

模型思想：

Steps: ① 通过已知类别的训练数据集，计算样本的先验概率。

② 利用贝叶斯概率公式，测算未知类别的样本属于某个类别的后验概率。
分母计算

③ 以最大后验概率所对应类别作为样本预测值。

朴素假设：未知样本(X)各属性值($X_1, X_2 \dots X_p$)对定类的影响
影响相互独立。

贝叶斯公式基础：

$$\text{条件概率: } P(B|A) = \frac{P(B \cdot A)}{P(A)}$$

$$\left\{ \begin{array}{l} \text{全概率公式: } P(A) = \sum_{i=1}^n P(B_i) \cdot P(A|B_i) \end{array} \right.$$

$$\rightarrow \text{贝叶斯公式: } P(C_i|X) = \frac{P(C_i \cdot X)}{P(X)} = \frac{P(C_i) \cdot P(X|C_i)}{\sum_{i=1}^k P(C_i) \cdot P(X|C_i)}$$

条件概率公式
全概率公式

已知先验概率情况下求条件概率，而正适合于我们通过训练数据集已知先验概率 $P(C_i) = \frac{N_{C_i}}{N} \Rightarrow$ 类别样本数 / 总样本数 而 $\sum_{i=1}^k P(C_i) \cdot P(X|C_i) = P(X)$ 为常量，因此 $\arg \max P(C_i|X) \Leftrightarrow \arg \max P(C_i) \cdot P(X|C_i)$

△ 其中 C_i 表示样本所属某个类别，因此 ③ 中所述样本最终所属

类别 C_p 为 $P(C_p|X)$ 最大，即后验概率最大的类别。

朴素贝叶斯模型：

假设 X 中自变量是条件独立的，

X_1, X_2, \dots, X_p 间

$$\therefore P(X|C_i) = P(X_1, X_2, \dots, X_p | C_i)$$

$$= P(X_1 | C_i) \cdot P(X_2 | C_i) \cdots P(X_p | C_i)$$

$$= \prod_{k=1}^p P(X_k | C_i)$$

针对数据类型：

高斯分类器： X 为连续数值型，且

$$X \sim Gaussian(\mu, \sigma^2)$$

伯努利 分类器： X 为 0-1 二元值。

{二元分类器}

多项式 分类器： X 均为离散型变量，且 $X \sim$ 多项式分布

高斯分类器：

$$P(X_j | C_i) = \frac{1}{\sqrt{2\pi} \sigma_{ji}} \exp \left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji}, σ_{ji} 为 X_j 在 C_i 下的正态分布

的参数 μ_{ji}, σ_{ji} 。

e.g.

CSF 朴素贝叶斯模型

高斯贝叶斯分类器

x_j	X	y
x_{j1}	x_{j2}	y
x_{j1}	x_{j2}	C_{i1}
x_{j1}	x_{j2}	C_{i2}

(s) 贝叶斯后验概率

$$\begin{aligned}
 & P(\text{loan} = 0 | \text{Age} = 24, \text{Income} = 8500) \\
 & = P(\text{loan} = 0) \times P(\text{Age} = 24 | \text{loan} = 0) \times P(\text{Income} = 8500 | \text{loan} = 0) \\
 & = 0.5 \times 0.0926 \times 1.0384 \times 10^{-6} = 4.8079 \times 10^{-8}
 \end{aligned}$$

$$\begin{aligned}
 & P(\text{loan} = 1 | \text{Age} = 24, \text{Income} = 8500) \\
 & = P(\text{loan} = 1) \times P(\text{Age} = 24 | \text{loan} = 1) \times P(\text{Income} = 8500 | \text{loan} = 1) \\
 & = 0.5 \times 0.0375 \times 1.1456 \times 10^{-4} = 2.1479 \times 10^{-6}
 \end{aligned}$$

分子最大

经过上面的计算可知，当客户的年龄为24岁，并且收入为8500时，被预测为不放贷的概率是 4.8079×10^{-8} ，放贷的概率为 2.1479×10^{-6} ，所以根据 $\text{argmax } P(C_i)P(X|C_i)$ 的原则，最终该金融公司决定给客户放贷。

多项式分类器：

$$P(X_j | C_i) = \frac{N_{ik} + \alpha}{N_{it} + n\alpha}$$

→ 平滑系数，为防止 $P=0$ 的可能，
通常 $\alpha=1$ ，表示拉普拉斯平滑
因变量类别个数为 n ，即 C_1, C_2, \dots, C_n .
 $\hookrightarrow C_i$ 类别样本个数

自变量 X_j 的取值。

贷款中年龄≤28岁人数

$$e.g.: P(X_{j1k1} | C_i) = \frac{N_{ik} + \alpha}{N_{it} + n\alpha}$$

↑
贷款
↓
年龄
28岁
↓
贷款总人数

$$P(X_{j2k2} | C_{i2}) = \frac{N_{i2k2} + \alpha}{N_{i2} + n\alpha}$$

↑
不贷款
↓
收入
8000
↓
不贷款中收入 8000 人素文
2

伯努利分类器：

$$P(X_j | C_i) = p_i \cdot X_j + (1-p_i)(1-X_j)$$

类别为 C_i 时 $X_j = 1$ 的概率。

$$p = P(X_j = 1 | C_i) = \frac{N_{Xj} + \alpha}{N_i + n\alpha}$$