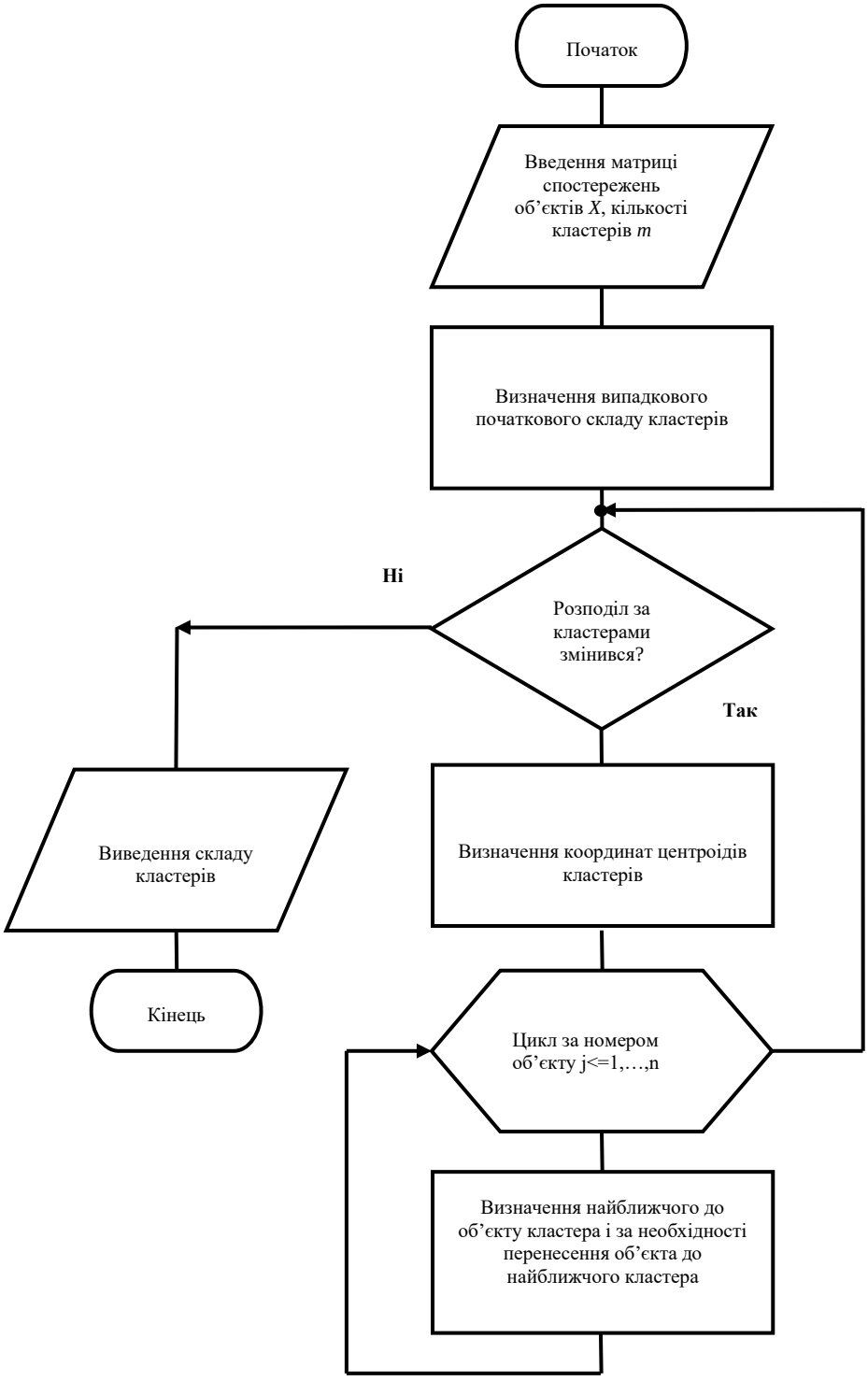

Рис. 2.5.2. Графічна схема алгоритму кластеризації за методом K- means Мак-Куїна

Кластерна структура, що формується надається матрицею приналежності об'єктів до кластерів $\mathbf{M} = \{M_{kj}\}_{\substack{k=1,\ldots,m \\ j=1,\ldots,n}}$. Тут M_{kj} – ступень належності j-го об'єкту k-му кластеру, яка задовольняє наступним вимогам:

– $M_{kj} \in [0;1] \ \forall k, j$;

– $\sum_{k=1}^{m} M_{kj} = 1 \ \forall j$ – тобто кожен об'єкт повинен бути повністю розподілений між кластерами;

– $\sum_{j=1}^{n} M_{kj} \in (0;n) \ \forall k$ – тобто жоден кластер не повинен бути порожнім або містити всі об'єкти.

Для оцінки якості розбиття використовується критерій розкиду, що показує суму відстаней від об'єктів до центрів кластерів з відповідними ступенями приналежності:

$$J = \sum_{k=1}^{m} \sum_{j=1}^{n} M_{kj}^{w} \cdot \rho\left(\vec{\bar{X}}_k, \vec{X}_j\right), \tag{2.5.13}$$

де $\rho\left(\vec{\bar{X}}_k, \vec{X}_j\right)$ – відстань між об'єктом $\vec{X}_j = \left(x_{j1},...,x_{jp}\right)$ та центроїдом k-го кластера $\vec{\bar{X}}_k = \left(x_{k1},...,x_{kp}\right)$, $w \in [0;+\infty)$ – експонентна вага, яка визначає нечіткість, розмитість кластерів. Чим більше це значення, тим значення матриці приналежності для кожного об'єкту більш розмиті по кластерам і при $w \to +\infty$ її елементи приймають значення $M_{kj} = \frac{1}{m}$, тобто всі об'єкти з однаковим ступенем розподілені по всім кластерам. При $w = 1$ алгоритм Fuzzy C-means вироджується в звичайний K-means. Теоретично обґрунтованого правила вибору ваги поки не існує, і зазвичай встановлюють $w = 2$. Елементи матриці координат центроїдів кластерів $\overline{\mathbf{X}} = \left\{\vec{\bar{X}}_k\right\}_{k=1,...,m} = \left\{\bar{x}_{ki}\right\}_{\substack{k=1,...,m \\ i=1,...,p}}$ визначаються як

$$\bar{x}_{ki} = \frac{\sum_{j=1}^{n} M_{kj}^{w} \cdot x_{ji}}{\sum_{j=1}^{n} M_{kj}^{w}} \tag{2.5.14}$$

Завданням нечіткої кластеризації є визначення значень матриці приналежності $\mathbf{M}$, яка мінімізує величину J, що визначається формулою (8). Значення елементів матриці приналежності об'єктів перераховуються за співвідношенням

$$M_{kj} = \begin{cases} \dfrac{1}{\rho_{kj}^{\frac{2}{w-1}} \sum_{l=1}^{m} \dfrac{1}{\rho_{lj}^{\frac{2}{w-1}}}} & при \quad \rho_{kj} > 0 \\ 1 \quad при \quad \rho_{kj} = 0 \end{cases} \tag{2.5.15}$$

Критерієм припинення ітерацій є умова малості зміни матриці приналежності при здійсненні чергової ітерації.

$$\left\| \mathbf{M} - \mathbf{M}^* \right\| = \sqrt{\sum_{k=1}^{m} \sum_{j=1}^{n} \left(M_{kj} - M_{kj}^* \right)^2} < \varepsilon . \qquad (2.5.16)$$

Тут $\mathbf{M}$, $\mathbf{M}^*$ – матриці приналежності на поточному та попередньому кроках ітерації, ε – наперед визначене мале число.

Графічна схема алгоритму кластеризації за методом нечіткої кластеризації Fuzzy C-means представлена на рис. 2.5.3.

Недоліком алгоритму є високий ступінь залежності результуючого розбиття об'єктів на кластери від початкової матриці приналежності.

Постановка задачі

Дано

Багатовимірна вибірка результатів діагностики хворих з підозрою на апендицит. Рядки вибірки відповідають хворим, а стовбці – результативній ознаці – істинному діагнозу та факторним ознакам – симптомам хвороби. Ознаки вибірки надаються у порядковій шкалі та стандартизовані.

Потрібно

Сформувати: з вибірки даних, що відображають результати діагностики апендициту, підвибірку згідно номеру варіанту.

Провести: кластеризацію даних вибірки ієрархічним агломеративним методом та методами K- means, Fuzzy C-means за значеннями факторних ознак (симптомів хвороби) для встановлення необхідності оперативного втручання.

Визначити: відносні частоти виникнення помилок першого та другого роду при встановленні необхідності оперативного втручання.

Дослідити: залежність середньої дисперсії номеру кластеру від коефіцієнта експонентної ваги.

Послідовність виконання роботи

В роботі використовуються ті ж дані, які були описані у роботі 4. Для проведення кластерного аналізу вони додатково стандартизовані та містяться у файлах *standard.txt, standard.xls*.

Відповідно до номеру варіанта, використовуючи середовище математичних розрахунків MathCAD та програмну оболонку інтерпретатора Python, наприклад Anaconda:

1. Імпортувати дані вибірки у середовище розрахунків та сформувати підвибірку з варіант з діапазоном індексів: №С+№Г-1÷ №С+№Г +3, №С+№Г +23 ÷ №С+№Г +27, №С+№Г +49 ÷ №С+№Г +53, №С+№Г +74 ÷ №С+№Г +78. Тут №С – номер студента по журналу, №Г – номер групи.

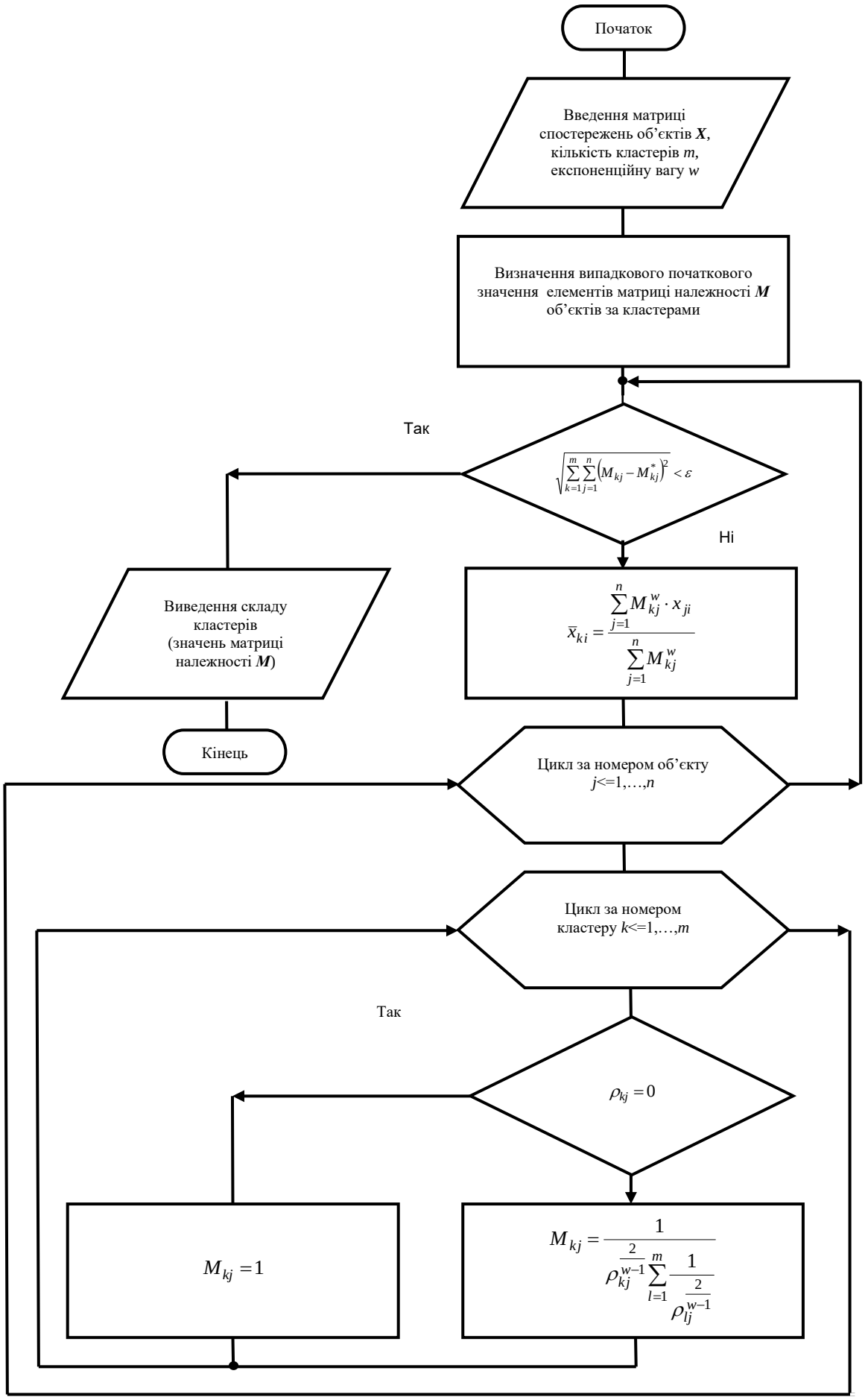

Рис. 2.5.3. Графічна схема алгоритму кластеризації за методом Fuzzy C-means

2. Провести кластеризацію даних методом K-means для випадку 4-х кластерів. Код відповідної MathCAD-програми міститься у файлі Kmeans.mcd, а фрагмент програмного блоку – у додатку 5. Визначити коефіцієнти парної кореляції факторних ознак (симптомів хвороби) з результативною ознакою – істинним діагнозом. Визначити по мінімальному значенню симптому який найбільш корельований з діагнозом, номеру кластеру, що відповідає непідтвердженому діагнозу. Вважаючи, що нульова гіпотеза формулюється як "апендициту немає, хірургічне втручання не потрібне", знайти відносну частота помилки першого роду – апендициту немає, але призначається втручання та відносну частоту помилки другого роду – апендицит є, але втручання не призначається.
3. Провести кластеризацію даних методом Fuzzy C-means для випадку 4-х кластерів і коефіцієнта експонентної ваги w=1.5. Код відповідної MathCAD-програми міститься у файлі *Cmeans.mcd*, а фрагмент програмного блоку – у додатку 5. Визначити коефіцієнти парної кореляції факторних ознак (симптомів хвороби) з результативною ознакою – істинним діагнозом. Визначити по мінімальному значенню симптому який найбільш корельований з діагнозом, номеру кластеру, що відповідає непідтвердженому діагнозу. Знайти відносні частоти помилок першого та другого роду. Побудувати графік залежності середньої дисперсії номеру кластеру від коефіцієнта експонентної ваги.
4. Провести кластеризацію даних ієрархічними агломеративними методами та побудувати дендрограму, на дендрограмі позначити горизонталь, що відповідає випадку 4-х кластерів. Група 1 використовує метод Уорда, група 2 – повного, група 3 – середнього зв'язку. Код Python-програми кластеризації даних ієрархічними агломеративними методами з побудуванням дендрограми міститься у додатку 6. Експортувати дендрограму у вигляді графічного файлу і за допомогою графічного редактору різними кольорами виділити варіанти, що відповідають можливим значенням істинного діагнозу.

Вимоги до звіту

Звіт роботі повинен містити:

1. Назву дисципліни та лабораторної роботи.
2. Прізвище, ім'я та по батькові студента, шифр групи, номер варіанта.
3. Об'єкт, предмет і мету лабораторної роботи.
4. Коди програм, що реалізують поставлені завдання.
5. Сформовану підвибірку стандартизованих результатів діагностики апендициту.
6. Результати кластеризації даних методами K-means та Fuzzy C-means для випадку 4-х кластерів.
7. Результати розрахунку відносних частот помилок першого та другого роду для обох методів.
8. Графік залежності середньої дисперсії номеру кластеру від коефіцієнта експонентної ваги для методу Fuzzy C-means.

9. Дендрограму кластеризації даних ієрархічними агломеративними методами з горизонталлю, що відповідає випадку 4-х кластерів та кольоровим виділенням варіант, що відповідають можливим значенням істинного діагнозу.
10. Висновки.

Контрольні питання і завдання

1. Сформулюйте математичну постановку завдання кластеризації.
2. Наведіть властивості мір відстані та подібності.
3. Знайдіть манхеттенську відстань між векторами $\vec{X}_r = (1,\ 0,\ -1,\ 2);\ \vec{X}_v = (3,\ 3,\ -2,\ -1)$.
4. Назвіть основні групи методів кластерного аналізу.
5. Чому дорівнює число варіантів кластеризації 4-х об'єктів на 3 кластери?
6. Для матриці спостережень $\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ знайти матрицю евклідових відстаней та побудувати дендрограму методом ближнього зв'язку.
7. Назвіть властивості матриці приналежності об'єктів до кластерів.

2.6 Лабораторна робота № 7
Пісочниця Apache Hadoop і команди файлової системи HDFS

Об'єкт – екосистема Apache Hadoop. Предмет – пісочниця Apache Hadoop і файлова система HDFS. Мета – ознайомлення з порядком інсталяції пісочниці Apache Hadoop і вивчення за її допомогою команд файлової системи HDFS.

