第4章 朴素贝叶斯法

主要内容

- 朴素贝叶斯法的学习和分类
- 朴素贝叶斯法的参数估计

4.1朴素贝叶斯法的学习和分类

最大似然估计：

假设外观相同的两个箱子，甲箱中有99个白球，1个黑球，乙箱中有99个黑球，1个白球，若某次取出来的是白球，请问最有可能是从哪个箱子取出的？

基本思想

首先基于**特征条件的独立**假设，学习输入/输出的联合概率;
然后基于此模型，对于给定输入x，利用贝叶斯定理，求出**后验概率最大**的输出y。

1、基本方法

输入空间X∈Rn为n维向量的集合，输出空间为类标记集合У={c_1,c_2,..c_n}，输入特征向量x∈ X，输出为类标签y∈ У，X是在输入空间上的随机向量，Y是在输出空间У上的随机变量，P(X,Y）是X和Y的联合概率分布，训练数据集T={(x_1,y_1), (x_2,y_2),…(x_N,y_N)}，由P(X,Y)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=c_k)。

学习先验概率分布P(Y=c_k),k=1,2,..K和条件概率分布P(X=x|Y=c_k)。

$$P(X，Y=c_k)=P(Y=c_k)\ P(X=x|Y=c_k)$$

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

条件概率分布$P(X=x|Y=c_k)$是 n维的条件概率，有指数级的参数，求解困难

假设$x^{(j)}$可能值有S_j个，j=1,2，...n，Y可能取值有K个，参数个数 $K\prod_{j=1}^{n}S_j$

条件独立性假设：$P(X=x|Y=c_k)$

$$
\begin{aligned}
&= P(X^{(1)}=x^{(1)}, X^{(2)}=x^{(2)}, .. X^{(n)}=x^{(n)}|Y=c_k) \\
&= \prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k)
\end{aligned}
$$

条件独立性假设：说明用于分类的特征，在类别确定的条件下是条件独立的。

简化了朴素贝叶斯法，牺牲了分类准确性。

后验概率$P(Y=c_k|X=x)$的计算：

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

代入独立假设条件

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

朴素贝叶斯分类器

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

对于所有的c_k，分母都相同。

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

2、后验概率最大化

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，假设选择0-1损失函数，f(X)为决策函数：

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

期望风险函数

$$\mathrm{R}_{\exp}(f) = E[L(Y, f(X))]$$

条件期望

$$\mathrm{R}_{\exp}(f) = E_{\mathrm{X}} \sum_{k=1}^{K} [L(\mathrm{c}_{\mathrm{k}}, f(X))] P(c_k \mid \mathrm{X})$$

只需对X=x逐个极小化，得：

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

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

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

$$= \arg\max_{y\in \mathrm{Y}} \sum_{k=1}^{K} P(y = c_k \mid \mathrm{X} = \mathrm{x})$$

后验概率最大化准则：

$$f(x) = \arg\max_{c_k} P(c_k \mid X = x)$$

4.2朴素贝叶斯法的参数估计

1、极大似然估计

先验概率$P(Y=c_k)$的极大似然估计：

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

设第j个特征$x^{(j)}$可能取值的集合为$\{a_{j1},a_{j2},...a_{jSj}\}$,条件概率的极大似然估计：

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

$j=1,2,..n$；$l=1,2,..S_j$；$k=1,2,..K$，$x_i^{(j)}$是第i个样本的第j个特征，a_{jl}是第j个特征可能取的第l个值

2、学习与分类算法

输入：训练数据$T=\{(x_1,y_1),(x_2,y_2),\ldots(x_N,y_N)\}$，其中$x_i=(x_i^{(1)}, x_i^{(2)},\ldots x_i^{(n)})^T$, $x_i^{(j)}$是第i个样本的第j个特征， $x_i^{(j)} \in\{a_{j1},a_{j2},\ldots a_{jsj}\}$，$a_{jl}$是第j个特征可能取得第l个值，j=1,2，…n，l=1,2，…S_j,$y_i\in\{c_1,c_2,\ldots c_k\}$，实例x；

