

朴素贝叶斯法

- 朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法；
- 给定训练数据集，目的是基于特征条件独立假设学习输入/输出的联合概率分布，然后基于此模型，对于给定的输入 x ，利用贝叶斯定理求出后验概率最大的输出 y ；
- 朴素贝叶斯（naive Bayes）与贝叶斯估计（Bayesian estimation）是不同的概念；

4.1 朴素贝叶斯的学习与分类

4.1.1 基本方法

- 朴素贝叶斯通过训练数据集学习**联合概率分布** $P(X, Y)$ ，具体地学习以下先验概率分布和条件概率分布，于是学习到联合概率分布

- 先验概率分布：

$$P(Y = c_k), k = 1, 2, \dots, K \quad (1)$$

- 条件概率分布：

$$P(X = x|Y = c_k) = P(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)}|Y = c_k), k = 1, 2, \dots, K \quad (2)$$

- 上述条件概率分布具有指数级数量的参数，其估计是不可行的（假设 $x^{(d)}$ 可取值 S_d 个，那么参数个数为 $K \prod_{d=1}^n S_d$ ）；
- **条件独立性假设**：一个强假设，因此得名

$$P(X = x|Y = c_k) = \prod_{d=1}^n P(X^{(d)} = x^{(d)}|Y = c_k) \quad (3)$$

- 朴素贝叶斯实际上学习到生成数据的机制，属于生成模型；
- **分类预测**：对于给定输入 x ，通过模型计算后验概率分布 $P(Y = c_k|X = x)$ ，将后验概率最大的类作为输出，根据贝叶斯定理

$$P(Y = c_k | X = x) = \frac{P(X = x | Y = c_k)P(Y = c_k)}{\sum_k P(X = x | Y = c_k)P(Y = c_k)} \\ = \frac{P(Y = c_k) \prod_d P(X^{(d)} = x^{(d)} | Y = c_k)}{\sum_k P(Y = c_k) \prod_d P(X^{(d)} = x^{(d)} | Y = c_k)}$$

- 朴素贝叶斯分类器（分母是全概率公式，结果一致，可以省略）：

$$y = f(x) = \arg \max_{c_k} P(Y = c_k) \prod_d P(X^{(d)} = x^{(d)} | Y = c_k) \quad (4)$$

4.1.2 后验概率最大化含义

- 可以证明取后验概率最大的类等价于期望风险最小化;

4.2 参数估计

4.2.1 极大似然估计

- 先验概率：

$$P(Y = c_k) = \frac{\sum_{i=1}^N I(y_i = c_k)}{N}, k = 1, 2, \dots, K \quad (5)$$

- 条件概率（设第 d 个特征 $x^{(d)}$ 的可能取值集合为 $\{a_{d1}, a_{d2}, \dots, a_{dS_d}\}$ ）：

$$P(X^{(d)} = a_{dl} | Y = c_k) = \frac{\sum_{i=1}^N I(x_i^{(d)} = a_{dl}, y_i = c_k)}{\sum_{i=1}^N I(y_i = c_k)}, l = 1, 2, \dots, S_d \quad (6)$$

4.2.2 学习与分类算法

- 朴素贝叶斯算法：

- 计算训练集的先验概率和条件概率；
- 对于给定的实例 $x = (x^{(1)}, x^{(2)}, \dots, x^{(n)})^T$ ，计算：

$$P(Y = c_k) \prod_d P(X^{(d)} = x^{(d)} | Y = c_k), k = 1, 2, \dots, K \quad (7)$$

- 选择后验概率最大的作为输出类；

4.2.3 贝叶斯估计

- 使用MLE可能会出现所要的估计概率为0的情况，会影响后验概率结果，使分类产生偏差；
- **条件概率的贝叶斯估计：**

$$P_{\lambda}(X^{(d)} = a_{dl} | Y = c_k) = \frac{\sum_{i=1}^N I(x_i^{(d)} = a_{dl}, y_i = c_k) + \lambda}{\sum_{i=1}^N I(y_i = c_k) + S_d \lambda}, l = 1, 2, \dots, S_d \quad (8)$$

其中 $\lambda \geq 0$ ，相当于在各个频数上赋予了一个基础正数 λ ；当 $\lambda = 0$ 时就是极大似然估计；经常取值 $\lambda = 1$ ，称为拉普拉斯平滑（Laplace smoothing）；可以证明经过上述调整后仍然是一种概率分布（ $p > 0, \sum p = 1$ ）；

- **先验概率的贝叶斯估计：**

$$P_{\lambda}(Y = c_k) = \frac{\sum_{i=1}^N I(y_i = c_k) + \lambda}{N + K \lambda}, k = 1, 2, \dots, K \quad (9)$$

后记-连续型属性

- 当属性值是连续型时，可以假设连续变量服从某种概率分布，然后使用训练数据估计分布的参数，高斯分布经常被用来表示连续属性的类条件概率分布：

$$P(X^{(d)} = a_d | Y = c_k) = \frac{1}{\sqrt{2\pi}\sigma_{dk}} \exp\left(-\frac{(a_d - \mu_{dk})^2}{2\sigma_{dk}^2}\right) \quad (10)$$

其中参数 μ_{dk} 与 σ_{dk}^2 可以用类 c_k 的所有训练数据关于特征 $X^{(d)}$ 的样本均值和方差估计；