Стислі теоретичні відомості

Загальні відомості про проект Apache Hadoop. Проект Hadoop заснований у 2005 році, у 2008-му набув статус проекту верхнього рівня фонду Apache Software Foundation є вільно розповсюджуваним набором утиліт, бібліотек і фреймворків для розробки і експлуатації розподілених програм, що працюють на кластерах з сотень і тисяч вузлів. Використовується для реалізації пошукових і тематичних механізмів багатьох високонавантажених веб-сайтів, в тому числі, для Yahoo! і Facebook. Розроблений на Java в рамках обчислювальної парадигми MapReduce, згідно з якою додаток поділяється на велику кількість однакових елементарних завдань, здійсненних на вузлах кластера та зводимих в кінцевий результат. Вважається однією з основоположних технологій Big Data. Протягом 2010 року кілька підпроектів Hadoop: Avro, HBase, Hive, Pig, Zookeeper послідовно стали проектами верхнього рівня фонду Apache, що послужило початком формування екосистеми навколо Hadoop, структура якої наведено на рис. 2.6.1.

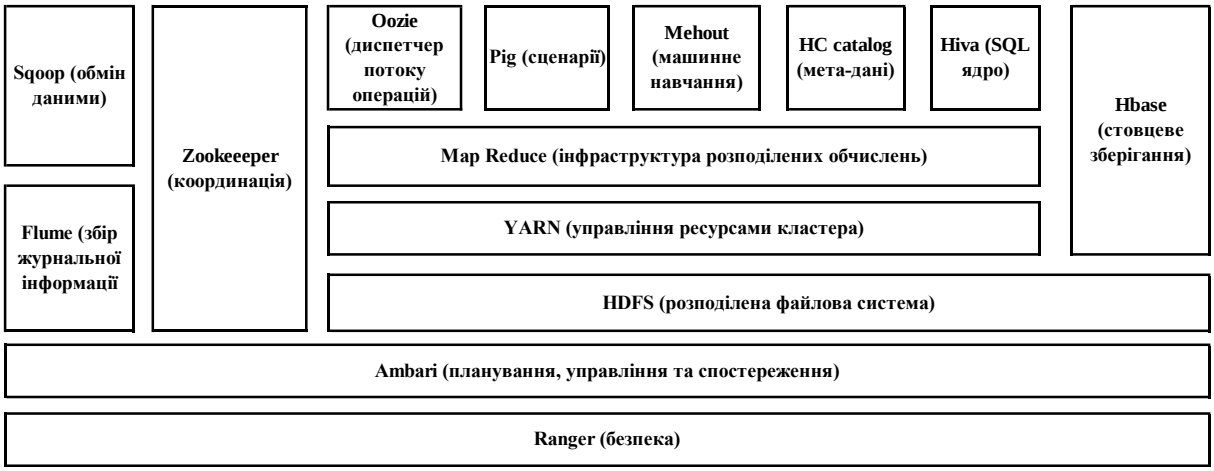

Рис. 2.6.1. Структура екосистеми Hadoop

Станом на 2014 рік, проект складався з чотирьох складових – Hadoop Common (сполучаюче програмне забезпечення – набір інфраструктурних програмних бібліотек і утиліт, використовуваних для інших модулів і споріднених проектів), HDFS (розподілена файлова система), YARN (система для планування завдань і управління кластером) і Hadoop MapReduce (платформа програмування і виконання розподілених MapReduce-обчислень), раніше в Hadoop входив цілий ряд інших проектів, які стали самостійними в рамках системи проектів Apache Software Foundation.

В аналітичному звіті Big Data Analytics Market Study 2017-го року приводиться діаграма (рис. 2.6.2) кількості інфраструктур Big Data, впроваджених на підприємствах, що представлена в розрізі розмірів підприємств.

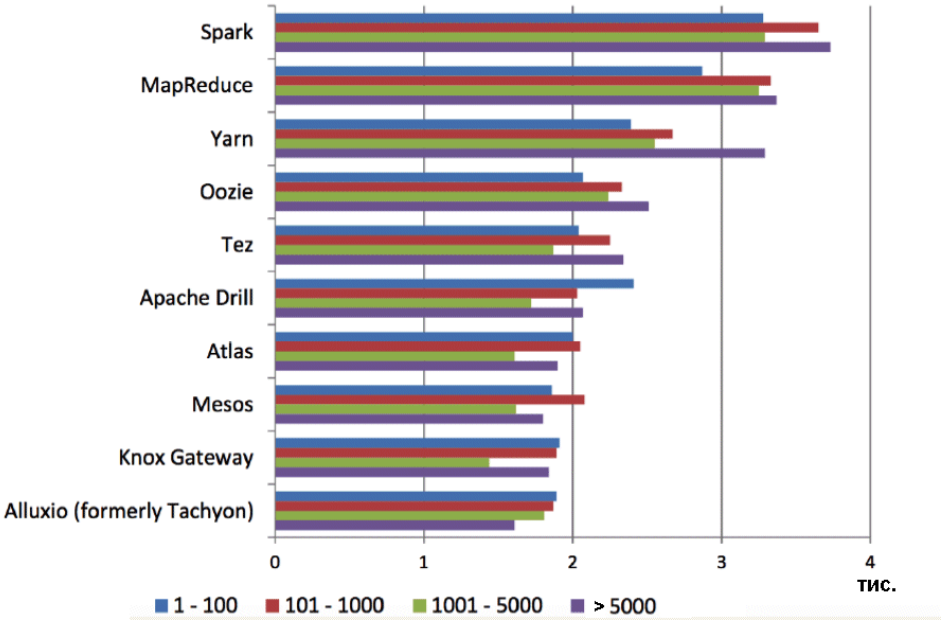

Рис. 2.6.2. Діаграма кількості інфраструктур Big Data, впроваджених на підприємствах станом на 2017 рік

Проект Hadoop має наступні цілі:

- портабельність (кросплатформеність) – можливість інтеграції в проект апаратно-програмного забезпечення різних комп'ютерних платформ.
- надійність, яка досягається резервним копіюванням даних;
- відмовостійкість – виявлення збоїв та автоматичне відновлення нормального функціонування систем;
- горизонтальна масштабованість – можливість кластера за допомогою додавання обладнання масового класу, Commodity hardware;

Hadoop Common. У цю складову інфраструктури входять бібліотеки управління файловими системами, підтримуваними Hadoop, і сценарії створення необхідної інфраструктури та управління розподіленою обробкою даних, для зручності виконання яких створено спеціалізований спрощений інтерпретатор командного рядка (FS shell, filesystem shell), що запускається з оболонки операційної системи сімейства Linux CentOS командою виду: hadoop fs -command URI, де -command – команда інтерпретатора, а URI – список ресурсів з префіксами, що вказують тип підтримуваної файлової системи, наприклад: hdfs://example.com/file1 або file:///tmp/local/file2. Велика частина команд інтерпретатора реалізована за аналогією з відповідними командами Unix (такі, наприклад, cat, chmod, chown, chgrp, cp, du, ls, mkdir, mv, rm, tail, притому, підтримані деякі ключі аналогічних Unix-команд, наприклад ключ рекурсивності -R для chmod, chown, chgrp), є команди, специфічні для Hadoop (наприклад, count підраховує кількість каталогів, файлів і байтів по заданому шляху, expunge очищає кошик, а setrep модифікує коефіцієнт реплікації для заданого ресурсу).

YARN (Yet Another Resource Negotiator) (ще один ресурсний посередник) – модуль, що з'явився з версією 2.0 (2013), що відповідає за управління ресурсами кластерів і планування завдань. Якщо в попередніх випусках ця функція була інтегрована в модуль MapReduce, де була реалізована єдиним компонентом (JobTracker), то в YARN функціонує логічно самостійний демон – планувальник ресурсів (ResourceManager), який абстрагує всі обчислювальні ресурси кластера і керує їх наданням додаткам розподіленої обробки. Працювати під керуванням YARN можуть як MapReduce-програми, так і будь-які інші розподілені додатки, що підтримують відповідні програмні інтерфейси; YARN забезпечує можливість паралельного виконання декількох різних завдань в рамках кластера і їх ізоляцію (за принципами мультіарендності). Розробнику розподіленого додатка необхідно реалізувати спеціальний клас управління додатком (ApplicationMaster), який відповідає за координацію завдань в рамках тих ресурсів, які надасть планувальник ресурсів; планувальник ресурсів же відповідає за створення екземплярів класу управління додатком і взаємодії з ним через відповідний мережевий протокол. YARN може бути розглянутий як кластерна операційна система в тому сенсі, що виступає інтерфейсом між апаратними ресурсами кластера і широким класом додатків, що використовують його потужності для виконання обчислень.

Hadoop MapReduce – програмний каркас для програмування розподілених обчислень в рамках парадигми MapReduce. Для паралельної обробки даних в інтерфейсі MPI (Message Passing Interface), що є поширеним стандартом для обміну даними при організації паралельних обчислень, була запропонована парадигма паралельної обробки наборів даних за допомогою використання функцій Map і Reduce. Розробнику додатка для Hadoop MapReduce необхідно реалізувати обробник, який на кожному обчислювальному вузлі кластера забезпечить перетворення вхідних даних в проміжний набір пар "ключ – значення" (клас, який реалізує інтерфейс Mapper, названий по функції Map) – фаза відображення і обробник, який зведе проміжний набір пар в остаточний, скорочений набір (згортку, клас, який реалізує інтерфейс Reducer, названий по функції Reduce) – фаза згортки. Для ілюстрації роботи цієї парадигми розглянемо наступний приклад. Фірма займається виробництвом іграшок, кожна з котрих фарбується у два кольори. Закази клієнтів приймаються через Web-сторінку і зберігаються в системі, яка підтримує технологію MapReduce. На основі цих даних слід підрахувати потреби у фарбах різних кольорів. Схема вирішення цього завдання за технологію MapReduce наведена на рис. 2.6.3.

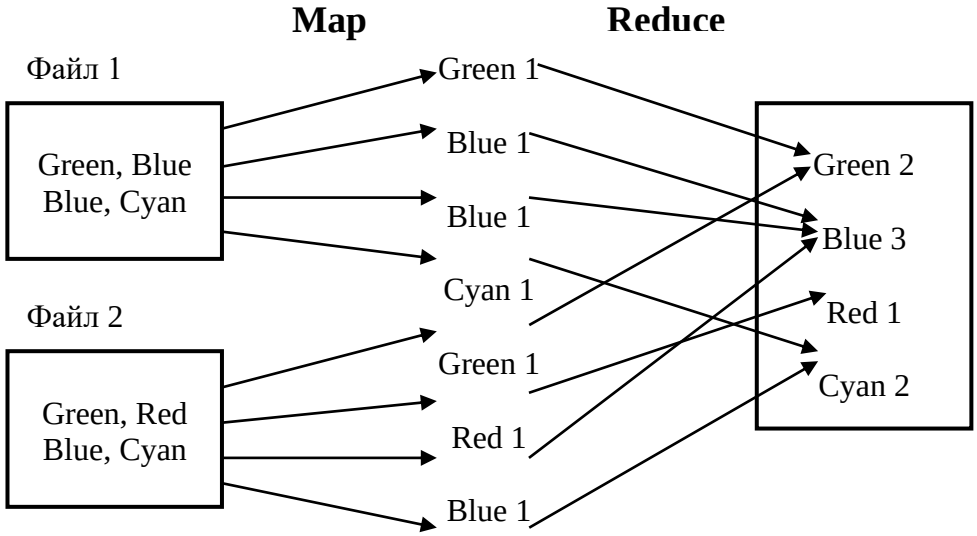

Рис. 2.6.3. Приклад реалізації технології MapReduce

На фазі згортки унікальні екземпляри даних групуються і, в залежності від функції згортки можуть бути отримані різні результати. В нашому випадку функція згортки повертає кількість входжень кожного кольору.

Для ефективного розпаралелювання вхідних даних функція згортки повинна задовольняти умовам комутативності і асоціативності, або принаймні умові асоціативності на множині визначення. У першому випадку є можливість у довільному порядку обирати пари елементів проміжного набору "ключ – значення", здійснювати потрібні паралельні обчислення та згортку отриманих результатів. Тобто, алгоритми, в яких функція згортки комутативна і асоціативна, ефективні при обробці великих як впорядкованих, так і

невпорядкованих наборів даних. У другому випадку для алгоритмів з асоціативними, але некомутативними функціями згортки є можливість обробки наборів даних, елементи яких впорядковані. В цьому випадку вихідний набір розбивається на кілька впорядкованих між собою послідовних піднаборів, наприклад, за кількістю наявних обчислювачів в кластері. Далі обробка здійснюється аналогічним чином, але при згортці суттєвим є порядок слідування результатів обчислення. Вказана вимога разом з вимогою впорядкованості вхідних даних робить такі алгоритми менш ефективними (більш трудомісткими).

Функція $f(a,b)$ задовольняє умові комутативності, якщо $f(a,b) = f(b,a)$, а асоціативності якщо $f(f(a,b),c) = f(a, f(b,c))$.

Наприклад, умовам комутативності і асоціативності задовольняють функції визначення суми $f(a,b) = a + b$, добутку скалярів $f(a,b) = a \cdot b$, функції Наприклад, умовам комутативності і асоціативності задовольняють функції визначення суми $f(a,b) = a + b$, добутку скалярів $f(a,b) = a \cdot b$, функція $f(a,b) = \sqrt{a^2 + b^2}$. Так, для останньої функції

$$f(a,b) = \sqrt{a^2 + b^2} = \sqrt{b^2 + a^2} = f(b,a) \qquad (2.6.1)$$

$$f(f(a,b),c) = \sqrt{\left(\sqrt{a^2 + b^2}\right)^2 + c^2} = \sqrt{a^2 + \left(\sqrt{b^2 + c^2}\right)^2} = f(a, f(b,c)). \qquad (2.6.2)$$

Ні умові комутативності, ні умові асоціативності не задовольняють функції визначення різниці $f(a,b) = a - b$, ділення $f(a,b) = a / b$.

Умові комутативності задовольняє, а умові асоціативності не задовольняє функція $f(a,b) = \sqrt{a \cdot b}$

$$f(a,b) = \sqrt{a \cdot b} = \sqrt{b \cdot a} = f(b,a), \qquad (2.6.3)$$

$$f(f(a,b),c) = \sqrt{\left(\sqrt{a \cdot b}\right) \cdot c} = \sqrt[4]{a \cdot b} \cdot \sqrt{c} \neq \sqrt{a} \cdot \sqrt[4]{b \cdot c} = \sqrt{a\left(\sqrt{b \cdot c}\right)} = f(a, f(b,c)) \qquad (2.6.4)$$

Умові комутативності не задовольняє, а умові асоціативності задовольняє функція добутку матриць $f(A,B) = A \cdot B.$

Умовою можливості обрахування такої функції (здійснення операції множення) є $Ac = Br$, де Ac та Br є відповідно кількість стовбців матриці A та кількість рядків матриці B. Умовою можливості обрахування функції $f(B,A) = B \cdot A$ є $Bc = Ar$, яка в загальному випадку не виконується при виконанні попередньої умови.

З іншого боку, результатом виконання функції $f(A,B) = A \cdot B$ є матриця розміром $Ar \times Bc$. Значення її елементів обчислюються як $\sum_{j=1}^{Ac} A_{ij} \cdot B_{jk}$. Умовою

можливості обрахування функції $f(f(A,B),C)=(A\cdot B)\cdot C$ є $Ac=Br$ та $Bc=Cr$. Значення її елементів обчислюються як $\sum_{j=1}^{Ac} A_{ij}\cdot B_{jk}\cdot\sum_{k=1}^{Bc} C_{kl}$.

Результатом виконання функції $f(B,C)=B\cdot C$ є матриця розміром $Br\times Cc$. Значення її елементів обчислюються як $\sum_{j=1}^{Bc} B_{ij}\cdot C_{jk}$. Умовою можливості обрахування функції $f(A,f(B,C))=A\cdot(B\cdot C)$ є $Bc=Cr$ та $Ac=Br$. Значення її елементів обчислюються як $\sum_{j=1}^{Ac} A_{ij}\cdot\sum_{k=1}^{Bc} B_{jk}\cdot C_{kl}$. В обох випадках умови можливості обчислення функції однакові. Також з урахуванням того, що елементи зазначених сум скаляри, множник можна вносити під знак суми, множення скалярів асоціативно, межі підсумовування однакові, то і результати обчислення сум, а як наслідок і результати обчислення елементів добутку трьох матриць, в обох випадках однакові.

Hadoop MapReduce дозволяє створювати завдання як з базовими обробниками, так і зі згортками, написаними без використання Java: утиліти Hadoop streaming дозволяють використовувати в якості базових обробників і згорток будь-який виконуваний файл, який працює зі стандартним введенням-виведенням операційної системи (наприклад, утиліти командної оболонки UNIX), є також прикладний інтерфейс програмування Hadoop pipes на C ++. Також, до складу дистрибутивів Hadoop входять реалізації різних конкретних базових оброблювачів і згорток, найбільш типово використовуваних в розподіленій обробці.

У перших версіях Hadoop MapReduce включав планувальник завдань (JobTracker), починаючи з версії 2.0 ця функція перенесена в YARN, і починаючи з цієї версії модуль Hadoop MapReduce реалізований поверх YARN. Програмні інтерфейси здебільшого збережені, однак повної сумісності немає (тобто для запуску програм, написаних для попередніх версій API, для роботи в YARN в загальному випадку потрібна їх модифікація або рефакторинг, і лише при деяких обмеженнях можливі варіанти зворотної двійкової сумісності).

