Вероятностни и статистически методи за разпознаване. Подход на Бейс. Гаусово разпределение и разстояние по Мехаланобис. Методи за разпознаване по разстояние

Иван Иванов, Изкуствен Интелект, 23610

I. Увод

Задачата за разпознаване може да се формулира така. Дадено е множество от обекти X. Всеки обект $x \in X$ се характеризира с множество от признаци (атрибути). Ако броят им е d то обекта x може да се разглежда като d-мерен вектор $x = (x_1, x_2, \ldots x_d)$. Дадено е и множество от категории (класове) от обекти $C = \{c_1, c_2, \ldots c_n\}$. Целта е да се намери функция (класификатор), която съпоставя категория $c \in C$ на всеки обект $x \in X$.

Едни от най-успешните методи за разпознаване са базирани на вероятностни и статистически подходи. При тях се избира класът, който е „най-вероятен“ за даден обект. Определянето на това колко е вероятен даден клас зависи от конкретните вероятностни и статистически модели, които се използват.

II. Подход на Бейс

Подходът на *Бейс* е основан на теоремата на *Бейс*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Доказателство:

$$\begin{aligned} P(B|A) &= \frac{P(B \cap A)}{P(A)} && \text{от дефиницията за условна вероятност} \\ &= \frac{P(A \cap B)}{P(A)} && \text{от комутативността на операцията сечение} \end{aligned}$$

или

$$P(A \cap B) = P(B|A)P(A) \qquad (1)$$

тогава

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && \text{от дефиницията за условна вероятност} \\ &= \frac{P(B|A)P(A)}{P(B)} && \text{след прилагане на равенство (1)} \end{aligned}$$

За целите на разпознаване на обекти теоремата на *Бейс* може да бъде записана като:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}$$

Това ни дава възможност да изразим *posterior*-ната вероятност $P(c_i \mid x)$ (колко е вероятна класификацията c_i, ако сме избрали обекта x) чрез prior-ните вероятности $P(c_i)$ (каква е вероятността случайно избран обект да принадлежи на класа c_i), $P(x)$ (каква е вероятността да бъде избран точно обекта x) и условната вероятност $P(x \mid c_i)$ (каква е вероятността да бъде избран обекта x при условие, че е избран обект принадлежащ на класа c_i). Теоремата може да бъде използвана за намирането на най-вероятната класификация. Всяка максимално вероятна класификация се нарича *maximum aposterior* класификация (*MAP*). *MAP* класификацията може да бъде определена чрез прилагане на теоремата на *Бейс*:

$$\begin{aligned} c_{MAP} &= \underset{c_i \in C}{argmax}\, P(c_i|x) \\ &= \underset{c_i \in C}{argmax}\, \frac{P(x|c_i)P(c_i)}{P(x)} \\ &= \underset{c_i \in C}{argmax}\, P(x|c_i)P(c_i) \qquad (2) \end{aligned}$$

Ако допуснем, че всяка класификация c_i е с еднаква *prior* вероятност ($P(c_i) = P(c_j)$ за всички класификации c_i и c_j), то формула *(2)* може да бъде допълнително опростена:

$$c_{ML} = \underset{c_i \in C}{argmax}\, P(x|c_i) \qquad (3)$$

В този случай класификацията c_{ML} се нарича *maximum likelihood (ML)* класификация.

III. Наивен Бейсов класификатор

Вече показахме, че:

$$c_{MAP} = \underset{c_i \in C}{argmax}\, P(x|c_i)P(c_i)$$

Тъй като x е d-мерен вектор горната формула може да бъде написана като:

$$c_{MAP} = \underset{c_i \in C}{argmax}\, P(x_1, x_2, \ldots x_d|c_i)P(c_i) \qquad (4)$$

При *наивният Бейсов класификатор* двата множителя в дясната част на *(4)* могат да бъдат емпирично апроксимирани с помощта на тренировъчно множество D, което се състои от наредени двойки *(x, c)* от примерни обекти и тяхната класификация. Вероятностите $P(c_i)$ се

апроксимират лесно, като се изчисли пропорцията на класовете c_i в тренировъчните примери. За съжаление, обаче, вероятността $P(x_1, x_2, \ldots x_d \mid c_i)$ не може лесно да се апроксимира. Поради тази причина *наивният Бейсов класификатор* прави опростяващо предположение, че отделните компоненти x_k са условно независими при дадена класификация. От условието за независимост имаме:

$$P(x_{1,} x_{2,} \ldots x_d | c_i) = \prod_{k=1}^{d} P(x_k | c_i) \qquad (5)$$

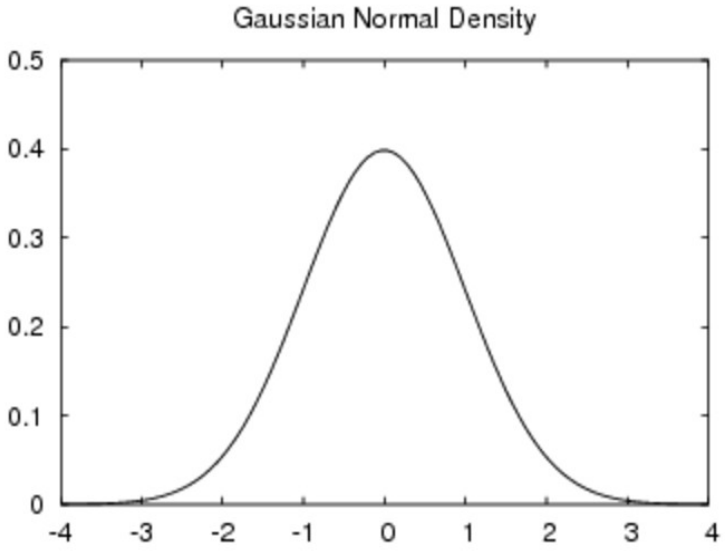
Като заменим в израз *(3)* получаваме:

