

第4章 贝叶斯分类器

导引

- 朴素贝叶斯(naive Bayes) 法是基于贝叶斯定理与特征条件独立假设的分类方法
- 对于给定的训练数据集
 - 模型：首先基于特征条件独立假设，学习输入输出的联合概率分布
 - 因采用的原理是贝叶斯定理，核心的判断依据是后验概率来判断，计算转换为
 - 先验概率 $P(y = c_k)$
 - 条件概率 $P(X = x|Y = c_k)$
 - 预测：【先验概率+观测数据=>后验概率】利用贝叶斯定理求出后验概率最大的输出
- 朴素贝叶斯应用于分类时
 - 使用最大后验概率（观察到新数据下的类别）的类别
- 注意：比较的相对大小，不一定要求出值
 - 常用技巧，缩放

概率(回顾)

➤ 定义: Ω 为试验 E 的样本空间, $B_1, B_2, \dots, B_n$ 为 E 的一组事件, 若

$$B_i \cap B_j = \emptyset, i, j = 1, 2, \dots, n$$

$$B_1 \cup B_2 \cup \dots \cup B_n = \Omega$$

则称 $B_1, B_2, \dots, B_n$ 为样本空间的一个分划

➤ 全概公式

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n) = \sum_{i=1}^n P(B)P(A|B_i)$$

➤ 乘法规则

$$P(x, y) = P(x)P(y | x)$$

联合概率=边沿概率*条件概率

➤ 贝叶斯定理

$$P(\theta | D) = \frac{P(\theta)P(D | \theta)}{P(D)}$$

1 朴素贝叶斯法的学习与分类

贝叶斯公式，以及条件独立性假设(特征的各个分量独立)

朴素贝叶斯法将实例分到后验概率最大的类中，等价于期望风险最小化

基本方法

输入空间: $\mathcal{X} \subseteq \mathbb{R}^n$, 输出空间为类标记集合: $\mathcal{Y} = \{c_1, c_1, \dots, c_k\}$

联合概率分布 $P(X, Y)$

训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 由联合概率分布独立同分布产生

朴素贝叶斯, 通过训练数据集学习联合概率分布 $P(X, Y)$

先验概率分布 (类别先验概率)

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

条件概率分布 (类条件概率)

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

朴素贝叶斯: 条件独立性假设(特征的各个分量独立)

$$P(X = x|Y = c_k) = P(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)}|Y = c_k) = \prod_{j=1}^n P(X^{(j)} = x^{(j)}|Y = c_k)$$

朴素贝叶斯方法用于分类

朴素贝叶斯：条件独立性假设(特征的各个分量独立, 朴素由来)

$$P(X = x|Y = c_k) = P(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)}|Y = c_k) = \prod_{j=1}^n P(X^{(j)} = x^{(j)}|Y = c_k)$$

对给定的输入 x , 通过学习到的模型计算后验概率分布 $P(X = x|Y = c_k)$, 将后验概率最大的类作为 x 的类输出

$$\begin{aligned} P(Y = c_k|X = x) &= \frac{P(X = x, Y = c_k)}{P(X = x)} = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_k P(X = x|Y = c_k)P(Y = c_k)} \\ &= \frac{P(Y = c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)}|Y = c_k)}{\sum_k P(Y = c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)}|Y = c_k)} \end{aligned}$$

朴素贝叶斯方法用于分类

朴素贝叶斯分类器(最大后验概率)

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

$$y = f(x) = \arg \max_{c_k} P(Y = c_k | X = x) = \arg \max_{c_k} \frac{P(Y=c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_k P(Y=c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)} | Y = c_k)}$$

对于所有 c_k 分母不变, 故上式变为

$$y = f(x) = \arg \max_{c_k} P(Y = c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)} | Y = c_k)$$

后验概率最大化的含义

朴素贝叶斯法将实例分到后验概率最大的类中，等价于期望风险最小化(结论)