HDFS (Hadoop Distributed File System) – розподілена файлова система Hadoop для зберігання файлів великих розмірів з можливістю потокового доступу до інформації, поблоково розподіленої за вузлами обчислювального кластера, який може складатися з довільного апаратного забезпечення. HDFS як і будь-яка файлова система є ієрархією каталогів з вкладеними в них підкаталогами і файлами.

HDFS є невід'ємної частиною Hadoop і основою інфраструктури Big Data. Однак, Hadoop підтримує роботу і з іншими розподіленими файловими системами, зокрема, Amazon S3 і CloudStore. Також деякі дистрибутиви Hadoop, наприклад, MapR, реалізують свою аналогічну розподілену файлову систему MapR File System.

HDFS може використовуватися не тільки для запуску MapReduce-завдань, а й як розподілена файлова система загального призначення, забезпечуючи

роботу розподілених СУБД (HBase) і масштабованих систем машинного навчання (Apache Mahout).

Архітектура HDFS представлена на рис. 2.6.4.

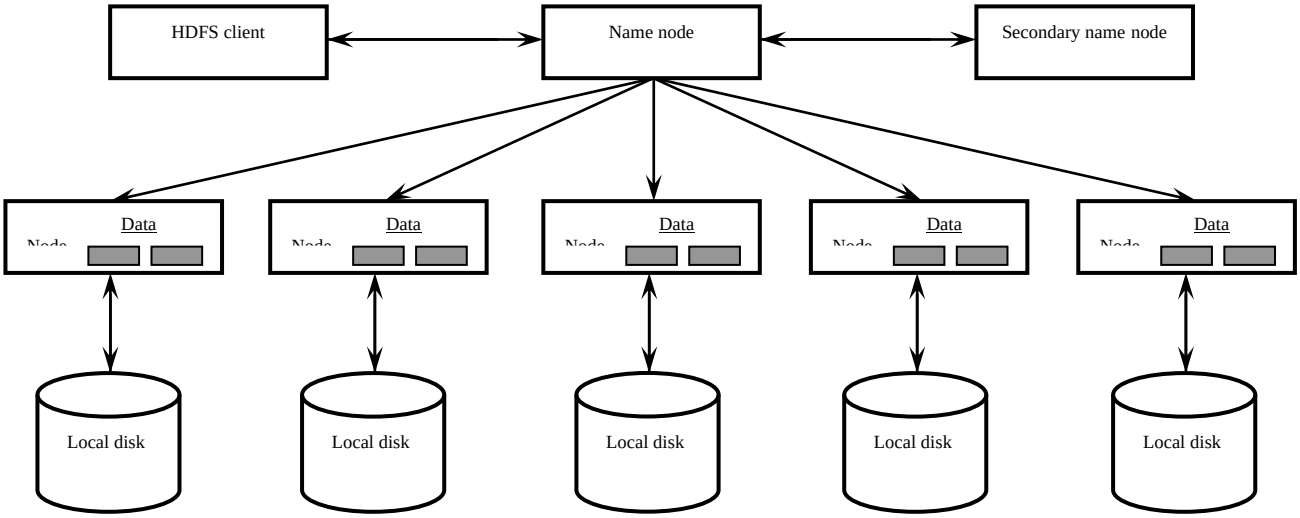

Рис. 2.6.4. Архітектура HDFS

Кластер HDFS включає наступні компоненти.

Керуючий вузол, вузол імен або сервер імен (NameNode) – окремий, єдиний в кластері, сервер з програмним кодом для управління простором імен файлової системи, який зберігає дерево файлів, а також метадані файлів і каталогів. NameNode – обов'язковий компонент кластера HDFS, який відповідає за відкриття і закриття файлів, створення і видалення каталогів, управління доступом з боку зовнішніх клієнтів і відповідність між файлами і блоками, дубльованими (репліційованими) на вузлах даних. Сервер імен розкриває для всіх бажаючих розташування блоків даних на машинах кластера.

Secondary NameNode – вторинний вузол імен, окремий сервер, єдиний в кластері, який копіює образ HDFS і лог транзакцій операцій з файловими блоками в тимчасову папку, застосовує зміни, накопичені в лозі транзакцій до образу HDFS, а також записує його на вузол NameNode і очищає лог транзакцій. Secondary NameNode необхідний для швидкого ручного восстанавленія NameNode в разі його виходу з ладу.

Вузол або сервер даних (DataNode, Node) – один із множини серверів кластера з програмним кодом, що відповідає за файлові операції і роботу з блоками даних. DataNode є обов'язковим компонентом кластеру HDFS, який відповідає за запис і читання даних, виконання команд від вузла NameNode по створенню, видалення і реплікації блоків, а також періодичну відправку повідомлення про стан (heartbeats) і обробку запитів на читання і запис, що надходять від клієнтів файлової системи HDFS. З інших вузлів кластера дані проходять до клієнта повз вузла NameNode.

Клієнт (client) – користувач або додаток, який взаємодіє через спеціальний інтерфейс (API – Application Programming Interface) з розподіленою файловою системою. При наявності достатніх прав, клієнту дозволені наступні операції з

файлами і каталогами: створення, видалення, читання, запис, перейменування і переміщення. Створюючи файл, клієнт може явно вказати розмір блоку файлу (за замовчуванням 64 Мб) і кількість створюваних реплік (за замовчуванням значення дорівнює 3-ом).

Завдяки реплікації блоків по вузлах даних, розподілена файлова система Hadoop забезпечує високу надійність зберігання даних і швидкість обчислень. Крім того, HDFS властива наступна низка відмінностей:

- великий розмір блоку в порівнянні з іншими файловими системами (> 64MB), оскільки HDFS призначена для зберігання великої кількості величезних (> 10GB) файлів;
- орієнтація на недорогі і, тому не найнадійніші сервера – відмовостійкість всього кластера забезпечується за рахунок реплікації даних;
- віддзеркалення і реплікація здійснюються на рівні кластера, а не на рівні вузлів даних;
- реплікація відбувається в асинхронному режимі – інформація розподіляється по декількох серверах прямо під час завантаження, тому вихід з ладу окремих вузлів даних не спричинить за собою повну втрату даних;
- HDFS оптимізована для потокових зчитувань файлів, тому застосовувати її для нерегулярних і довільних зчитувань недоцільно;
- клієнти можуть зчитувати і записувати файли HDFS безпосередньо через програмний інтерфейс Java;
- файли пишуться одноразово, що виключає внесення в них будь-яких довільних змін;
- принцип WORM (Write-once and read-many, один раз записати – багато разів прочитати) повністю звільняє систему від блокувань типу «запис-читання». Запис в файл в один час доступна тільки одному процесу, що виключає конфлікти множинної записи.
- HDFS оптимізована під потокову передачу даних;
- стиснення даних і раціональне використання дискового простору дозволило знизити навантаження на канали передачі даних, які найчастіше є вузьким місцем в розподілених середовищах;
- самодіагностика – кожен вузол даних через певні інтервали часу відправляє діагностичні повідомлення вузлу імен, який записує логи операцій над файлами в спеціальний журнал;
- всі метадані сервера імен зберігаються в оперативній пам'яті.

У зв'язку з особливостями архітектури та принципом дії, для HDFS характерні наступні недоліки:

- сервер імен є центральною точкою всього кластера і його відмова спричинить збій системи цілком;
- відсутність повноцінної реплікації Secondary NameNode;
- відсутність можливості дописувати або залишити відкритим для запису файли в HDFS, за рахунок чого в класичному дистрибутиві Apache Hadoop неможливо оновлювати блоки вже записаних даних;

- відсутність підтримки реляційних моделей даних;
- відсутність інструментів для підтримки посилальної цілісності даних, що не гарантує ідентичність реплік. HDFS перекладає перевірку цілісності даних на клієнтів. При створенні файлу клієнт розраховує контрольні суми кожні 512 байт, які в подальшому зберігаються на сервері імен. При зчитуванні файлу клієнт звертається до даних і контрольних сумах. У разі їх невідповідності відбувається звернення до іншої репліці.

Масштабованість проекту. Однією з основних цілей Hadoop спочатку було забезпечення горизонтальної масштабованості кластера за допомогою додавання недорогих вузлів (обладнання масового класу, Commodity hardware), без вдавання до потужних серверів і дорогих мереж зберігання даних.

Станом на 2011 рік типовий кластер будувався з однопроцесорних багатоядерних x86-64-вузлів під управлінням Linux з 3-12 дисковими пристроями зберігання, пов'язаних мережею з пропускною спроможністю 1 Гбіт/с.

Функціонуючі кластери розміром в тисячі вузлів підтверджують здійсненність і економічну ефективність таких систем, так, станом на 2011 рік відомо про великі кластери Hadoop в Yahoo (більше 4 тис. вузлів з сумарною місткістю зберігання 15 Пбайт (Пета 1015)), Facebook (близько 2 тис. вузлів на 21 Пбайт) і Ebay (700 вузлів на 16 Пбайт).

Проте, вважається, що горизонтальна масштабованість в Hadoop-системах обмежена, для Hadoop до версії 2.0 максимально можливо оцінювалася в 4 тис. вузлів при використанні 10 MapReduce-завдань на вузол. Багато в чому цьому обмеження сприяла концентрація в модулі MapReduce функцій з контролю за життєвим циклом завдань, вважається, що з виносом її в модуль YARN в Hadoop 2.0 і децентралізацією – розподілом частини функцій з моніторингу на вузли обробки, горизонтальна масштабованість підвищилася.

Ще одним обмеженням Hadoop-систем є розмір оперативної пам'яті на вузлі імен (NameNode), що зберігає весь простір імен кластера для розподілу обробки, до того ж загальна кількість файлів, яку здатний обробляти вузол імен – 100 млн. Для подолання цього обмеження ведуться роботи з розподілу вузла імен, єдиного в поточній архітектурі на весь кластер, на кілька незалежних вузлів. Іншим варіантом подолання цього обмеження є використання розподілених СУБД поверх HDFS, таких як HBase, роль файлів і каталогів в яких з точки зору програми грають записи в одній великій таблиці бази даних.

Пісочниця Apache Hadoop від Hortonworks. Пісочниця Apache Hadoop від Hortonworks є локальним середовищем розробки для ознайомлення з Hadoop, розподіленої файлової системою Hadoop (HDFS) і відправкою завдань, яка дозволяє ознайомитися з екосистемою Hadoop. Розглянемо налаштування на віртуальній машині Apache Hadoop від Hortonworks. Додаткові відомості з цього питання можуть бути отримані за посиланням https://docs.microsoft.com/ru-ru/azure/hdinsight/hadoop/apache-hadoop-emulator-get-started.

Для доступу до пісочниці слід встановити оболонку віртуальних машин Oracle VM VirtualBox, яка дозволяє у середовищі встановленої в комп'ютері операційної системи емулювати інші операційні системи та віртуальну машину HDP Sandbox, що містить віртуальний образ операційної системи CentOS зі встановленим інтерфейсом до Hadoop. Інсталятор першого програмного продукта можна отримати за посиланням https://www.virtualbox.org/wiki/Downloads. Після встановлення VirtualBox platform packages потрібно також встановити пакет розширення VirtualBox Extension Pack тієї самої ж версії, що і основний. Другий програмний продукт можна отримати за адресою https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html. При цьому слід мати на увазі, що доступні наразі версії HDP Sandbox містяться у файлах великих розмірів: версія 2.5 – 11,5 Гб, 2.6.5 – 15,7 Гб, 3.0.1 – 21,6 Гб. В розгорнутому вигляді вони відповідно потребують: версія 2.5 – 21,8 Гб, 2.6.5 – 62,5 Гб дискового простору.

Образ віртуальної машини HDP Sandbox слід імпортувати у середовище VirtualBox. Для цього слід скористатися пунктом меню *Файл – Імпорт конфігурацій* (Ctrl+I) (рис. 2.6.5) і в відкришімся вікні діалогу вказати файл HDP Sandbox (цей файл має відкритий формат віртуалізації ova).

Далі потрібно налаштувати віртуальну машину, натиснувши на кнопку *Налаштувати* (Ctrl+S). Слід зазначити, що віртуальна машина HDP Sandbox висуває значні вимоги до апаратної частини комп'ютера, на який вона встановлюється. Приведені нижче на рис. 2.6.6, 2.6.7 параметри комп'ютера слід вважати близькими до мінімальної необхідної конфігурації для встановлення віртуальної машини HDP Sandbox.

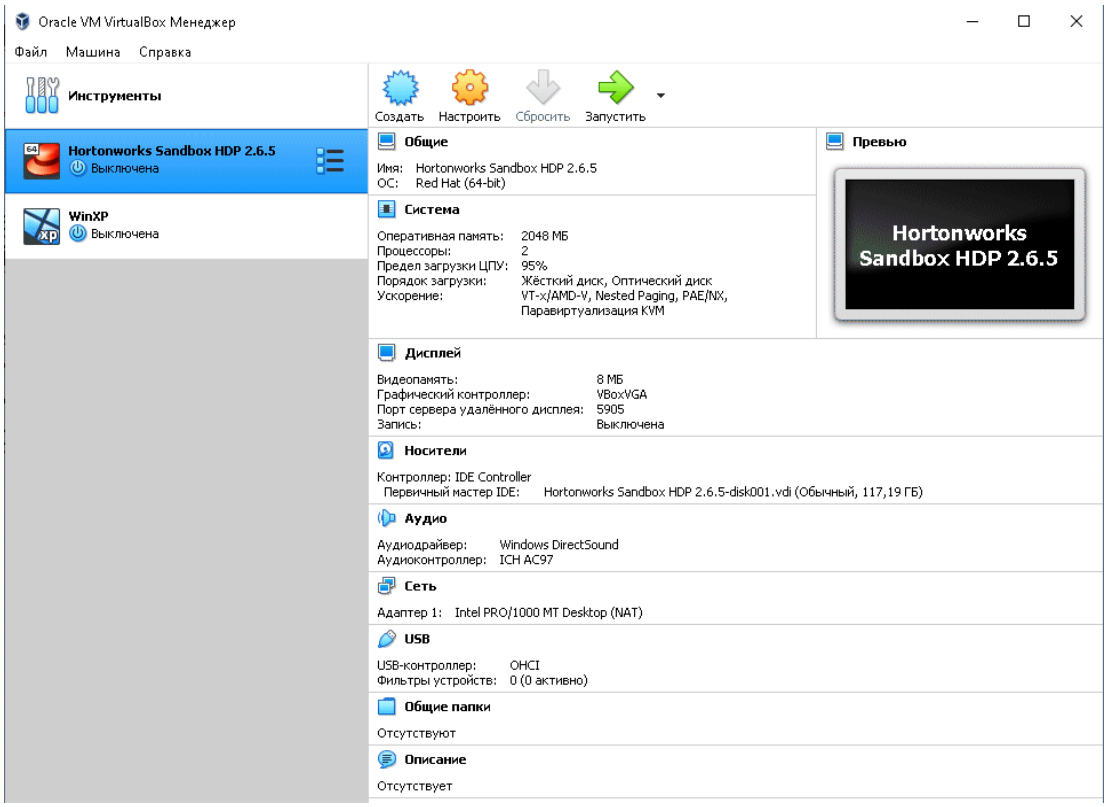

Рис. 2.6.5. Діалогове вікно середовища Oracle VM VirtualBox

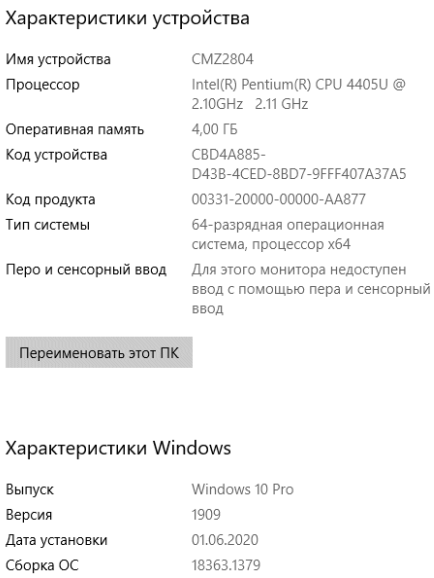

Рис. 2.6.6. Параметри комп’ютера, на який встановлюється віртуальна машина HDP Sandbox (початок)

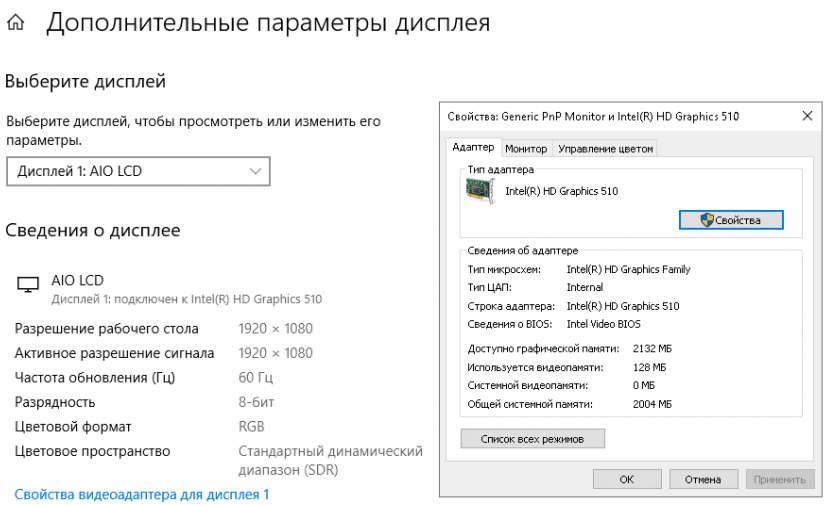

Рис. 2.6.7. Параметри комп’ютера, на який встановлюється віртуальна машина HDP Sandbox (продовження)

Найбільш критичними при налаштуванні є розділи *Система – Материнська* плата и *Система – Процесор*. Відомості про невірні налаштування містяться у спливаючих вікнах, які розташовуються знизу вікна діалогу налаштування та мають вигляд, представлений на рис. 2.6.8, 2.6.9.

