Інтелектуальний аналіз даних

МЕТОД БАЙЄСА

Виконав:студент групи ОІ-36 Пироженко Назар

ЗМІСТ

ЩО ТАКЕ DATA MINING?

Data Mining - це набір методів, алгоритмів і засобів видобування із сирих і неопрацьованих даних необхідної інформації (необхідних знань)

Класифікація у контексті Data Mining - це віднесення об'єктів, процесів або явищ до певних класів за їхніми характеристиками

ЗАДАЧІ DATA MINING:
класифікація, кластеризація, регресія, побудова асоціативних правил

ЯКІ ІСНУЮТЬ МЕТОДИ КЛАСИФІКАЦІЇ?

1.Байєсова (“наївна”) класифікація;
2.Класифікація за допомогою штучних нейронних мереж;
3.Класифікація методом опорних векторів;
4.Статистичні методи, зокрема, лінійна регресія;
5.Класифікація за допомогою методу найближчого сусіда;
6.Класифікація CBR-методом;
7.Класифікація за допомогою генетичних алгоритмів

ОБРАНИЙ ДАТАСЕТ

Для виконання лабораторної роботи я обрав
набір даних для класифікації листів.

Мета: Класифікувати електронні листи як спам або не спам.
Розмір: 10 спостережень (рядків).
Ознаки:
«купити» — чи містить лист слово “купити” (так/ні)
«знижка» — чи містить лист слово “знижка” (так/ні)
Цільова змінна: Клас (спам / не спам)
Особливість: Невеликий ілюстративний датасет,
що дозволяє вручну розрахувати ймовірності
та побудувати частотні й правдоподібні таблиці

ПРИКЛАД

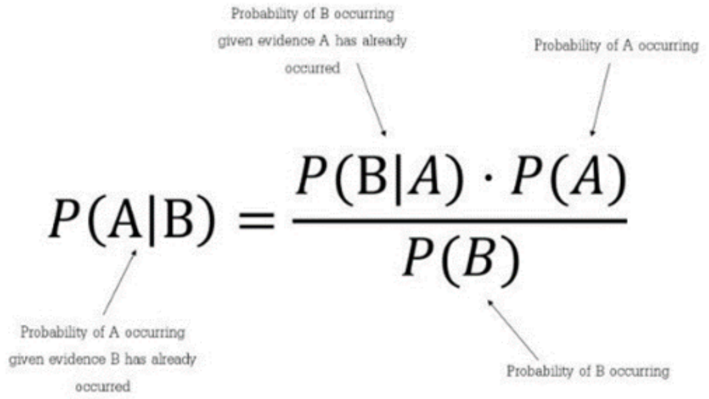
№	Містить "купити"	Містить "знижка"	Рішення
1	так	так	спам
2	ні	так	не спам
3	так	ні	спам
4	ні	ні	не спам
5	так	так	спам
6	ні	так	не спам
7	так	ні	спам
8	ні	ні	не спам
9	так	так	спам
10	ні	ні	не спам

ОБРАНИЙ МЕТОД

Для виконання лабораторної роботи я обрав **Байєсову класифікацію**.

“Наївна” Байєсова класифікація - це алгоритм класифікації, заснований на теоремі Байєса з припущенням про незалежність ознак.

Теорема Баєса задається наступною формулою:



Властивості наївної класифікації:

Використання всіх змінних і визначення всіх залежностей між ними;

Наявність двох припущень щодо змінних;

Всі змінні є однаково важливими;

Всі змінні є статистично незалежними, тобто значення однієї змінної нічого не говорить про значення іншої.

РЕАЛІЗАЦІЯ

Крок 1. Перетворити набір даних в частотну таблицю (frequency table).

Таблиця частот — це таблиця, яка показує, скільки разів (тобто з якою частотою) зустрічається кожне значення ознак у наборі даних, зокрема в межах кожного класу.

