Лабораторна робота №4

Програмна розробка методу класифікації

Мета роботи: Здобути навички програмної розробки методу класифікації.

Завдання: Розробити програму для класифікації даних.

Вступ

У першій лабораторній роботі Ви здобули навички використання алгоритмів класифікації, які доступні у пакеті WEKA. У даній лабораторній роботі Вам потрібно власноруч розробити програмну реалізацію одного з методів класифікації — “Наївний класифікатор Байєса”.

Метод класифікації “Наївний класифікатор Байєса”

Нехай вирішується задача класифікації листів з електронної скриньки на наявність спаму. Деякі електронні листи вже мають відмітки “Спам”, а де які “Не спам”. Необхідно розробити програму, яка буде аналізувати назви нових листів і повідомляти про можливість спаму з певним відсотком вірогідності.

З першої лабораторної роботи Вам вже відомо, що для класифікації необхідно побудувати модель класифікації. Модель класифікації є результатом аналізу зв'язків в певному навчальному наборі даних.

У якості прикладу розглянемо назви електронних листів (навчальні дані):

Назва 1: New meeting tomorrow (file)
Назва 2: Corporate party tomorrow

Назва 3: Free sales party (SPAM)
Назва 4: Free file for you (SPAM)
Назва 5: New greeting text
Назва 6: Free file upload (SPAM)

Рис. 1. Навчальні дані

Проводити аналіз можна за різними метриками слів, у тому числі: за кількістю букв у слові, відношенням слова до певної частини мови, за змістом. Також можна проводити аналіз зустрічаємості слів у назвах.

Для оцінки зустрічаємості слів складається словник з переліку слів та кількості їх повторень, визначаються класи, у нашому прикладі їх буде два - “Спам” та “Не спам”.

Таблиця 1. Словник слів та кількість їх повторень

Слово	Кількість
tomorrow	2
free	3
file	3
new	2
party	2
Інші	1

Розрахована кількість повторень слів розділяється на повторення кожного з перерахованих класів.

Таблиця 2. Словник слів та кількість їх повторень за класами

Слово	Кількість “Спам”	Кількість “Не спам”
tomorrow	2	0
free	0	3
file	1	2
new	2	0
party	1	1

Використовуючи кількість повторень слів можна розрахувати вірогідність їх відношення до певного класу.

Таблиця 3. Словник слів та їх вірогідність за класами

Слово	Кількість “Спам”, 0.5	Кількість “Не спам”, 0.5
tomorrow	1	0
free	0	1
file	0.33	0.66
new	1	0
party	0.5	0.5

У нашому прикладі побудова моделі класифікації “Наївний класифікатор Байєса” ґрунтується на статистиці слів у назвах електронних листів.

Розрахунок загальної вірогідності

Для того, щоб класифікувати назву листа необхідно мати словник слів та їх вірогідності. Якщо класів слів два (як у нашому прикладі), необхідно для кожного класу з вірогідністю 0.5 розрахувати загальну вірогідність приналежних слів за наступною формулою:

$$p_{повідом} = p_{клас} \cdot \prod_{i=1}^{n} p_{слово}[i], \quad (1)$$

де $p_{повідом}$ — вірогідність приналежності повідомлення до класу, $p_{клас}$ — вірогідність класу, $p_{слово}$ — вірогідність приналежності слова до класу.

Нормування вірогідності класів

Якщо вірогідність хоча б одного слова у назві рівна нулю загальна вірогідність також буде рівна нулю. Щоб уникнути цього, усі вірогідності у словнику можна нормувати за наступною формулою:

$$p_{норм} = \frac{n_{слово} \cdot p_{ненорм} + 0.5}{n_{слово} + 1}, \quad (2)$$

де $p_{норм}$ — нормована вірогідність приналежності слова до класу, $p_{ненорм}$ — ненормована вірогідність приналежності слова до класу, $n_{слово}$ — кількість повторень слова у навчальному наборі.

Таблиця 4. Словник слів та їх нормовані вірогідності за класами

Слово	Кількість “Спам”, 0.5	Кількість “Не спам”, 0.5
tomorrow	0.83	0.16
free	0.13	0.87
file	0.37	0.62
new	0.83	0.16
party	0.5	0.5

Класифікація даних

Нехай модель класифікаціях вже побудована і необхідно визначити відноситься назва нового листа до класу “Спам”, чи ні.

Назва листа: Free file tommorow

Розрахуємо загальну вірогідність класу “Спам”:

$$p\_спам = 0.5 \cdot (0.87 \cdot 0.62 \cdot 0.16) = 0.043$$

Розрахуємо загальну вірогідність класу “Не спам”:

$$p\_не\_спам = 0.5 \cdot (0.13 \cdot 0.37 \cdot 0.83) = 0.019$$

Оскільки вірогідність класу “Спам” більша, ніж “Не спам”, можна зробити висновок, що електронних лист відноситься до класу “Спам”.

Звіт по роботі

1. Розробити програму;
2. Вибрати набір даних для навчання (відмінний від прикладу);
3. Вибрати набір даних для тестування (відмінний від прикладу);
4. Побудувати криву залежності якості класифікації відповідно до кількості екземплярів для навчання та тестування;
5. Навести код програми;
6. Сформулювати висновок.