Після налаштування віртуальну машину можна запустити кнопкою Запустить. Коли віртуальна машина стартує на екрані з’явиться вікно, представлене на рис. 2.6.10.

Для доступу до пісочниці використовується протокол SSH, широко використовуваний в операційних системах сімейства Linux для віддаленого доступу між машинами. При доступі до віддаленого комп'ютера Linux з Windows в останній повинен бути встановлений клієнт SSH. Але для версій 2.5 та 2.6.5 можна використовувати протокол SSH, наданий віртуальною машиною за адресою http://localhost:4200/.

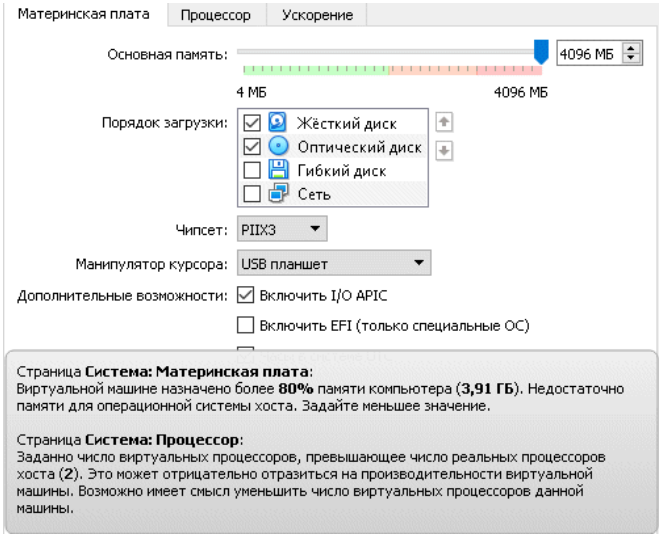

Рис. 2.6.8 Відомості про невірні налаштування комп’ютера, на який встановлюється віртуальна машина HDP Sandbox (початок)

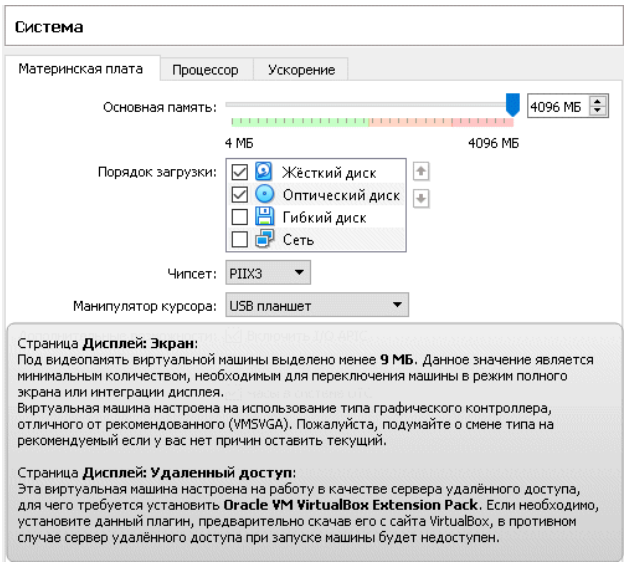

Рис. 2.6.9. Відомості про невірні налаштування комп’ютера, на який встановлюється віртуальна машина HDP Sandbox (продовження)

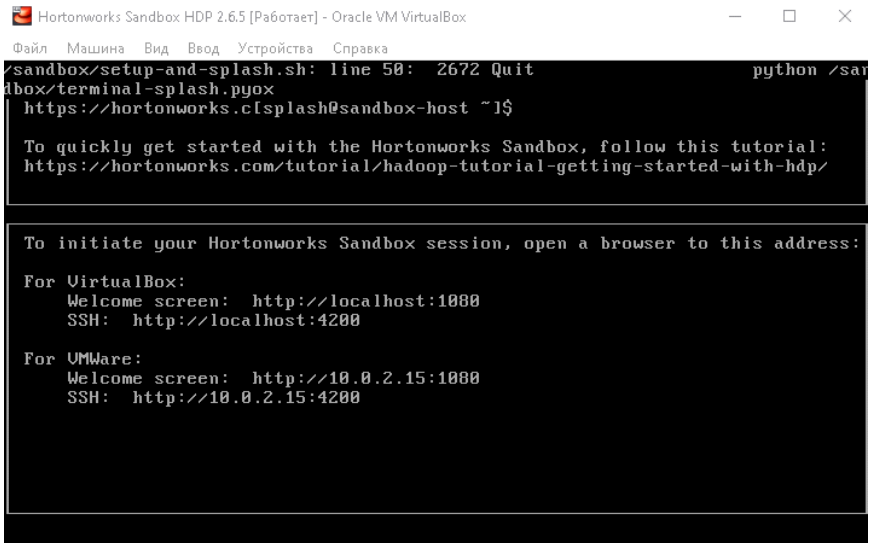

Рис. 2.6.10. Ініціалізація роботи з пісочницею Apache Hadoop

При першому підключенні до пісочниці за допомогою SSH треба ввести ім‘я користувача root та пароль за замовченням hadoop, після чого буде запропоновано змінити пароль для облікового запису root, у відповідь на що треба ввести новий пароль, який подалі буде використовуватися при вході. Після цього відкривається доступ до командного рядка файлової системи HDFS.

Формат команд HDFS (shell команд) має наступний вигляд:

Версія 1 hadoop fs -command -<options> <URI>

Версія 2 hdfs dfs -command -<options> <URI>

Далі використовується друга версія формату.

Тут:

- hdfs – утиліта роботи з HDFS;
- dfs – спеціальний параметр, який позначає, що ми працюватимемо безпосередньо з розподіленою файловою системою. Можуть бути вказані інші параметри;
- <command> – команда, яку ми хочемо застосувати до файлової системи;
- <options> – опції команди (можуть відсутні);
- <URI> – шлях у вигляді URI-схеми. URI (Uniform Resource Identifier) – уніфікований ідентифікатор ресурсу – символьний рядок, що дозволяє ідентифікувати будь-який ресурс: документ, зображення, файл, службу, скриньку електронної пошти тощо, насамперед, йдеться про ресурси мережі Інтернет.

Деякі команди файлової системи HDFS в форматі версії 2 наведено в табл. 2.6.1.

Таблиця 2.6.1. Команди файлової системи HDFS

Команда	Дія, яка виконується командою
Команди виведення списків	
hdfs dfs	Виведення повного списку команд HDFS
hdfs dfs -ls /	Виведення списку всіх підкаталогів/файлів в кореневому каталозі HDFS
hdfs dfs -ls /DirectoryName	Виведення списку всіх підкаталогів/файлів в локальному каталозі або каталозі HDFS
hdfs dfs -ls -R /DirectoryName	Рекурсивне виведення списку всіх підкаталогів/файлів локальному каталозі або каталозі HDFS (виведення змісту всього дерева підкаталогів для якого зазначений каталог є вершиною)
Команди управління каталогами і файлами	
hdfs dfs -rmdir /DirectoryName	Вилучення локального каталогу або каталогу HDFS
hdfs dfs -mkdir /DirectoryName	Створення локального каталогу або каталогу HDFS
hdfs dfs -chmod code /DirectoryName[/FileName]	Зміна прав доступу до каталогу/файлу. При значенні code=777 надається повний доступ (читання r, запис w, запуск x)
hdfs dfs -cp [-f] [-p] /DirectoryName1/FileName1 /DirectoryName2/ FileName2	Копіювання файла з одного локального каталогу або каталогу HDFS в інший каталог. При копіюванні ім’я файлу може бути змінено. Необов’язковий параметр f

	дозволяє перезаписувати файл з тим же ім'ям у каталогу призначення, якщо він вже існує, параметр p дозволяє зберігати режим доступу до файла
hdfs dfs -mv /DirectoryName1/FileName /DirectoryName2	Переміщення файла з одного локального каталогу або каталогу HDFS в іншій каталог.
hdfs dfs -touchz /DirectoryName/FileName	Створення порожнього файлу в заданому локальному каталозі або каталозі HDFS
hdfs dfs -put /localDirectoryName/localfFileName /hdfsDirectoryName/hdfsFileName	Копіювання файла з каталогу локальної файлової системи в каталог HDFS. При копіюванні ім'я файлу може бути змінено.
hdfs dfs -get /hdfsDirectoryName/hdfsFileName /localDirectoryName/localfFileName	Копіювання файла з каталогу HDFS в каталог локальної файлової системи. При копіюванні ім'я файлу може бути змінено.
hdfs dfs -cat /DirectoryName/FileName	Виведення змісту локального файла або файла HDFS
hdfs dfs -text /DirectoryName/FileName [| head -n кількість рядків]	Виведення змісту локального файла або файла HDFS. Якщо файл є архівом, то відбувається його розархівування. Необов'язкова опція, яка записана у дужках, може бути використана для виведення заданої кількості перших рядків файла
Адміністративні команди	
hadoop version	Виведення поточної версії Hadoop
hdfs fsck /	Перевірка стану HDFS
hdfs dfs -df /	Виведення обсягу загального, вільного і використовуваного простору файлової системи

Кириличні імена каталогів та файлів підтримуються.

HDP Sandbox оперує двома файловими системами: локальною і розподіленою, при заданні шляху до каталогу або файлу в першій з них власно шляху повинна передувати так звана схема file://. Так, наприклад команди перегляду змісту каталогів верхнього рівня в локальній і розподіленій файловій системах будуть відповідно виглядати як

hdfs dfs -ls file://

hdfs dfs -ls /

Обидві системи доступні із середовища гостьової операційної системи, якою є Linux-подібна система Cent OS. Файловий обмін між зазначеними системами можливий за допомогою команд HDFS get та put. Для завантаження файлів з зовнішнього ресурсу в локальну файлову структуру може бути використана команда

Linux wget [--no-check-certificate] URL ресурса

при цьому завантаження відбувається в каталог file:///root. Необов'язкова опція, яка записана у дужках, може бути використана якщо при завантаженні виникає помилка, яка обумовлена неможливістю перевірки сертифікату, з якого відбувається завантаження.

В приведеному на рис. 2.6.11 прикладі послідовно здійснюється:

– перегляд змісту кореневого каталогу;

– створення каталогу student;

– надання дозволу на повний доступ до каталогу student;
– створення підкаталогу documents каталогу student;
– повторний перегляд змісту кореневого каталогу;
– перегляд змісту каталогу student.

```
Last login: Fri Jan 10 17:00:43 2025 from 172.18.0.2
[root@sandbox-hdp ~]# hdfs dfs -ls /
Found 12 items
drwxr-xr-x   - root   hdfs            0 2025-01-10 16:56 /anton
drwxrwxrwx   - yarn   hadoop          0 2018-06-18 15:18 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 16:13 /apps
drwxr-xr-x   - yarn   hadoop          0 2018-06-18 14:52 /ats
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 14:52 /hdp
drwx------   - livy   hdfs            0 2018-06-18 15:11 /livy2-recover
y
drwxr-xr-x   - mapred hdfs            0 2018-06-18 14:52 /mapred
drwxrwxrwx   - mapred hadoop          0 2018-06-18 14:52 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 15:59 /ranger
drwxrwxrwx   - spark  hadoop          0 2025-01-10 17:07 /spark2-histor
y
drwxrwxrwx   - hdfs   hdfs            0 2018-06-18 16:06 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 16:08 /user
[root@sandbox-hdp ~]# hdfs dfs -mkdir /student
[root@sandbox-hdp ~]# hdfs dfs -chmod 777 /student
[root@sandbox-hdp ~]# hdfs dfs -mkdir /student/documents
[root@sandbox-hdp ~]# hdfs dfs -ls /
Found 13 items
drwxr-xr-x   - root   hdfs            0 2025-01-10 16:56 /anton
drwxrwxrwx   - yarn   hadoop          0 2018-06-18 15:18 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 16:13 /apps
drwxr-xr-x   - yarn   hadoop          0 2018-06-18 14:52 /ats
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 14:52 /hdp
drwx------   - livy   hdfs            0 2018-06-18 15:11 /livy2-recover
y
drwxr-xr-x   - mapred hdfs            0 2018-06-18 14:52 /mapred
drwxrwxrwx   - mapred hadoop          0 2018-06-18 14:52 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 15:59 /ranger
drwxrwxrwx   - spark  hadoop          0 2025-01-16 11:10 /spark2-histor
y
drwxrwxrwx   - root   hdfs            0 2025-01-16 11:09 /student
drwxrwxrwx   - hdfs   hdfs            0 2018-06-18 16:06 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 16:08 /user
[root@sandbox-hdp ~]# hdfs dfs -ls /student
Found 1 items
drwxr-xr-x   - root hdfs          0 2025-01-16 11:09 /student/documents
[root@sandbox-hdp ~]#
```

Рис. 2.6.11. Приклад роботи з командами файлової системи HDFS

Постановка задачі

Дано

Інсталятор віртуальної машини HDP Sandbox.

Індивідуальна адреса архіву даних у репозиторії машинного навчання.

Потрібно

Інсталювати віртуальну машину HDP Sandbox на персональному комп’ютері.

Провести маніпуляції з каталогами та файлів в локальній і розподіленій файлових системах HDP Sandbox.

Рекомендації щодо виконання роботи

1. Встановити на персональному комп'ютері віртуальну машину HDP Sandbox.
2. Вивести обсяги загального, вільного і використовуваного простору файлової системи.
3. Отримати зміст каталогів верхнього рівня в розподіленій і локальній файловій структурах, порівняти їх.
4. Створити в розподіленій і локальній файловій структурах вкладені каталоги, верхній з яких має назву у вигляді прізвища студента латиницею, а нижній – кирилицею.
5. Надати права повного доступу до створених каталогів.
6. Завантажити до локальної файлової структури архів з репозиторію машинного навчання згідно варіанту завдання (табл. 2.6.2). Переконатися, що каталог file:///root містить завантажений архів і уточнити його ім’я. Перемістити його до створеного каталогу нижнього рівня локальної файлової структури.
7. Вивести перші 5 рядків архіву за допомогою команди text.
8. В створеному каталогу розподіленої файлової структури створити порожній файл який має назву у вигляді імені студента кирилицею.
9. Скопіювати завантажений архів в створений каталог розподіленої файлової структури, а створений файл – у відповідний каталог локальної файлової структури.
10. Рекурсивно вивести зміст створених гілок розподіленої і локальної файлової структур.

Вимоги до звіту

Звіт по роботі повинен містити:

1. Назву дисципліни та лабораторної роботи.
2. Прізвище, ім’я та по батькові студента, код групи.
3. Номер варіанту завдання і адреса архіву даних у репозиторії машинного навчання яка йому відповідає.
4. Знімки зображень екрану, які виникають при встановленні і ініціалізації на персональному комп'ютері віртуальної машини HDP Sandbox з підписами (Рис. 2.6.5 – 2.6.10) та знімки зображень екрану, які виникають при виконанні п. 2, 3, 6, 7, 10 рекомендацій щодо виконання роботи з підписами.
5. Висновки.

Контрольні питання і завдання

1. Назвіть основні складові екосистеми Hadoop.
2. Назвіть і охарактеризуйте компоненти кластеру HDFS.
3. Які основні переваги і недоліки притаманні файловій системі HDFS?
4. Наведіть загальні формати двох версій для команд HDFS.

5. Наведіть приклад будь-якої команди HDFS для каталогу або файлу в локальній і її аналог в розподіленій файловій системах.

6. Назвіть основні групи команд файлової системи HDFS.

