统计学习第四章

2022年8月10日 星期三 14:21

朴素贝叶斯法

设：输入空间 $\mathcal{X} \subseteq R^n$ 为 n 维向量的集合.

输出空间为标记集合 $\mathcal{Y} = \{c_1, c_2, \cdots, c_K\}$

训练集为：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$

$P(X, Y)$ 为 X 和 Y 的联合概率分布

先验概率分布为：$P(Y = c_k)$

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

联合概率分布 $P(X, Y)$

独立性假设：$P(X = x \mid Y = c_k) = P(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)} \mid Y = c_k)$

由贝叶斯定理.

$$P(Y = c_k \mid X = x) = \frac{P(X = x \mid Y = c_k) P(Y = c_k)}{\sum_k P(X = x \mid Y = c_k) P(Y = c_k)}$$

$$y = f(x) = \arg\max_{c_k} \frac{P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}{\sum P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}$$

后验概率最大化：

$$L(Y, f(X)) = \begin{cases} 1 & Y \neq f(X) \\ 0 & Y = f(X) \end{cases}$$

经变换得：$f(x) = \arg\max_{c_k} P(c_k \mid X = x)$

·极大似然估计：

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

朴素贝叶斯算法

$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$

其中 $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T$

$x_i^{(j)}$ 是第 i 个样本的第 j 个特征

$y_i \in \{c_1, c_2, \cdots, c_K\}$

朴素贝叶斯算法实现

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

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

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

确定实例 x 的类：$y = \arg\max_{c_k} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} \mid Y = c_k)$

eg.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1