		Рішення	
		спам	Не спам
Містить “купити”	так	5	0
	ні	0	5

		Рішення	
		спам	Не спам
Містить “знижка”	так	3	2
	ні	2	3

РЕАЛІЗАЦІЯ

Крок 2. Створимо таблицю правдоподібності (likelihood table), розрахувавши відповідні ймовірності.

Таблиця частот — Це таблиця, у якій показано ймовірність появи певної ознаки в межах кожного класу.

		Рішення	
		спам	Не спам
Містить “купити”	так	1	0
	ні	0	1

		Рішення	
		спам	Не спам
Містить “знижка”	так	0.6	0.4
	ні	0.4	0.6

РЕАЛІЗАЦІЯ

Крок 3. За допомогою теореми Байєса розраховуємо апостеріорну ймовірність для кожного класу при заданих умовах. Клас з найбільшою апостеріорною ймовірністю буде результатом розв'язку задачі.

Це ймовірність того, що об'єкт належить до певного класу, після того, як ми врахували всі наявні ознаки.

P(спам) = 5/10 = 0.5
P(не спам) = 5/10 = 0.5

Для ознаки "купити":
$P(так_1 \mid спам) = 1.0$
$P(ні_1 \mid спам) = 0.0$
$P(так_1 \mid не\ спам) = 0.0$
$P(ні_1 \mid не\ спам) = 1.0$

Для ознаки "знижка":
$P(так_2 \mid спам) = 0.6$
$P(ні_2 \mid спам) = 0.4$
$P(так_2 \mid не\ спам) = 0.4$
$P(ні_2 \mid не\ спам) = 0.6$

Для класу "спам":
$P(спам \mid так, так) \propto P(так \mid спам) \times P(так \mid спам) \times P(спам) =$
$= 1.0 \times 0.6 \times 0.5 = 0.30$

Для класу "не спам":
$P(не\ спам \mid так, так) \propto P(так \mid не\ спам) \times P(так \mid не\ спам) \times P(не\ спам) =$
$= 0.0 \times 0.4 \times 0.5 = 0.00$

РЕАЛІЗАЦІЯ

Сума = 0.30 + 0.00 = 0.30
P(спам | так, так) = 0.30 / 0.30 = 1.0 (100%)
P(не спам | так, так) = 0.00 / 0.30 = 0.0 (0%)

Результат:

При умові "купити" = так і "знижка" = так, класифікатор передбачає: "спам" з ймовірністю 100%.

ОСОБЛИВОСТІ МЕТОДУ

Переваги методу:

В моделі визначаються залежності між усіма змінними, це дозволяє легко обробляти ситуації, в яких значення деяких змінних невідомі;

байєсовський метод дозволяє природним чином поєднувати закономірності, виведені з даних, і, наприклад, експертні знання, отримані в явному вигляді;

Недоліки методу

Перемножувати умовні ймовірності коректно лише тоді, коли всі вхідні змінні дійсно статистично незалежні; хоча часто даний метод показує досить гарні результати при недотриманні умови статистичної незалежності, але теоретично така ситуація повинна оброблятися більш складними методами, заснованими на навчанні байєсівських мереж;

На результати класифікації впливають тільки індивідуальні значення вхідних змінних, комбінований вплив пар або трійок значень різних атрибутів тут не враховується. Це могло б покращити якість класифікаційної моделі з точки зору її прогнозуючої точності, проте, збільшило б кількість варіантів, які необхідно перевірити.

ВИСНОВОК

У цій роботі ми розглянули застосування наївного байєсівського класифікатора для задачі бінарної класифікації електронних листів на спам і не спам.

Незважаючи на свою простоту, наївний байєсівський класифікатор широко використовується в реальних завданнях: фільтрація спаму, класифікація текстів, медична діагностика, розпізнавання образів. Метод забезпечує хороший баланс між точністю, швидкістю роботи та простотою реалізації, що робить його одним з базових алгоритмів машинного навчання.