Таблиця 2.6.2. Індивідуальні адреси архівів даних у репозиторії машинного навчання

Спільний початок всіх адрес https://archive.ics.uci.edu/static/public/

№	Продовження адреси
1	484/travel+reviews.zip
2	697/predict+students+dropout+and+academic+success.zip
3	397/las+vegas+strip.zip
4	468/online+shoppers+purchasing+intention+dataset.zip
5	519/heart+failure+clinical+records.zip
6	292/wholesale+customers.zip
7	560/seoul+bike+sharing+demand.zip
8	601/ai4i+2020+predictive+maintenance+dataset.zip
9	236/seeds.zip
10	267/banknote+authentication.zip
11	856/higher+education+students+performance+evaluation.zip
12	374/appliances+energy+prediction.zip
13	597/productivity+prediction+of+garment+employees.zip
14	529/early+stage+diabetes+risk+prediction+dataset.zip
15	936/national+poll+on+healthy+aging+(npha).zip
16	863/maternal+health+risk.zip
17	967/phiusiil+phishing+url+dataset.zip
18	1025/turkish+crowdfunding+startups.zip
19	563/iranian+churn+dataset.zip
20	878/cirrhosis+patient+survival+prediction+dataset-1.zip
21	225/ilpd+indian+liver+patient+dataset.zip
22	383/cervical+cancer+risk+factors.zip
23	542/internet+firewall+data.zip
24	857/risk+factor+prediction+of+chronic+kidney+disease.zip
25	571/hcv+data.zip
26	572/taiwanese+bankruptcy+prediction.zip
27	864/room+occupancy+estimation.zip
28	915/differentiated+thyroid+cancer+recurrence.zip
29	244/fertility.zip
30	837/product+classification+and+clustering.zip
31	547/algerian+forest+fires+dataset.zip
32	603/in+vehicle+coupon+recommendation.zip
33	396/sales+transactions+dataset+weekly.zip
34	851/steel+industry+energy+consumption.zip

35	409/daily+demand+forecasting+orders.zip
36	760/multivariate+gait+data.zip
37	849/power+consumption+of+tetouan+city.zip
38	488/facebook+live+sellers+in+thailand.zip
39	591/gender+by+name.zip
40	485/tarvel+review+ratings.zip
41	911/recipe+reviews+and+user+feedback+dataset.zip
42	537/cervical+cancer+behavior+risk.zip
43	211/communities+and+crime+unnormalized.zip
44	386/fma+a+dataset+for+music+analysis.zip
45	498/incident+management+process+enriched+event+log.zip
46	229/skin+segmentation.zip
47	722/naticusdroid+android+permissions+dataset.zip
48	567/covid+19+surveillance.zip
49	713/auction+verification.zip
50	482/parking+birmingham.zip

3. **КРИТЕРІЇ ОЦІНЮВАННЯ**

Навчальні досягнення здобувачів вищої освіти за результатами вивчення курсу оцінюватимуться за шкалою, що наведена нижче:

Рейтингова шкала	Інституційна шкала
90–100	відмінно
74–89	добре
60–73	задовільно
0–59	незадовільно

Здобувачі можуть отримати підсумкову оцінку з навчальної дисципліни на підставі поточного оцінювання знань за умови, якщо набрана кількість балів складатиме не менше як 60 балів.

Максимальне оцінювання становить:

Теоретична частина	Лабораторні роботи	Самостійна робота з підготовки до лабораторних робіт	**Разом**
51	42	7	**100**

Теоретична частина оцінюється за результатами online тесту, який містить 27 запитань.

Критерії оцінювання тесту

Для кожного запитання тесту потрібно вибрати єдину правильну відповідь. Кількість варіантів відповіді в залежності від характеру запитання може складати від 5-ти до 10-ти. Кількість балів, які нараховуються за правильну відповідь залежності від складності запитання може складати від 1-го до 4-х. У разі неправильної відповіді на запитання, здобувач отримує за нього 0 балів. Максимальна кількість балів, яку можна набрати за виконання тесту становить 51.

Критерії оцінювання лабораторної роботи

За кожну лабораторну роботу здобувач може отримати до 6 балів (усього 42 бали), а саме:

6 балів: програма обробки даних правильно функціонує, супроводжується достатньою кількістю коментарів, звіт з роботи оформлений згідно з методичними рекомендаціями і містить повні, інформативні та обґрунтовані висновки.

4 бали: програма обробки даних правильно функціонує, але містить незначні помилки, які суттєво не впливають на отримувані результати, або кількість коментарів недостатня, або звіт з роботи оформлений із незначними відхиленнями від методичних рекомендацій або висновки не повністю

задовольняють вимогам повноти, інформативності та обґрунтованості.

2 бали: програма обробки даних функціонує, але містить помилки, які суттєво впливають на отримувані результати, або коментарі відсутні, або звіт з роботи оформлений із суттєвими відхиленнями від методичних рекомендацій, або висновки не задовольняють вимогам повноти, інформативності та обґрунтованості.

0 балів: програма обробки даних не функціонує, або звіт з роботи відсутній, або не містить висновків, або висновки протирічать фактично отриманим результатам.

У разі отримання позитивної оцінки студент може підвищити її шляхом очної співбесіди з викладачем. При цьому за кожну правильну відповідь на запропоновані запитання нараховується 1 бал.

Критерії оцінювання самостійної роботи

За підготовку вхідних даних згідно з варіантом завдання до кожної лабораторної роботи здобувач отримує:

1 бал (разом до 7 балів).

0 балів: підготовку вхідних даних не виконано, дані розраховані невірно або не відповідають варіанту завдання.

СПИСОК ВИКОРИСТАНОЇ ТА РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Zgurovsky M.Z. Big Data: Conceptual Analysis and Applications. [Текст] / M.Z. Zgurovsky, Y.P. Zaychenko // Springer, 2020, 298 p.
2. Akerkar R. Models of Computation for Big Data [Текст] / R. Akerkar // Springer, 2018, 110 p.
3. Ghavami P. Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics [Текст] / P. Ghavami // CreateSpace Independent Publishing Platform, 2016, 204 p.
4. Feeney K. Engineering Agile Big-Data Systems [Текст] / K. Feeney, J. Davies, J. Welch, S. Hellmann, C. Dirschl, A. Koller, P. Francois, A. Marciniak // River Publishers, 2018, 436 p.
5. Big Data Fundamentals courses [Электрон. ресурс]. Режим доступа: https://cognitiveclass.ai/learn/big-data (дата звернення: 21.10.2024).
6. Big Data Analytics [Электрон. ресурс]. Режим доступа: https://cognitiveclass.ai/learn/analytics/ (дата звернення: 21.10.2024).
7. Aziukovskyi O., Udovyk I., Kozhevnykov A., Powroźnik T. Creating using the mathcad system of laboratory experimentation on the subject «intelligent data analysis» [Текст] / Матеріали XVI міжнародної конференції "Проблеми використання інформаційних технологій у сфері освіти, науки та промисловості" – Дніпро: НТУ «Дніпровська політехніка», 2021. – С. 21–26.
8. Кожевников А.В., Удовик І.М. Створення відкритої нейронної мережі бінарного класифікатора засобами системи Mathcad [Текст] / Матеріали XVI міжнародної конференції "Проблеми використання інформаційних технологій у сфері освіти, науки та промисловості" – Дніпро: НТУ «Дніпровська політехніка», 2021. – С. 27–32.
9. Кожевников А.В., Удовик І.М. Створення відкритої нейронної мережі предиктора лінійного часового ряду засобами системи Mathcad [Текст] / Матеріали XVI міжнародної конференції "Проблеми використання інформаційних технологій у сфері освіти, науки та промисловості" – Дніпро: НТУ «Дніпровська політехніка», 2021. – С. 45–49.
10. UCI Machine Learning Repository [Електрон. ресурс]. Режим доступа: http://archive.ics.uci.edu/ml/index.php/ (дата звернення: 21.10.2024).
11. HDFS Commands Guide – Apache Hadoop 3.4.1 [Электрон. ресурс]. Режим доступа: https://hadoop.apache.org/docs/r3.4.1/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html (дата звернення: 21.10.2024).

ДОДАТОК 1

Таблиця Д.1.1 – Вхідні дані для кореляційного і кластерного аналізу

Індекс варіанти	Діагноз	x1	x2	x3	x4	x5	x6	x7	x8
1	1	1	2	1	1	1	1	1	1
2	1	1	1	2	1	1	1	1	1
3	1	2	1	1	2	1	1	1	1
4	1	1	2	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1
6	1	1	2	1	1	1	1	1	1
7	1	1	2	1	2	1	1	1	1
8	1	2	1	1	1	1	1	1	1
9	1	1	1	1	2	1	1	1	1
10	1	1	4	1	1	1	1	1	1
11	1	1	3	1	1	1	1	1	1
12	1	1	4	1	1	1	1	1	1
13	1	1	2	1	1	1	1	1	1
14	1	1	2	1	1	1	1	1	1
15	1	1	1	2	1	1	1	1	1
16	1	1	2	1	1	1	1	1	1
17	1	1	1	2	1	1	1	2	1
18	1	2	1	1	1	1	1	2	1
19	1	1	3	1	1	2	1	1	1
20	1	1	1	2	1	2	1	1	1
21	1	1	2	1	2	2	1	1	1
22	1	1	2	1	2	2	1	1	1
23	1	2	1	1	2	2	1	1	1
24	1	1	1	1	2	2	1	1	1
25	2	1	4	1	2	1	1	1	2
26	2	1	4	1	1	1	1	2	1
27	2	1	2	1	1	1	1	2	2
28	2	2	3	1	3	1	1	2	2
29	2	2	4	1	1	1	2	1	1
30	2	1	3	1	1	1	2	2	1
31	2	2	3	1	2	1	2	2	2
32	2	2	4	1	2	2	1	1	1
33	2	2	4	1	2	2	1	1	1
34	2	1	3	1	2	2	1	2	2
35	2	2	3	1	2	2	1	2	2
36	2	2	4	1	2	2	1	2	2
37	2	2	4	2	2	2	2	1	1
38	2	2	3	1	1	2	2	1	2
39	2	2	4	2	2	2	2	1	2
40	2	2	3	1	3	2	2	2	1
41	2	2	1	1	1	2	2	2	1
42	2	2	1	2	2	2	2	2	1
43	2	2	3	1	2	2	2	2	1
44	2	1	3	1	1	2	2	2	1
45	2	1	4	2	1	2	2	2	1
46	2	2	3	1	2	2	2	2	1
47	2	1	4	1	2	2	2	2	1
48	2	2	4	1	2	2	2	2	2
49	2	2	3	2	2	2	2	2	2
50	2	2	4	1	2	2	2	2	2
51	3	2	4	2	3	1	1	2	2
52	3	1	4	2	1	1	2	1	2

53	3	2	4	1	2	1	2	1	2
54	3	2	4	1	2	1	2	2	2
55	3	2	4	1	3	1	2	2	2
56	3	2	4	1	3	1	2	2	2
57	3	2	3	1	2	1	2	2	2
58	3	1	2	1	2	2	1	1	2
59	3	2	3	1	2	2	1	2	1
60	3	2	3	1	3	2	1	2	2
61	3	1	4	2	1	2	1	2	2
62	3	2	4	2	3	2	1	2	2
63	3	2	4	2	2	2	2	1	1
64	3	2	3	1	2	2	2	1	1
65	3	1	3	1	1	2	2	1	2
66	3	2	3	1	2	2	2	2	1
67	3	2	4	1	2	2	2	2	1
68	3	2	3	1	3	2	2	2	1
69	3	1	4	2	2	2	2	2	2
70	3	1	2	1	2	2	2	2	2
71	3	1	4	1	1	2	2	2	2
72	3	1	4	2	2	2	2	2	2
73	3	2	4	1	2	2	2	2	2
74	3	2	4	2	2	2	2	2	2
75	3	1	3	2	2	2	2	2	2
76	4	2	3	1	3	1	1	2	2
77	4	2	2	1	3	1	2	1	2
78	4	1	1	2	2	1	2	2	2
79	4	2	3	2	2	1	2	2	2
80	4	2	2	3	1	2	1	2	2
81	4	2	3	2	2	2	1	2	2
82	4	2	1	1	3	2	1	2	2
83	4	2	2	2	2	2	2	1	2
84	4	2	2	2	2	2	2	1	2
85	4	2	3	3	2	2	2	1	2
86	4	2	2	1	2	2	2	1	2
87	4	2	3	2	2	2	2	2	1
88	4	2	1	1	3	2	2	2	1
89	4	2	3	2	3	2	2	2	1
90	4	2	3	1	2	2	2	2	2
91	4	2	3	1	3	2	2	2	2
92	4	2	4	1	3	2	2	2	2
93	4	1	2	2	3	2	2	2	2
94	4	2	3	2	2	2	2	2	2
95	4	2	3	2	2	2	2	2	2
96	4	2	4	2	2	2	2	2	2
97	4	2	2	1	3	2	2	2	2
98	4	2	3	2	3	2	2	2	2
99	4	2	3	3	2	2	2	2	2
100	4	2	3	1	2	2	2	2	2
101	4	2	3	2	3	2	2	2	2
102	4	2	3	1	3	2	2	2	2
103	4	2	3	1	2	2	2	2	2

ДОДАТОК 2

Фрагмент програми перетворення вибірки, в якій значення ознак представлені в звичайній порядковій шкалі в вибірку, в якій значення ознак представлені рангами

B - вибірка, в якій значення ознак представлені в звичайній порядковій шкалі в вибірку,
BR - вибірка, в якій значення ознак представлені рангами
a, b - мінімальне та максимальне значення ознаки в звичайних порядкових одиницях,
c - матриця значень, які можуть приймати ознаки

$$j := 1..\,cols(B)$$
$$a_j := min\left(B^{\langle j \rangle}\right)$$
$$b_j := max\left(B^{\langle j \rangle}\right)$$
$$i := 1..\,max(b - a) + 1$$
$$c_{i,j} := if\left(a_j + i - 1 \le b_j, a_j + i - 1, 0\right)$$

$$c = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 0 & 3 & 3 & 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

K, KN - матриці частот та накоплених частот прийняття ознакою значень, що визначаються матрицею c

$$K_{i,j} := \sum_{k=1}^{rows(B)} if\left(a_j + i - 1 \le b_j, if\left(B_{k,j} = c_{i,j}, 1, 0\right), 0\right)$$

$$K = \begin{pmatrix} 5 & 17 & 4 & 26 & 9 & 15 & 12 & 10 & 11 \\ 12 & 27 & 8 & 16 & 25 & 29 & 32 & 34 & 33 \\ 11 & 0 & 15 & 2 & 10 & 0 & 0 & 0 & 0 \\ 16 & 0 & 17 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$KN_{i,j} := if\left(i \le b_j, \sum_{p=1}^{i} K_{p,j}, 0\right)$$

$$KN = \begin{pmatrix} 5 & 17 & 4 & 26 & 9 & 15 & 12 & 10 & 11 \\ 17 & 44 & 12 & 42 & 34 & 44 & 44 & 44 & 44 \\ 28 & 0 & 27 & 44 & 44 & 0 & 0 & 0 & 0 \\ 44 & 0 & 44 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

RG - матриця рангів, яка відповідає значенням ознак в матриці c і виражених

у відносних одиницях

$$RG_{i,j} := \text{if}\left(i \le b_j, \text{if}\left(i = 1, \frac{KN_{i,j} + 1}{2}, \frac{KN_{i,j} + KN_{i-1,j} + 1}{2}\right), 0\right)$$

$$RG = \begin{pmatrix} 3 & 9 & 2.5 & 13.5 & 5 & 8 & 6.5 & 5.5 & 6 \\ 11.5 & 31 & 8.5 & 34.5 & 22 & 30 & 28.5 & 27.5 & 28 \\ 23 & 0 & 20 & 43.5 & 39.5 & 0 & 0 & 0 & 0 \\ 36.5 & 0 & 36 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$BR(B) := \left| \begin{array}{l} \text{for } i \in 1..\text{rows}(B) \\ \quad \text{for } j \in 1..\text{cols}(B) \\ \quad\quad \text{for } k \in 1..\text{rows}(c) \\ \quad\quad\quad BR_{i,j} \leftarrow RG_{k,j} \text{ if } B_{i,j} = c_{k,j} \\ BR \end{array} \right.$$

