ЛАБОРАТОРНА РОБОТА № 6
ПРОГНОЗУВАННЯ, ЗАДАЧА РЕГРЕСІЇ

Мета роботи

На практиці вивчити роботу алгоритмів, що вирішують задачу регресії, і навчитися інтерпретувати результати їх роботи.

Основні теоретичні відомості

У роботі розглядаються такі методи (у дужках наведено назву в WEKA):

– лінійна регресія (functions.LinearRegression);
– регресійні дерева і модельні дерева (trees.M5P);
– метод опорних векторів, модифікований для вирішення задач регресії (functions.SMOreg);
– метод найближчих сусідів (lazy.IBk).

Параметри налаштування алгоритмів

Розглянемо параметри налаштування використовуваних алгоритмів у WEKA (табл. 3.1).

Таблиця 3.1 – Параметри налаштування методів

Метод	**Параметри**
LinearRegression	*attributeSelectionMethod* – метод відбору атрибутів. *eliminateColinearAttributes* – виключити колінеарні атрибути. *ridge* – штраф за великі значення коефіцієнтів регресії (регуляризація Тихонова).
M5P	*buildRegressionTree* –регресійне або модельне дерево. *minNumInstances* – мінімальна кількість примірників у листі. *saveInstances* – зберігати примірники у вузлах для візуалізації. *unpruned* – будувати дерево без відсікання *useUnsmoothed* – незгладжене прогнозування.
SMOreg	Основні параметри можна знайти в табл. 2.1. *regOptimizer* – алгоритм навчання.

Методи оцінки якості прогнозування

Наведені в другій роботі параметри більш корисні для опису задач класифікації ніж для завдань регресії. Для задачі регресії помилки прогнозування не просто присутні або відсутні, а мають різні числові значенні. Для оцінки успішності числових прогнозів можуть бути використані альтернативні міри, деякі з яких наведено в табл. 3.2.

Таблиця 3.2 – Міри оцінки якості вирішення задачі регресії

Параметр	**Формула для расчета**
Середній квадрат помилки (mean-squared error)	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
Середньоквадратична помилка (root mean-squared error)	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
Середня абсолютна помилка (mean-absolute error)	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
Відносний квадрат помилки (relative-squared error)*	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\overline{a})^2+\ldots+(a_n-\overline{a})^2}$
Root relative-squared error*	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\overline{a})^2+\ldots+(a_n-\overline{a})^2}}$
Relative-absolute error*	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\overline{a}\rvert+\ldots+\lvert a_n-\overline{a}\rvert}$
Коефіцієнт кореляції (correlation coefficient)**	$\frac{S_{PA}}{\sqrt{S_P S_A}}$,
$S_{PA}=\frac{\sum_i (p_i-\overline{p})(a_i-\overline{a})}{n-1}$, $S_P=\frac{\sum_i (p_i-\overline{p})^2}{n-1}$, $S_A=\frac{\sum_i (a_i-\overline{a})^2}{n-1}$	

де p_1 , p_2 , …, p_n – прогнозовані значення для цільового атрибута тестової вибірки;

a_1, a_2 , …, a_n – реальні значення цільового атрибута;

$\overline{a}$ – середнє арифметичне (* – навчальної вибірки, ** – тестової).

Завдання на лабораторну роботу

1. Оберіть в таблиці Б.2 два набори даних. Виконайте для них наступні завдання.
2. Завантажте дані та за необхідності виконайте попередню обробку даних.
3. Вирішіть задачу регресії за допомогою наступних методів:
   - Linear regression;
   - SMOreg;
   - M5P (model trees and regression trees) з наступними параметрами налаштування:
     - build regression tree: True, unpruned: True, useUnsmoothed: True;
     - build regression tree: True, unpruned: False, useUnsmoothed: True;
     - build regression tree: False, unpruned: True, useUnsmoothed: True;
     - build regression tree: False, unpruned: False, useUnsmoothed: True;
   - kNN.
4. Запишіть отримані моделі і порівняйте їхню ефективність (точність передбачення).
5. Наведіть результати прогнозування для 5 довільних екземплярів.
6. Які з атрибутів є найбільш значущими для передбачення значень цільового атрибуту, судячи з побудованих моделей? Чому? Як зміниться точність передбачення, якщо залишити лише значущі атрибути?

Контрольні питання

1. У чому полягає задача регресії? Наведіть практичний приклад?
2. Чим задача регресії схожа і чим відрізняється від задачі класифікації?
3. Що таке навчання з учителем і без учителя? До якого типу належить завдання регресії?
4. Задача регресії є описовою або прогнозуючою і чому?
5. Опишіть один з розглянутих методів, що вирішують завдання регресії.

6. Як оцінити якість побудованої моделі для завдання регресії?

Зміст звіту

1. Тема і мета роботи.
2. Завдання до роботи.
3. Результати виконання завдань.
4. Відповіді на контрольні запитання.
5. Висновки, що відображують результати виконання роботи та їх критичний аналіз.

Додаток Б
Варіанти індивідуальних завдань

Обрати з таблиці за номером варіанту (N) з журналу набор даних для дослідження. Дослідити поставлену задачу, характеристики набору даних (атрибути), за необхідності провести попередню обробку даних та зменшити кількість об'єктів у вибірці, виділити аномалії та викиди, обрати стратегію роботи з об'єктами з пропусками, визначити стратегію тестування навчених алгоритмів. Для кожного з алгоритмів провести дослідження їх роботи на поставленій задачі, змінюючи параметри налаштування алгоритму.

Таблиця Б.1 – Набори даних для задачі класифікації

1	adult.arff	9	wine.arff
2	bank-data.arff	10	credit.arff
3	breast-cancer.arff	11	vote.arff
4	breast-w.arff	12	spambase.arff
5	labor.arff	13	zoo.arff
6	postoperative.arff	14	tic-tac-toe.arff
7	heart-statlog.arff	15	mushroom.arff
8	diabetes.arff	16	vehicle.arff

Для вирішення задачі регресії обрати одну задачу за номером варіанту (N mod 7)+1. Іншу задачу обрати на власний смак.

Таблиця Б.2 – Набори даних для задачі регресії

1	cpu.arff’	5	housing.arff
2	auto_mpg.arff	6	bodyfat.arff
3	winequality-red.csv	7	fishcatch.arff
4	autoprice.arff	8	auto93.arff