НАЇВНИЙ БАЄСІВСЬКИЙ КЛАСИФІКАТОР

Онищук Роман

КЛАСИФІКАЦІЯ

- Мета класифікації:
  - Відношення значення до певного класу
  - Класифікація даних

- Області застосування
  - банківський бізнес (кредити)
  - маркетинг
  - медицина

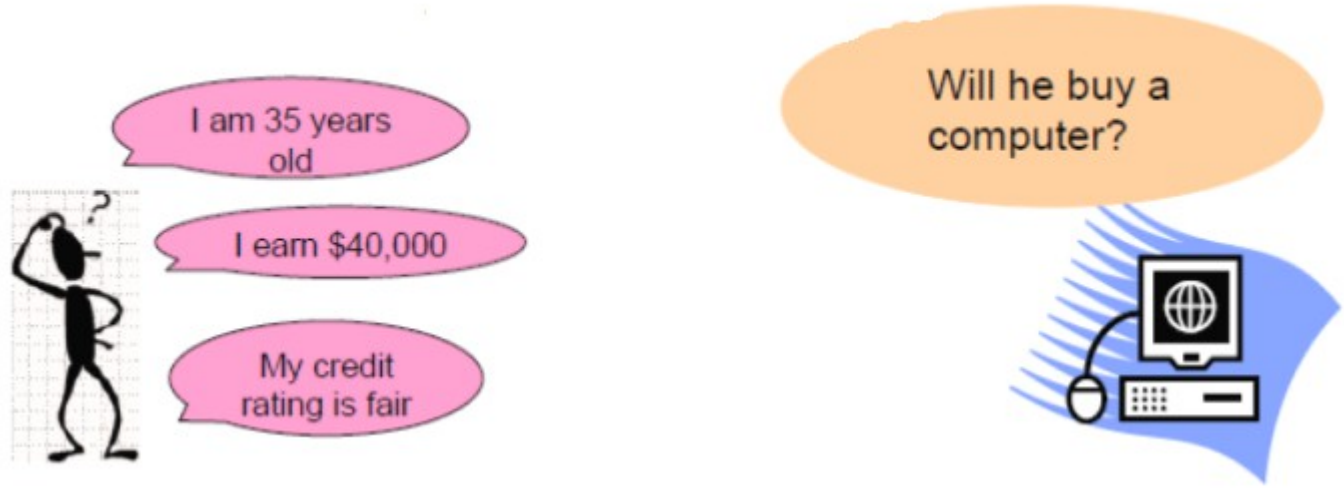
ЗАГАЛЬНИЙ ВИГЛЯД КЛАСИФІКАТОРА

- Побудова моделі: опис набору визначених класів
  - Кожен з атрибутів виборки належить до заздалегіть визначеного класу
  - Певний набір кортежів використовується для побудови моделі: training set
- Використання моделі: класифікація даних
  - Оцінювання точності заданої моделі
    - Значення отримане з тренувального набору порівнюється з відповідним значенням тестового набору
    - Оцінювана точність – відношення (зазвичай у відсотках) значень тестового набору, котрі були “коректно” класифіковані
    - Бажано, щоб тренувальний набір не залежав від тестового

ВВЕДЕННЯ В БАЄСІВСЬКУ КЛАСИФІКАЦІЮ

- Отож, що це?
  - Статистичний метод класифікації.
  - Один з базових методів використовуваних в машинному навчанні та інтелектуальному аналізі.
  - В його основі лежить теорема Байєса.
  - Вирішує проблеми, пов'язані як з дискретними так і з номінальними атрибутами.

Найпоширеніший приклад



X: 35 – річний чоловік, з зарплатою 40 тис. і хорошою кредитною історією

H: Гіпотеза, що цей чоловік купить комп’ютер.

ТЕОРЕМА БАЄСА

- Теорема Баєса:
  - P(H|X)= P(X|H) P(H)/ P(X)

- P(H|X) : умовна ймовірність, того, що клієнт X буде купувати комп’ютер за умови, що ми знаємо вік і дохід клієнта. (Постеріорна ймовірність H)
- P(H) : Ймовірність того, що клієнт буде купувати комп'ютер, незалежно від віку та доходів (Апріорна ймовірність H)
- P(X|H) : Ймовірність того, що особа 35 років, заробляє $40,000 точно купила комп’ютер
- P(X) : Ймовірність того, що людина з нашої виборки клієнтів 35 років і заробляє $ 40,000.

BAYESIAN CLASSIFIER

- Нехай X = {x1,x2, . . . , xn} – вибірка, компоненти якої являють собою значення, n-мірні вектори.
- Припустимо існує m класів: C1, C2,…Cm
- Баєсівський класифікатор передбачає, що X належить до класу Ci, якщо $P(Ci|X) > P(Cj|X)$, при $1 \leq i \leq m, i \neq j$.
- Отримуємо, максимальну апостеріорну ймовірність:

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

- Враховуючи наївне припущення взаємонезалежності, отримаємо, що P(X) – const.

NAÏVE BAYESIAN CLASSIFIER...