$$BR := BR(B)$$

ДОДАТОК З

Вхідні дані для регресійного аналізу

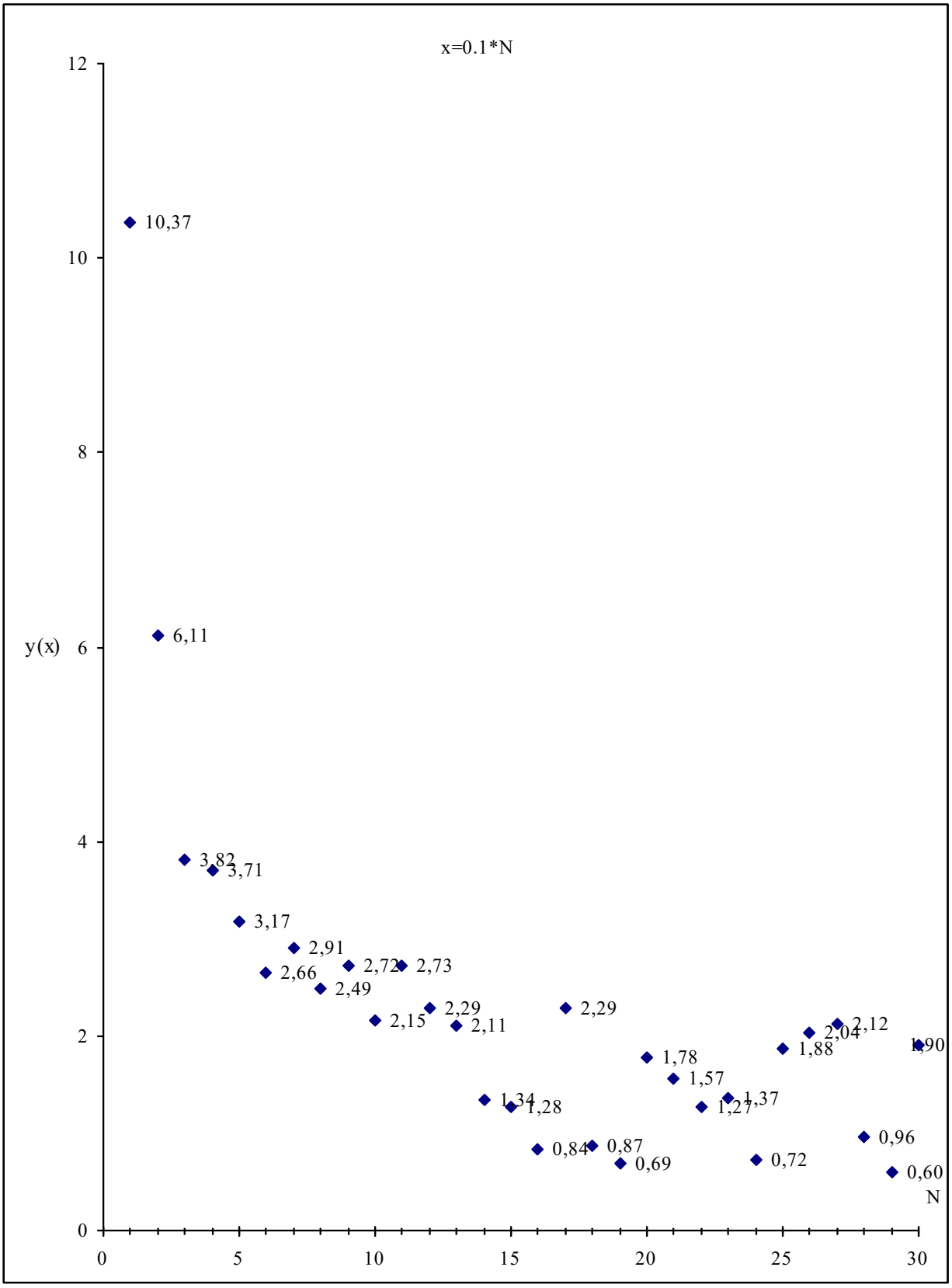

Рис. Д.2.1 Дані для гіперболічної регресії. Група 1

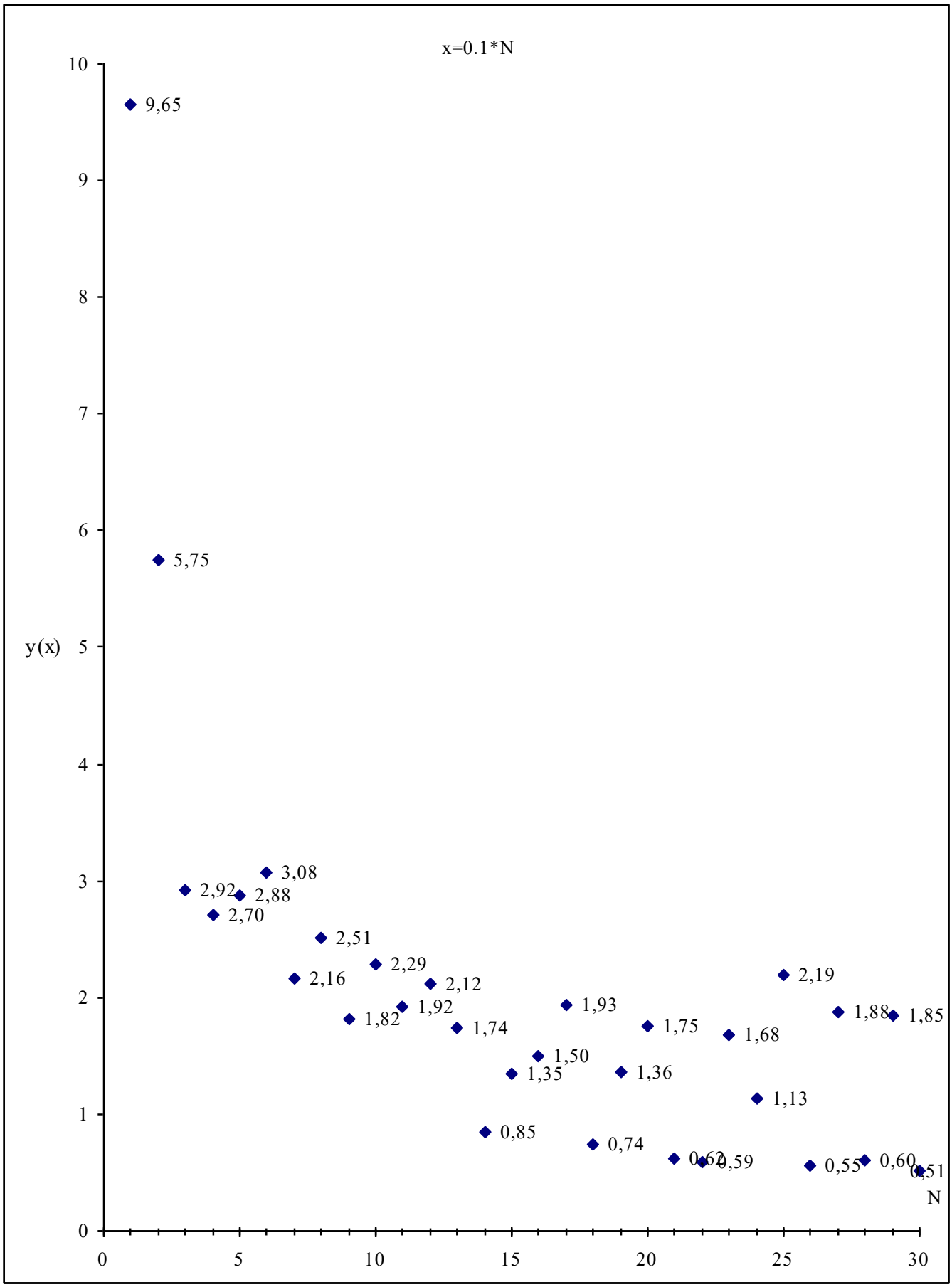

Рис. Д.2.2 Дані для гіперболічної регресії. Група 2

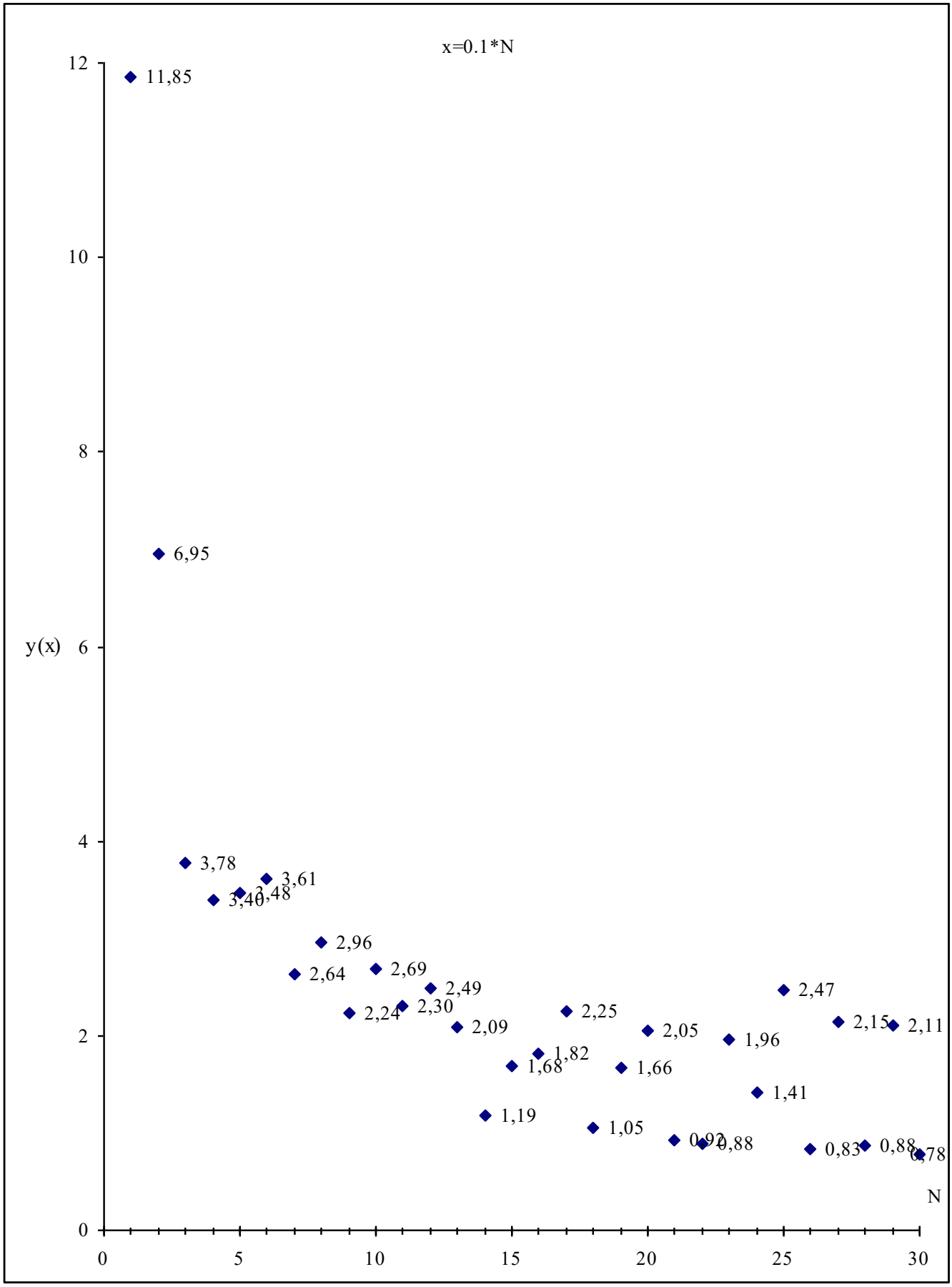

Рис. Д.2.3 Дані для гіперболічної регресії. Група 3

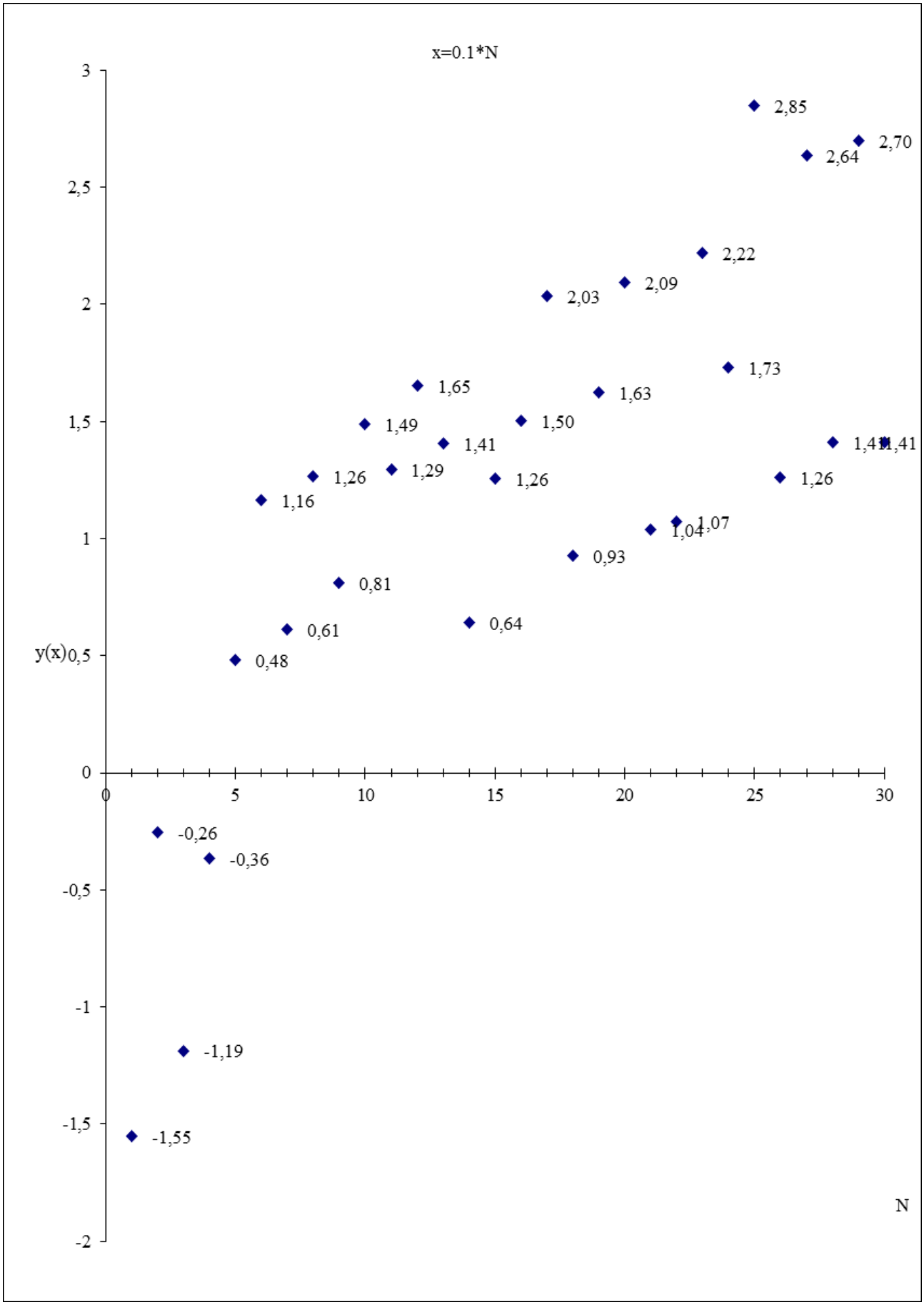

Рис. Д.2.4 Дані для логарифмічної регресії. Група 1

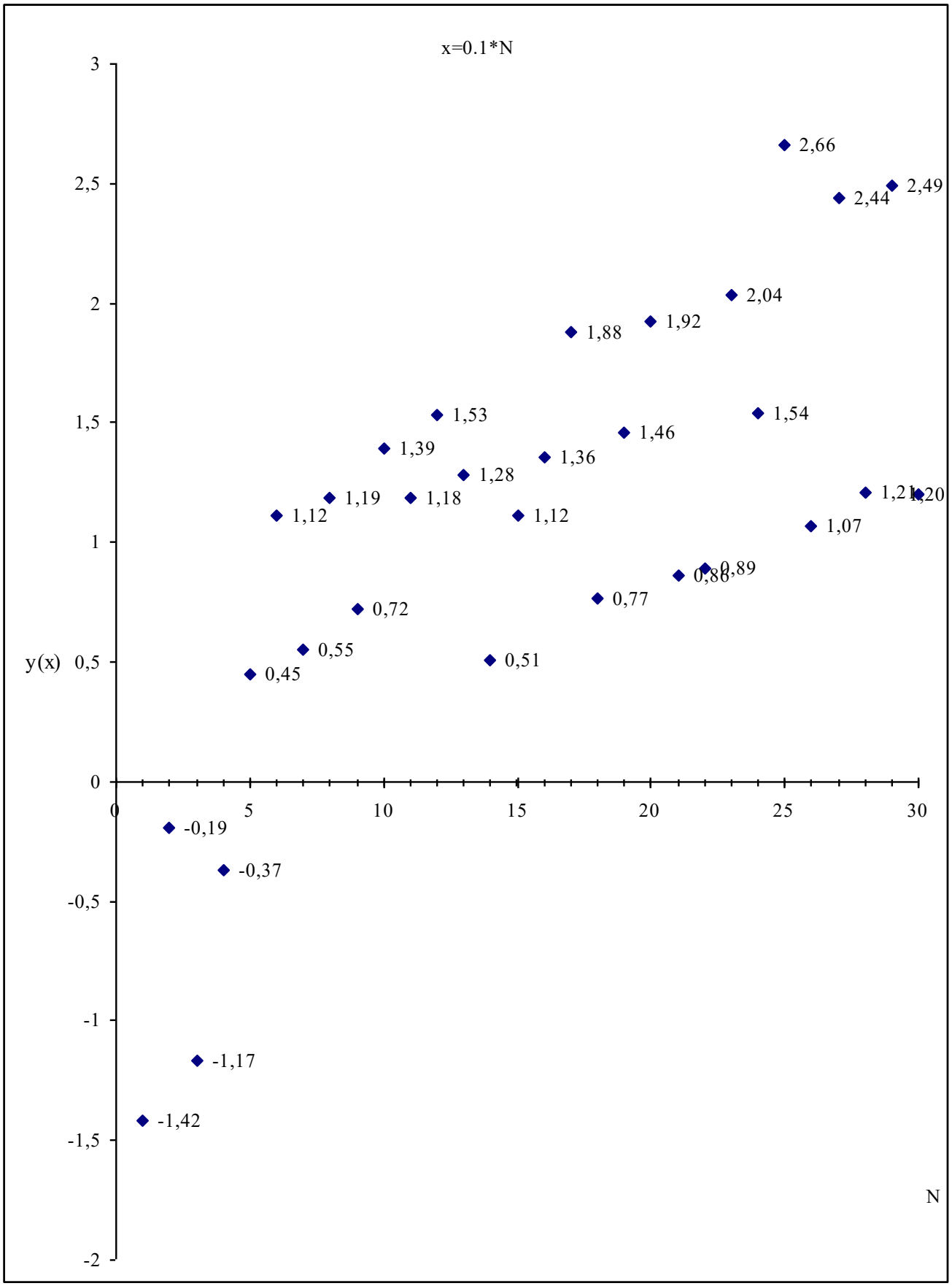

Рис. Д.2.5 Дані для логарифмічної регресії. Група 2

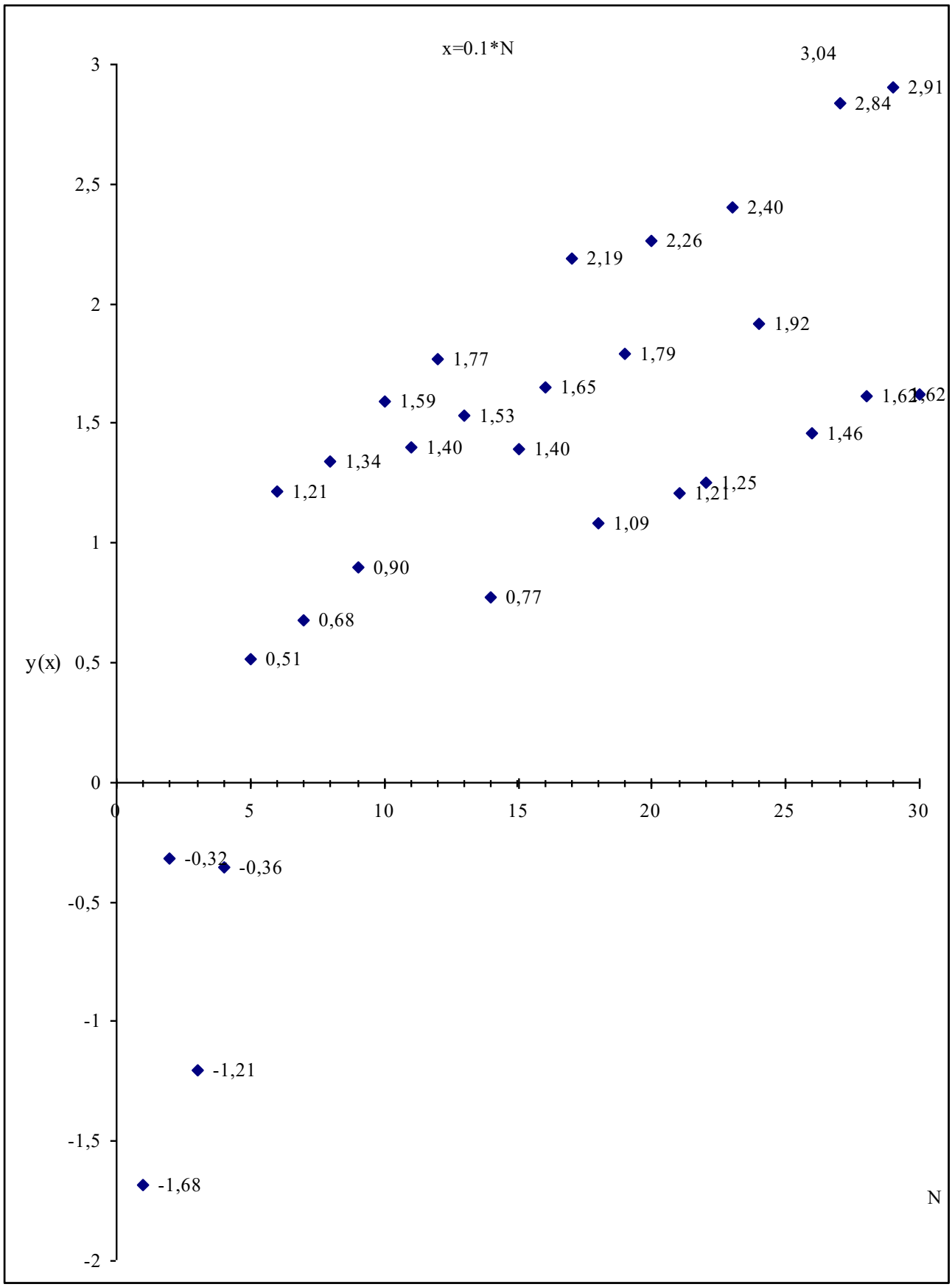

Рис. Д.2.6 Дані для логарифмічної регресії. Група 3

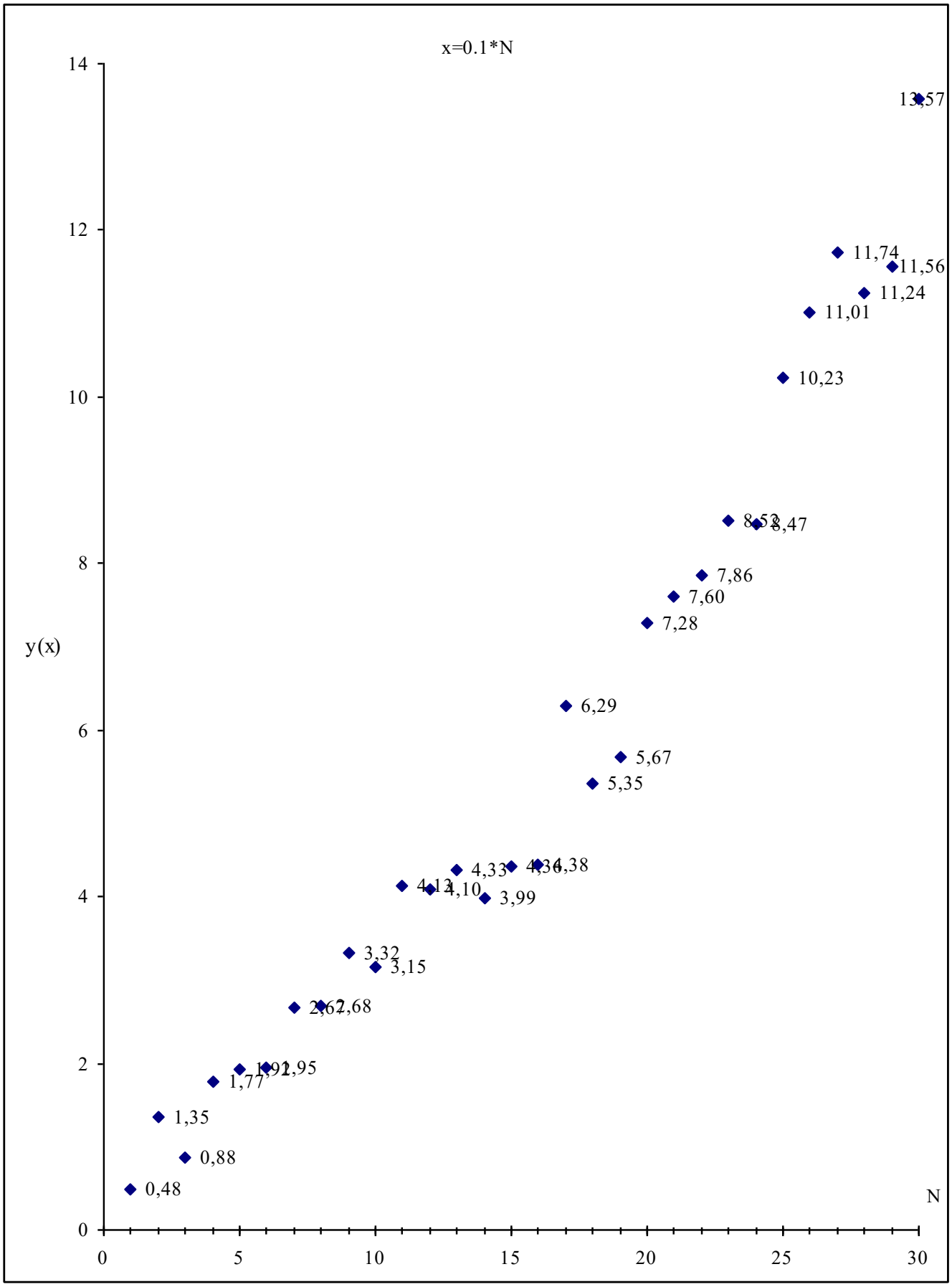