$$c_{NB} = \underset{c_i \in C}{argmax}\, P(c_i) \prod_{k=1}^{d} P(x_k | c_i) \qquad (6)$$

C_{NB} е класификацията препоръчана от *наивният Бейсов класификатор*. След прилагането на опростяващото предположение вероятностите $P(x_k \mid c_i)$ могат да бъдат апроксимирани като се анализират тренировъчните примери с класификация c_i.

IV. Гаусово разпределение

Гаусовото разпределение е непрекъснато вероятностно разпределение, което моделира данни скупчени около определена средна стойност. Графиката на разпределението има формата на камбана, която достига своят връх в средната стойност.



Фигура 1: Гаусовото разпределение има формата на камбана

Probability Density Function-ът на *Гаусовото разпределение* се задава чрез израза:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad (7)$$

където параметърът μ е средното аритметично (върхът на камбаната), а параметърът σ^2 е вариацията. В случай когато $\mu = 0$ и $\sigma^2 = 1$ разпределението се нарича *нормално разпределение*. *Нормалното разпределение* е най-разпространеното разпределение в статистиката.

Важни свойства на *Гаусовото разпределение* са:

- $\int_{-\infty}^{+\infty} f(x;\mu,\sigma)\,dx = 1$, тъй като *f* е *Probability Density Function;*
- Графиката на функцията *f* е симетрична спрямо средното аритметично μ*;*
- Медианата и модата съвпадат със средното аритметично μ*;*

V. Ковариационна матрица

Ковариационна матрица е матрица, чиито елемент на позиция *(i, j)* съответства на ковариацията между *i*-тия и *j*-тия елемент на вектор от случайни величини. По-формално ако $x = (x_1, x_2, \ldots x_n)$ е вектор от случайни величини, всяка, от които има крайна вариация, то корелационната матрица *S* е матрица чийто елемент *(i, j)* е ковариацията:

$$S_{i,j} = Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] \qquad (8)$$

VI. Разстояние по Мехаланобис

Формално *разстоянието по Мехаланобис* на вектора $x = (x_1, x_2, \ldots x_n)^T$ от група вектори с вектор от средни стойности $\mu = (\mu_1, \mu_2, \ldots \mu_n)^T$ с ковариационна матрица *S* е дефинирано като:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)} \qquad (9)$$

Разстоянието на Мехаланобис може да бъде дефинирано и като мярка за различие между два вектора от случайни величини *x* и *y* с корелационна матрица *S*:

$$d(x, y) = \sqrt{(x-y)^T S^{-1} (x-y)} \qquad (10)$$

Ако ковариационната матрица е матрицата на идентитета *I* то разстоянието *по Мехаланобис* се изражда в *Евклидово разстояние*.

Разтоянието по Мехаланобис често се прилага при класификацията на вектор *x*, като принадлежащ към един от *n* класа. За целта за всеки клас се изчислява ковариационна матрица, като се анализират примерни инстанции на класа. За да се класифицира векторът *x* се изчислява *разстоянието му по Мехаланобис* до всеки един от класовете. Класификаторът избира този клас, за който разстоянието е минимално.

VII. Методи за разпознаване по разстояние

Обектите които подлежат на разпознаване (класификация) са представени по обичайния начин са представени - като *d*-мерни вектори от признаци: $x = (x_1, x_2, \ldots x_d)$. Ако разгледаме векторите на всевъзможните обекти като точки в *d*-мерното пространство, често се оказва, че обектите от един и същ клас имат тенденцията да са скупчени заедно. Това свойство на обектите от един и същ клас позволява да изградим методи за класификация на обекти според това колко близо или далече класифицираният обект се намира до обектите, принадлежащи на даден клас.

Един от най-разпространените алгоритми за разпознаване по разстояние е *k-Nearest Neighbor* алгоритъмът. При него разполагаме с множество от тренировъчни примери *D* от наредени двойки *(x, c),* задаващи примерни обекти и теяхната класификация. За да бъде класифициран обект *x* се намират *k*-те най-близки до него обекта от тренировъчните примери и се избира тази класификация, която е най-често срещаната сред тях.

Често използван частен случай на *k-Nearest Neighbor* алгоритъма е *1-Nearest Neighbor* алгоритъма. При него за класифициране на обект *x* се използва класа на най-близкият до *x* обект от тренировъчните примери. При *k-Nearest Neighbor* алгоритмите най-често се използва евклидовото разстояние. Евклидовото разстояние между два вектора *x* и *y* е дефинирано като:

$$d(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \qquad (11)$$