输入空间：$\mathcal{X} \in R^n$

输出空间：$\mathcal{Y} = \{c_1, c_2 \cdots, c_K\}$

输入：$x \in \mathcal{X}$ feature vector

输出：$y \in \mathcal{Y}$ class label

先验概率 $P(Y = c_k) \quad k = 1, 2 \cdots, K$

条件概率 $P(X = x \mid Y = c_k) = P(X^{(1)} = x^{(1)}, X^{(2)} = x^{(2)} \cdots, X^{(n)} = x^{(n)} \mid Y = c_k)$

条件独立性假设：$P(X = x \mid Y = c_k)$

$$= \prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k)$$

后验概率：$P(Y = c_k \mid X = x) = \dfrac{P(X = x \mid Y = c_k) P(Y = c_k)}{\sum_{k=1}^{K} P(X = x \mid Y = c_k)}$

(应用 Bayes 公式)

$$= \frac{\prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k) P(Y = c_k)}{\sum_{k=1}^{K} \prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k)}$$

朴素贝叶斯分类：$y = f(x) = \arg\max_{c_k} \dfrac{\prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k) P(Y = c_k)}{\sum_{k=1}^{K} \prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k)}$

$$= \arg\max_{c_k} \prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k) P(Y = c_k)$$

(分母被所有 c_k 共享)。

后验概率最大化 $\Leftrightarrow$ 期望风险最小化

$$\underset{0-1\text{损失函数}}{L(Y, f(X))} = \begin{cases} 1 & Y \neq f(X) \\ 0 & Y = f(X) \end{cases}$$

$$R_{exp}(f) = E[L(Y, f(X))]$$

期望风险函数

$$= E_X \sum_{k=1}^{K} [L(c_k, f(X))] P(c_k|X)$$

$$f(x) = \arg\min_{y \in Y} \sum_{k=1}^{K} L(c_k, y) P(c_k|X=x)$$

$$= \arg\min_{y \in Y} \sum_{k=1}^{K} P(y \neq c_k|X=x)$$

$$= \arg\min_{y \in Y} 1 - P(y = c_k|X=x)$$

$$= \arg\max_{y \in Y} P(y = c_k|X=x)$$

极大似然估计：

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I\{y_i = c_k\}}{N}$$

$$P(x^{(j)} = a_{jr} | Y = c_k) = \frac{\sum_{i=1}^{N} I\{x_i^{(j)} = a_{jr}, y_i = c_k\}}{\sum_{i=1}^{N} I\{y_i = c_k\}}$$

（a_{jr} 表示 x 第 j 个特征的 r 个取值）

朴素贝叶斯算法：本质上 Bernoulli Event Model

输入：$T = \{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$ N个样本

$x_i = (x_i^{(1)}, x_i^{(2)} \dots, x_i^{(n)})$ n个特征

$x_i^{(j)} \in \{a_{j1}, a_{j2} \cdots, a_{jp}\}$ 每个特征p种取值

$y_i \in \{c_1, c_2 \cdots c_K\}$ K个分类

(1) 计算先验概率

$$P(Y=c_k) = \frac{\sum_{i=1}^{N} I\{y_i = c_k\}}{N}$$

$$P(X^{(j)} = a_{jp} \mid Y = c_k) = \frac{\sum_{i=1}^{N} I\{x^{(j)} = a_{jp}, y_i = c_k\}}{\sum_{i=1}^{N} I\{y_i = c_k\}}$$

($j = 1, 2 \dots, n$

$p = 1, 2 \dots, p$

$k = 1, 2 \dots, N$)

(2) 对于给定x，计算

$$P(Y=c_k) \prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k)$$

(3) 确定x的类

$$y = \arg\max_{c_k} P(Y = c_k) \prod_{i=1}^{n} P(X^{(i)} = x^{(i)} \mid Y = c_k)$$

贝叶斯估计

$$P(X^{(j)} = a_{jp} \mid Y = c_k) = \frac{\sum_{i=1}^{N} I\{x_i^{(j)} = a_{jp}, y_i = c_k\} + \lambda}{\sum_{i=1}^{N} I\{y_i = c_k\} + P \cdot \lambda}$$

P: 特征可能取值数

$\lambda = 1$ 时: Laplace Smoothing

① $P(X^{(j)} = a_{jp} \mid Y = c_k) > 0$

② $\sum_{p=1}^{P} P(X^{(j)} = a_{jp} \mid Y = c_k) = 1$

同理，先验概率变为 $P(Y = c_k) = \frac{\sum_{i=1}^{N} I\{y_i = c_k\} + \lambda}{N + K\lambda}$

实际中通常不用，因为数据集中 $P(Y = c_k) \neq 0$ $k = 1, 2, \ldots, K$

补充: 将连续值转化为离散值(如设立10个左右区间)即可以应用 NB 算法

补充: 多项式概率分布 Multinomial Event Model

$$P(X^{(j)} = a_{jp} \mid y = c_k) = \frac{\sum_{i=1}^{N} \sum_{m=1}^{n} I\{X_m^{(j)} = a_{jp}, y = c_k\}}{\sum_{i=1}^{N} I\{y = c_k\} \cdot n}$$

n: x 的特征维度

(以文本为例，假设每个单词与其出现的位置无关)

应用 Laplace 变换: 同上