Рис Д.2.7 Дані для квадратичної регресії. Група 1

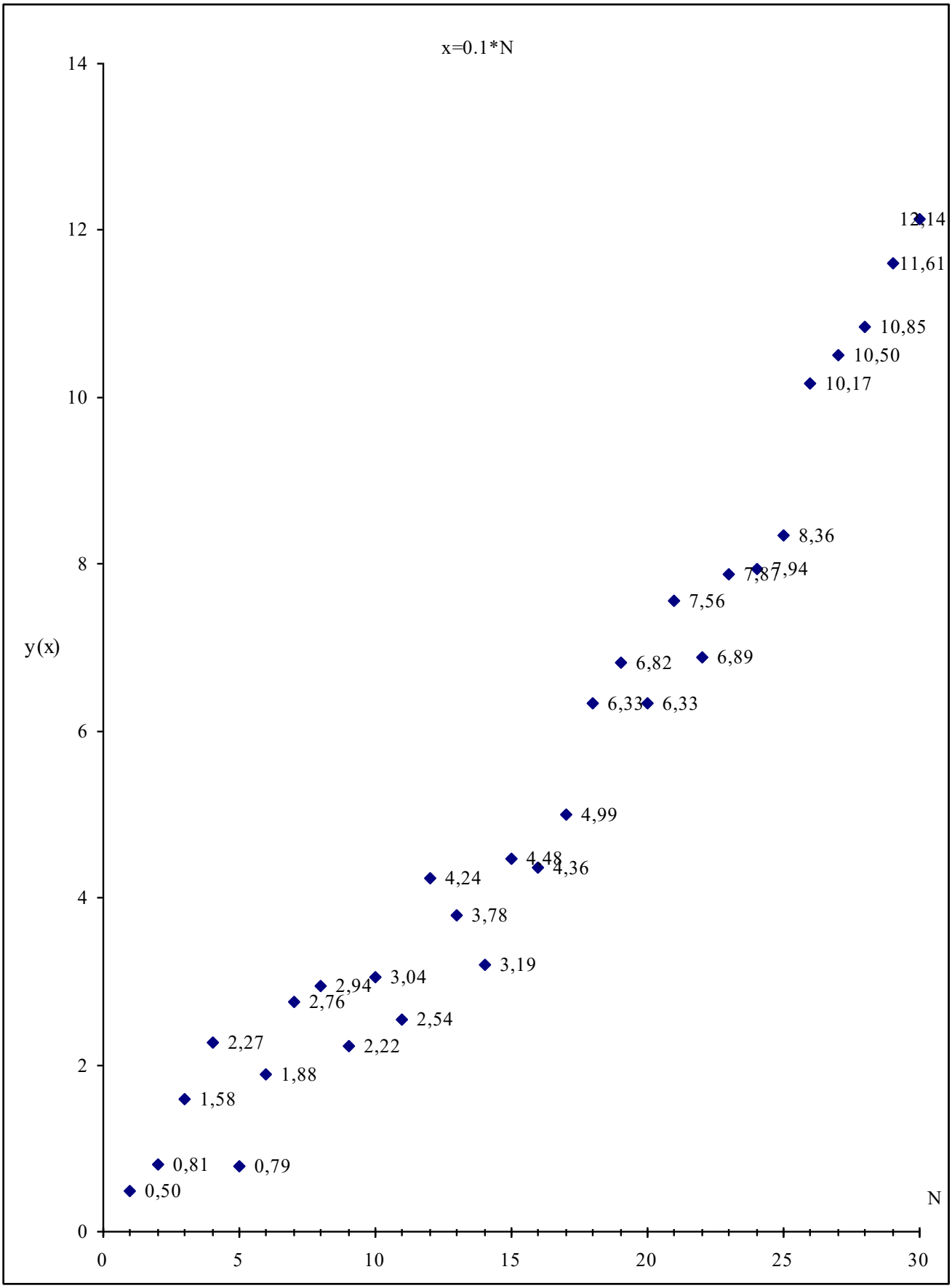

Рис Д.2.8 Дані для квадратичної регресії. Група 2

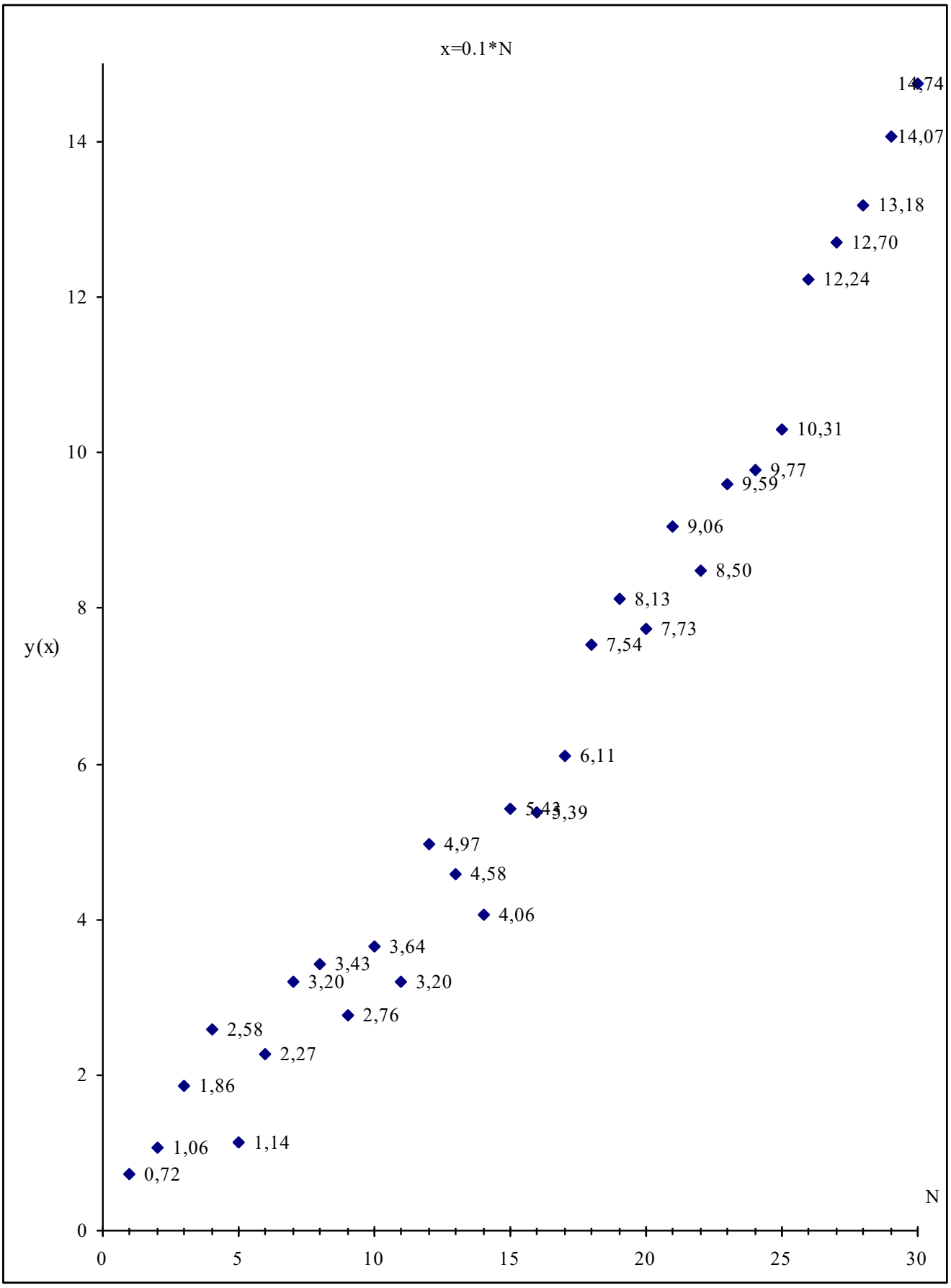

Рис Д.2.9 Дані для квадратичної регресії. Група 3

ДОДАТОК 4

Програма визначення номерів варіант вибірки для проведення регресійного аналізу за номером варіанта студента

NC - номер студента за журналом групи, NG - номер групи у потоці

$NC := 2$

$NG := 1$

$A_0 := \mathrm{mod}(NC + NG, 3)$

$A_1 := \mathrm{mod}\left(\mathrm{floor}\left(\frac{NC + NG}{3}\right), 3\right)$

$A_2 := \mathrm{mod}\left(\mathrm{floor}\left(\frac{\mathrm{floor}\left(\frac{NC + NG}{3}\right)}{3}\right), 3\right)$