【证明】选择0-1损失 $L(Y, f(X))$ ， $f(X)$ 为决策函数，期望风险 $R_{\text{exp}(f)}$

$$\begin{aligned} R_{\text{exp}(f)} &= E[L(Y, f(X))] = E_X \left(\sum_Y L(c_k, f(X)) P(Y|X) \right) \\ &= E_X \left(\sum_{k=1}^K [L(c_k, f(X)) P(c_k|X)] \right) \end{aligned}$$

$R_{\text{exp}(f)}$ 最小化：对 $E_X(\quad)$ 中的 $\sum_{k=1}^K [L(c_k, f(X)) P(c_k|X)]$ 的每一个 $X = x$ 都最小化

$$\begin{aligned} f(x) &= \arg \min_{y \in \mathcal{Y}} \sum_{k=1}^K [L(c_k, y) P(c_k|X = x)] = \arg \min_{y \in \mathcal{Y}} \sum_{k=1}^K P(y \neq c_k|X = x) \\ &= \arg \min_{y \in \mathcal{Y}} (1 - P(y = c_k|X = x)) = \arg \max_{y \in \mathcal{Y}} P(y = c_k|X = x) \end{aligned}$$

即最大化后验概率

2 朴素贝叶斯法的参数估计

古典概率公式计算经验概率

朴素贝叶斯法的参数估计

朴素贝叶斯法中，学习意味着估计 $P(y = c_k)$ 和 $P(X = x|Y = c_k)$

先验概率 $P(y = c_k)$ 的极大似然估计

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

设第 j 个特征 $x^{(j)}$ 可能取值的集合为： $\{a_{j1}, a_{j2}, \dots, a_{jS_j}\}$,

条件概率 $P(X^{(j)} = a_{jl}|Y = c_k)$ 的极大似然估计

$$P(X^{(j)} = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^N I(x_i^{(j)} = a_{jl}, y_i = c_k)}{\sum_{i=1}^N I(y_i = c_k)}, j = 1, \dots, n; l = 1, \dots, S_j; k = 1, \dots, K$$

学习与分类算法(Naïve Bayes Algorithm)

输入：训练集 $T = \{(x_1, y_1), \dots, (x_N, y_N)\}$ ，其中 $x_i = (x_i^{(1)}, \dots, x_i^{(n)})^T$ ， $x_i^{(j)} \in \{a_{j1}, \dots, a_{jS_j}\}$

输出： x 的分类

1. 计算先验概率和条件概率

$$P(Y = c_k) = \frac{\sum_{i=1}^N I(y_i = c_k)}{N}, k = 1, 2, \dots, K$$
$$P(X^{(j)} = a_{jl} | Y = c_k) = \frac{\sum_{i=1}^N I(x_i^{(j)} = a_{jl}, y_i = c_k)}{\sum_{i=1}^N I(y_i = c_k)}$$
$$j = 1, 2, \dots, n; l = 1, 2, \dots, S_j; k = 1, 2, \dots, K$$

2. 对于给定的实例 $x = (x^{(1)}, x^{(2)}, \dots, x^{(i)}, \dots, x^{(n)})^T$ ，计算各个类别的

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

3. 确定 x 的类别： $y = \arg \max_{c_k} P(Y = c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)} | Y = c_k)$

贝叶斯估计

用极大似然估计可能会出现所要估计的概率值为0的情况，从而影响后验概率的计算结果，使分类产生偏差，可采用贝叶斯估计解决

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

式中 $\lambda \geq 0$ 。等价于在随机变量各个取值的频数上赋予一个正数 $\lambda > 0$

当 $\lambda = 0$,极大似然估计; $\lambda = 1$,称为拉普拉斯平滑(Laplacian smoothing)

对任何 $l = 1, \dots, S_j, k = 1, 2, \dots, K$, 有

$$P_{\lambda}(X^{(j)} = a_{jl} \mid Y = c_k) > 0$$
$$\sum_{l=1}^{S_j} P_{\lambda}(X^{(j)} = a_{jl} \mid Y = c_k) = 1$$

表明确为一种概率分布。同样, 先验概率的贝叶斯估计为

$$P_{\lambda}(Y = c_k) = \frac{\sum_{i=1}^N I(y_i = c_k) + \lambda}{N + K \lambda}$$