输出：实例x的分类

（1）计算先验概率和条件概率

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

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

j=1,2,..n；l=1,2,..S_j；k=1,2,..K

（2）对于给定实例x=($x_i^{(1)}, x_i^{(2)}, \ldots x_i^{(n)})^T$, 计算

$$P(Y=c_k)\prod_{j=1}^{n} P(X^{(j)}=x^{(j)} \mid Y=c_k) \qquad k=1,2,..K$$

（3）确定实例x的类

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

3、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，这时会影响到后验概率的计算结果，使分类产生偏差.解决这一问题的方法是采用贝叶斯估计

条件概率的贝叶斯估计和先验概率的贝叶斯估计

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

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

$j=1,2,..n$；$l=1,2,..S_j$；$k=1,2,..K$

【例4.1】用训练数据学习一个朴素贝叶斯分类器，并确定x=(2,S)T的类标记y，X(1),X(2)为特征，取值的集合分别为A1={1,2,3}，A2={S,M,L},Y为类标记, Y∈C={1,-1}。按照拉普拉斯平滑估计概率（λ=1）

	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

解：A1={1,2,3}，A2={S,M,L}，C={1,-1}，
计算条件概率和先验概率下的贝叶斯估计:

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

P(Y=1)=10/17，P(Y=-1)=7/17

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

$P(X^{(1)}=1|Y=1)=3/12$， $P(X^{(1)}=1|Y=1)=4/12$， $P(X^{(1)}=1|Y=1)=5/12$

$P(X^{(2)}=S|Y=1)=2/12$， $P(X^{(2)}=M|Y=1)=5/12$， $P(X^{(2)}=L|Y=1)=5/12$

【例4.1】用训练数据学习一个朴素贝叶斯分类器，并确定x=(2,S)T的类标记y，X(1),X(2)为特征，取值的集合分别为A1={1,2,3}，A2={S,M,L},Y为类标记, Y∈C={1,-1}。按照拉普拉斯平滑估计概率（λ=1）

	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

$P(X^{(1)}=1|Y=-1)=4/9$， $P(X^{(1)}=1|Y=-1)=3/9$， $P(X^{(1)}=1|Y=-1)=2/9$

$P(X^{(2)}=S|Y=-1)=4/9$， $P(X^{(2)}=M|Y=-1)=3/9$， $P(X^{(2)}=L|Y=-1)=2/9$

对于给定$x=(2,S)^T$计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)\,P(X^{(2)}=S|Y=1)=0.0327$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)\,P(X^{(2)}=S|Y=-1)=0.0610$$

由于$P(Y=-1)P(X^{(1)}=2|Y=-1)\,P(X^{(2)}=S|Y=-1)$值最大，所以$y=-1$

伯努利模型

在伯努利模型中，每个特征的取值是布尔型的(true和false)，或者1和0。即一个特征有在一个样本中是否出现。

先验概率p(c)=属于类c样本总数/整个训练样本的样本总数

类条件概率p(tk|c)=(类c下包含特征tk的样本总数+1)/(类c下的样本总数+2)

	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

$P(Y=1)=?$ ，$P(Y=-1)=?$

$P(X^{(1)}=1|Y=1)=?$ ， $P(X^{(1)}=1|Y=1)=?$ ， $P(X^{(1)}=1|Y=1)=?$

$P(X^{(2)}=S|Y=1)=?$ ， $P(X^{(2)}=M|Y=1)=?$ ， $P(X^{(2)}=L|Y=1)=?$

高斯模型

假设男性和女性的身高、体重和脚掌是正态分布，通过样本分别计算出均值μ和方差σ^2

身高(英寸)	体重(磅)	脚掌(英寸)	性别
6	180	12	男
5.92	190	11	男
5.98	170	12	男
5.92	165	10	男
5	100	6	女
5.5	150	8	女
5.42	130	7	女

预测某人身高为6英尺，体重179磅，脚掌10英寸的性别。则：

$$\mathrm{p(height|male)} = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(6-\mu)^2}{2\sigma^2}}$$

据此可以求出其他特征的类条件概率，从而最终推断出该人是男性还是女性。

朴素贝叶斯分类的优缺点

优点：
（1） 算法逻辑简单,易于实现
（2）分类过程中时，开销小（假设特征相互独立，只会涉及到二维存储）

缺点：
在属性个数比较多或者属性之间相关性较大时，分类效果不好。