$j := 0..9$

$$A = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}$$

$B_j := 3 \cdot j + A_0 + 1$

$B_{j+1} := 3 \cdot j + A_1 + 4$

$B_{j+2} := 3 \cdot j + A_2 + 7$

$Varianta := \mathrm{submatrix}(B, 0, 9, 0, 0)^T$

$Varianta =$

	0	1	2	3	4	5	6	7	8	9
0	1	5	7	10	13	16	19	22	25	28

ДОДАТОК 5

Програмні блоки кластеризації ітеративними методами групування

A - стандартизована матриця координат об'єктів в просторі ознак
kmax - число кластерів
n - число об'єктів
m - число ознак (симптомів за якими ставиться діагноз)
b - вектор розподілу об'єктів за кластерами на попередньому кроці ітерацій
b1 - на вході в програмний блок - випадковий вектор початкового розподілу об'єктів за кластерами; в програмному блоці - вектор розподілу об'єктів за кластерами на поточному кроці ітерацій
M, M1 - матриці приналежності об'єктів кластерам на попередньому та поточному кроках ітерацій
w - експонентна вага
e - похибка між нормою різниці матриць приналежностей на послідовних кроках при досягненні якої ітерації припиняються
count - кількість ітерацій
centr - матриця центроідів кластерів, в якій перший стовбець - це кількість об'єктів в кластері (для методу K-means), а решта стовбців - координати центроіда в просторі ознак
S - норма різниці матриць приналежностей на останньому та передостанньому кроках ітерацій

Програмний блок кластеризації методом K-means

$KMEANS(A, kmax, n, m, b1) :=$

$count \leftarrow 0$

$S \leftarrow 1$

"The condition for the end of iterations is the invariability"

"of the distribution of objects by clusters at successive steps"

while $S > 0$

$count \leftarrow count + 1$

$S \leftarrow 0$

$b \leftarrow b1$

for $k \in 1..kmax$

for $j \in 1..m$

$centr_{k,j} \leftarrow 0$

"Clusters centroids coordinates determining"

for $i \in 1..n$

for $k \in 1..kmax$

if $b_i = k$

$centr_{k,1} \leftarrow centr_{k,1} + 1$

for $j \in 2..m$

$centr_{k,j} \leftarrow centr_{k,j} + A_{i,j-1}$

for $k \in 1..kmax$

for $j \in 2..m$

$centr_{k,j} \leftarrow \frac{centr_{k,j}}{centr_{k,1}}$

"Determining the new belonging of objects to clusters"

for $i \in 1..n$

$b1_i \leftarrow 1$

$rmin \leftarrow \sqrt{\sum_{j=3}^{m} \left(centr_{1,j} - A_{i,j-1}\right)^2}$

for $k \in 2..kmax$

if $rmin > \sqrt{\sum_{j=3}^{m} \left(centr_{k,j} - A_{i,j-1}\right)^2}$

$rmin \leftarrow \sqrt{\sum_{j=3}^{m} \left(centr_{k,j} - A_{i,j-1}\right)^2}$

$b1_i \leftarrow k$

"Determination of the distance between the vectors of objects"

"belonging to clusters at successive steps of iterations"

$S \leftarrow \sqrt{\sum_{i=1}^{n} \left(b1_i - b_i\right)^2}$

$\begin{pmatrix} augment(b, b1) \\ centr \\ count \end{pmatrix}$

Програмний блок кластеризації методом Fuzzy C-means

$$
\begin{aligned}
&\mathrm{CMEANS}(A, kmax, n, m, M1, w, \varepsilon) := \\
&\quad \mathrm{count} \leftarrow 0 \\
&\quad S \leftarrow 1 \\
&\quad \text{while } S > \varepsilon \\
&\qquad \mathrm{count} \leftarrow \mathrm{count} + 1 \\
&\qquad S \leftarrow 0 \\
&\qquad M \leftarrow M1 \\
&\qquad \text{for } k \in 1..kmax \\
&\qquad\quad \text{for } j \in 1..m \\
&\qquad\qquad \mathrm{centr}_{k,j} \leftarrow 0 \\
&\qquad \text{for } k \in 1..kmax \\
&\qquad\quad \text{for } j \in 1..m \\
&\qquad\qquad \mathrm{centr}_{k,j} \leftarrow \dfrac{\sum_{i=1}^{n} \left(M_{i,k}\right)^{w} \cdot A_{i,j}}{\sum_{i=1}^{n} \left(M_{i,k}\right)^{w}} \\
&\qquad \text{for } i \in 1..n \\
&\qquad\quad \text{for } k \in 1..kmax \\
&\qquad\qquad r_{k,i} \leftarrow \sqrt{\sum_{j=2}^{m} \left(\mathrm{centr}_{k,j} - A_{i,j}\right)^2} \\
&\qquad\quad \text{for } k \in 1..kmax \\
&\qquad\qquad M1_{i,k} \leftarrow \dfrac{1}{\left(r_{k,i}\right)^{\frac{2}{w-1}} \cdot \sum_{p=1}^{kmax} \dfrac{1}{\left(r_{p,i}\right)^{\frac{2}{w-1}}}} \quad \text{if } r_{k,i} \neq 0 \\
&\qquad\qquad M1_{i,k} \leftarrow 1 \quad \text{otherwise} \\
&\qquad S \leftarrow \sqrt{\sum_{k=1}^{kmax} \sum_{i=1}^{n} \left(M1_{i,k} - M_{i,k}\right)^2} \\
&\quad \begin{pmatrix} M1 \\ \mathrm{centr} \\ \mathrm{count} \\ S \end{pmatrix}
\end{aligned}
$$

ДОДАТОК 6

Програма кластеризації даних ієрархічними агломеративними методами з побудуванням дендрограми

Кластеризація здійснюється за всіма факторними ознаками всіх 103-х варіант таблиці

```
# Імпорт модуля pyplot бібліотеки matplotlib та бібліотек numpy,  pandas
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
# Імпорт даних з файлу Excel
xls = pd.ExcelFile('c:\work\standard.xls')
df = pd.read_excel(xls, 'Standard', header=0)
data=np.array(df)
# Для кластеризації обираються дані всіх 103 рядки таблиці
X=data[0:102]
# Імпорт з бібліотеки SciPy функції побудування дендрограми dendrogram и функцій
# кластеризации різними методами: Уорда, повного та середнього зв'язку -
# ward, complete, average
from scipy.cluster.hierarchy import dendrogram, ward, complete, average
# Функція complete повертає масив зв'язків з відстанями,
# обчисленими з використанням евклідової метрикою в ході виконання
# агломеративної кластеризації методом повного зв'язку
linkage_array = complete(X)
# Побудування дендрограми
dendrogram(linkage_array)
# Позначки на дереві для відстаней, що відповідають чотирьом кластерам
ax = plt.gca()
```

```
        bounds = ax.get_xbound()
        ax.plot(bounds, [5.65, 5.65], '--', c='k')
        ax.text(bounds[1], 5.65, ' Чотири кластери', va='center', fontdict={'size':
15})
        # Підписи на осях
        plt.xlabel("Індекс спостереження")
        plt.ylabel("Міжкластерна відстань")
```

ДОДАТОК 7

Brief researchs abstracts

Research № 1, 2.

*Name:*Probabilistic distributions of discrete and continuous random variables and point estimates of their parameters.

Input data:

– names and formulas for differential theoretical distributions of discrete and continuous random variables;

– formulas: calculation of distribution parameters depending on the option number, mathematical expectations and variances of distributions;

– names of the MathCAD functions used.

Brief content of research:

– determination of the effective range of variation of a random value;

– generation of a random values sample with a given probability distribution;

– plotting of theoretical and empirical differential probability distributions of a random value;

– plotting of theoretical and empirical cumulative probability distributions of a random value;

– determination of random value parameter estimates: mean, variance, asymmetry, excess mode and median of a random variable using built-in MathCAD, according to sample data and empirical distribution;

– plotting of the dependence mean and statistical variance of a random value on the volume of the sample.

Typical conclusions:

– polygons (plots) of differential theoretical probability distributions of a random value, constructed using the distribution formulas and the built-in MathCAD function matches;

– polygons (plots) of differential and cumulative theoretical probability distributions of random value matches with the corresponding polygons (histograms) of empirical distributions within the statistical error;

– the values of random value statistical parameters (mean, variance, asymmetry, excess, and median), calculated using the built-in MathCAD functions and calculated based on sample data matches, and calculated based on empirical distribution differ slightly from them;

– as the sample size increases, the values of the statistical parameters of the random value (mean and variance) are directed, respectively, to the values of the mathematical expectation and variance of the general population.

Research № 3.

Name: Special functions of mathematical statistics, interval estimates of random values parameters and tests to verification statistical hypotheses

Input data are the data sets for research variations № 1, 2.

Brief content of research:

– plotting of the dependencies of the confidence interval of the mathematical expectation on the confidence probability using the inverse cumulative functions of the normal distribution and the Student's distribution;

– generation array of random value samples with given probability distributions;

– determining the estimations of the confidence probabilities of a random value mathematical expectation and the variance confidence intervals using the random trials method. Comparison of the obtained estimations with the confidence probability for which confidence intervals were calculated;

– calculating mean and variance for the sum large number of identically distributed random values;

– plotting of theoretical and empirical differential probability distributions for the mean large number of identically distributed random values;

– compliance check between theoretical and empirical differential probability distributions for the mean large number of identically distributed random values using by χ^2 criterion.

Typical conclusions:

– confidence interval increases with increasing confidence probability and decreases with increasing sample size;

– with a known variance of the general population and other identical conditions, the values of the confidence interval for the mathematical expectation are less than if the variance is calculated from the sample;

– the probabilities estimation for falling in the confidence intervals mathematical expectation and variance determined by the method of statistical trials differ slightly from the confidence probabilities;

– the mean and statistical variance of the empirical probability distributions for the mean large number of identically distributed random values coincide, respectively, with the mathematical expectation and variance of the theoretical distribution;

– the empirical distribution for the mean large number of identically distributed random values visually corresponds to the normal distribution. The χ^2 test also shows that the empirical distribution for the mean large number of identically distributed random values corresponds to the normal distribution.

Research № 4.

Name: Correlation analysis.

Initial data is a subsample of an appendicitis diagnostics results sample. The sample is a data of observation matrices which presented in the ordinal scale. Matrices are given in two forms: with ordinary ordinal data and with rank data. Each row of the matrix corresponds to the patient. The columns of the matrices contain the codes of the resultant sign – the clinically confirmed absence of appendicitis or its morphological form and 8 independent signs – the values of the diagnostic attributes.

Brief content of research:

– formation of a data subsample from a sample in accordance with the option;

– calculating matrices of pairwise Pearson's correlation coefficients and vectors of multiple correlation coefficients for ordinary ordinal data and for rank data;

– checking the significance of the minimum in modulus paired and minimum multiple correlation coefficients;

– calculating matrices of Spearman's rank correlation coefficients and Kendall's rank correlation coefficients for ordinary ordinal data and for rank data with and without tied ranks correction.

Typical conclusions:

– obtained correlation indices vary within the relevant ranges of values: Pearson's coefficient and Spearman's and Kendall's coefficients with and without tied ranks corrections from -1 to 1. The multiple correlation coefficient varies from 0 to 1;

– all correlation matrices are symmetric, on their main diagonals are ones;

– the resultant sign has high correlation coefficients with independent signs, which indicates the correct diagnostic (independent) attributes choice;

– Pearson's coefficients have similar values for the input data presented in ordinary ordinal and rank units. The Kendel's coefficients give the same values in these cases. To calculate the Spearman's coefficients, only rank values can be used;

– Spearman's coefficients, with/without corrections for tied ranks, differ significantly due to the large number of tied ranks. The values of the Pearson's and the Spearman's coefficients with corrections for tied ranks coincide. The calculation of Kendall's coefficients is possible only with the corrections for tied ranks;

– the minimum in modulus paired and minimum multiple correlation coefficients are not significant.

Research № 5.

Name: Regression analysis.

Initial data are a nonlinear regression model and a subsample of the quantitative data sample. Each item, of the sample, contain resultant and one independent sign.

Brief content of research:

– formation of a data subsample from a sample in accordance with the option;

– composing of normal equations systems with respect to unknown coefficients estimates for linear and nonlinear regression;

– calculation of coefficients estimates for linear and nonlinear pairs regression by solving of normal equations systems;

– calculation of values for total, regression and residual variances and coefficients of determination for linear and nonlinear regression;

– plotting in a common field of the subsample initial data, graphs of linear and nonlinear regression;

– checking the significance of the linear regression equation and its coefficients.

Typical conclusions:

– the result of plotting in a common field of the subsample initial data, graphs of linear and nonlinear regression shows good match between them. Also

graphics shows that nonlinear regression is better than linear to approximation of initial data;

– better quality of nonlinear regression is confirmed by the results of calculations which show that nonlinear regression has a lower residual variance and a higher coefficient of determination;

– the linear regression equation is statistically significant, as are both of its coefficients.

Research № 6.

Name: Cluster analysis.

Initial data is the data set for research variation № 4.

Brief content of research:

– formation of a data subsample from a sample in accordance with the option;

– clustering of the initial data by the method K-means;

– calculation of coefficients of pair correlation of diagnostic (independent) attributes with the result – the true diagnosis;

– determination of the cluster number corresponding to an unconfirmed diagnosis of appendicitis. For this cluster, the diagnostic attribute, most correlated with the resultant one, has the minimum value;

– calculation of probabilities estimates for 1: (a healthy patient belongs to one of the clusters with diagnosed appendicitis) and 2: (a patient who has appendicitis belongs to a cluster with an unconfirmed diagnosis) kind errors;

– clustering of the initial data by the method fuzzy C-means;

– determination of the cluster number corresponding to an unconfirmed diagnosis of appendicitis;

– calculation of probabilities estimates for 1: (a healthy patient belongs to one of the clusters with diagnosed appendicitis) and 2: (a patient who has appendicitis belongs to a cluster with an unconfirmed diagnosis) kind errors;

– plotting the dependence of the average variance of the cluster's number on the exponential weighting coefficient.

Typical conclusions:

– clustering of the patients diagnosis results with suspected appendicitis, which were performed by the methods of K-means and Fuzzy C-means (for exponential weighting coefficient w=1,5) show that the chosen diagnostic attributes system allows to clearly distinguish clinically unconfirmed cases and cases of appendicitis;

– the probabilities estimates for the 1-st kind errors is equal zero and the 2-nd kind consist of no more than 0,1;

– the dependence of the average variance of the cluster's number on the exponential weighting coefficient in the range of argument values (1.1. – 2.0) has monotonically increasing character.

Research № 7.

Name: Introduction to the Apache Hadoop Sandbox and the HDFS file system.

Initial data are HDP Sandbox virtual machine installer and custom data archive address in the Machine Learning Repository.

Brief content of research:

– installing and initializing the HDP Sandbox virtual machine;

– manipulation with directories and files in the local and distributed file systems of the HDP Sandbox by various shell HDFS commands .

Typical conclusions:

– installing and initializing the HDP Sandbox virtual machine requires significant hardware resources;

– HDP Sandbox operates on two file systems: local and distributed (HDFS);

– File system resources can be managed using shell commands from the command line;

– Both systems are accessible from the guest operating system environment, which is a Linux-like Cent OS system;

– File sharing between these systems is possible, and files can be loaded from an external resource into the local file structure;

– HDFS file system commands have a common format with a DOS operating system commands;

– the functionality of the HDFS file system commands is similar to that of the DOS operating system commands.

ЗМІСТ

1. ЗАГАЛЬНІ ПОЛОЖЕННЯ 3
2. ПОСТАНОВКА ЗАДАЧ ЛАБОРАТОРНИХ РОБІТ, МЕТОДИЧНІ РЕКОМЕНДАЦІЇ ДЛЯ ЇХ ВИКОНАННЯ ТА ІНДИВІДУАЛЬНІ ЗАВДАННЯ 5
2.1. Лабораторні роботи № 1, 2
Імовірнісні розподіли дискретних і безперервних випадкових величин і точкові оцінки їх параметрів 5
2.2. Лабораторна робота № 3
Спеціальні функції математичної статистики, інтервальні оцінки параметрів випадкових величин і тести для перевірки статистичних гіпотез 13
2.3. Лабораторна робота № 4
Кореляційний аналіз 18
2.4. Лабораторна робота № 5
Регресійний аналіз 22
2.5. Лабораторна робота № 6
Кластерний аналіз 30
2.6. Лабораторна робота № 7
Пісочниця Apache Hadoop і команди файлової системи HDFS 43
3. КРИТЕРІЇ ОЦІНЮВАННЯ 61
СПИСОК ВИКОРИСТАНОЇ ТА РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ 63
Додаток 1. Вхідні дані для кореляційного і кластерного аналізу 64
Додаток 2. Фрагмент програми перетворення вибірки, в якій значення ознак представлені в звичайній порядковій шкалі в вибірку, в якій значення ознак представлені рангами 66
Додаток 3. Вхідні дані для регресійного аналізу 68
Додаток 4. Програма визначення номерів варіант вибірки для проведення регресійного аналізу 77
Додаток 5. Програмні блоки кластеризації ітеративними методами групування 78
Додаток 6. Програма кластеризації даних ієрархічними агломеративними методами з побудуванням дендрограми 81
Додаток 7. Brief researchs abstracts 83

Навчальне видання

Кожевников Антон Вячеславович

МЕТОДИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ (BIG DATA)

Методичні рекомендації до виконання лабораторних робіт
для здобувачів ступеня бакалавра
освітньо-професійної програми «Інформаційні системи та технології»
зі спеціальності 126 Інформаційні системи та технології

Видано в авторській редакції

Електронний ресурс.
Підписано до видання 18.02.2025. Авт. арк. 6,6.

Національний технічний університет «Дніпровська політехніка».
49005, м. Дніпро, просп. Дмитра Яворницького, 19.