- Звідси й отримаєм:

$$P(X \mid C_i) = P(x_1, x_2, \ldots, x_n \mid C_i)$$

$$= P(x_1 \mid C_i) * P(x_2 \mid C_i) * \ldots\ldots\ldots * P(x_n \mid C_i)$$

$$= \prod_{k=1}^{n} P(x_k \mid C_i)$$

NAÏVE BAYESIAN CLASSIFIER...

- Ймовірності P(x1 | Ci),P(x2 | Ci),...,P(xn | Ci) можуть бути легко оцінені по навчальній вибірці, де xk - це значення атрибуту Ak виборки X. Звідси,ми отримуєм два випадки:
- Якщо значення Ak - дискретна(категорійована) змінна (тобто Ak прийматиме значення з певної фіксованої множини), то P(xk|Ci) це кількість зразків класу Ci в множинні значень навчальної вибірки T, що мають значення xk для атрибуту Ak, розділене на частоту(Ci,T),кількості Ci-тих зразів класу C в T.
- Якщо значення Ak безперервна (цифрова) змінна (тобто така, яка може приймати будь-яке значення в безперервному діапазоні), то зазвичай припускається, що змінні мають Гауссівський розподіл з середнім μ і стандартним відхиленням σ, і визначається формулою:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp - \frac{(x-\mu)^2}{2\sigma^2},$$

$$P(x_k|C_i) = g(x_k, \mu C_i, \sigma C_i).$$

ПРОБЛЕМА "НУЛЬОВОЇ ЧАСТОТИ"

- Що робити, якщо X має значення атрибуту xk, таке, що для нього не існує відповідного значення атрибуту в класі Ci? У цьому випадку P(xk|Ci) = 0,що призводить до P(X|Ci) = 0, хоча P(xk|Ci) для всіх інших атрибутів в X може бути великим і матимуть вплив на загальну ймовірність. Існує простий трюк, щоб уникнути цієї проблеми. Ми можемо припустити, що наш навчальний набір настільки великий, що додавши до кожної змінної, значення котрої нам треба підрахувати, одиницю привнесе мізерні зміни в розрахунок ймовірностей, але допоможе оминути отримання нульової ймовірності.

ПРИКЛАД: ВИБІРКА ПОКУПЦІВ КОМП’ЮТЕРІВ

Вік	Дохід	Кредитний статус	Купити комп’ютер
Молодий	Високий	Так	Ні
Молодий	Високий	Ні	Ні
Дорослий	Високий	Так	Так
Пенсіонер	Середній	Так	Так
Пенсіонер	Низький	Так	Так
Пенсіонер	Низький	Ні	Ні
Дорослий	Низький	Ні	Так
Молодий	Середній	Так	Ні
Молодий	Низький	Так	Так
Пенсіонер	Середній	Так	Так
Молодий	Середній	Ні	Так
Дорослий	Середній	Ні	Так
Дорослий	Високий	Так	Так

ПРИКЛАД…

- Нехай X = (Вік = Молодий , Дохід = Середній, Кредитний рейтинг = Так)
- *P (Купити комп = Так)* = 9/14 = 0,643
- *P (Купити комп = Ні)* = 5/14 = 0,357

- P (Вік = Молодий | Купити комп *= Так) = 2/9 = 0,222*
- P (Вік = Молодий | Купити комп *= Ні) = 3/5 = 0,6*

- P *(*Дохід = Середній | Купити комп *= Так) = 4/9 = 0,444*
- P *(*Дохід = Середній | Купити комп *= Ні) = 2/5 = 0,4*

- P (Кредитний рейтинг = *Так* | Купити комп *= Так) = 0,667*
- P (Кредитний рейтинг = *Так* | Купити комп *= Ні) = 0,4*

ПРИКЛАД…

- P(X | Купити комп = *Так)* =

  P (Вік = Молодий | Купити комп = *Так)**

  P *(*Дохід = Середній | Купити комп = *Так)**

  P (Кредитний рейтинг = *Так* | Купити комп = *Так)*

  *= 0.222*0.444*0.667 = 0.066*

  - P(X | Купити комп = *Ні) = 0,096*

  *Шукаєм клас, який максимізує P(X|Ci)*P(Ci)*

  - P(X | Купити комп = *Так) ** P(Купити комп = *Так) = 0.0424*
  - P(X | Купити комп = *Ні) ** P(Купити комп = *Ні) = 0,034*

Переваги та недостатки

Переваги:

- Легкий в реалізації
- Потребує невеликої тренувальної виборки для навчання
- Дуже високі результати класифікації

Недоліки:

- Через припущення взаємонезалежності можлива втрата точності

Висновки

✓Наївний Баєсовський метод є надзвичайно привабливим через його простоту та надійності.

✓ Це один з найстаріших алгоритмів класифікації і навіть у найпростішому вигляді він є на диво ефективним.

✓Він широко використовується в таких областях, як класифікація тексту і фільтрації спаму.

✓Існує велика кількість модифікацій даного алгоритму.