第8章 朴素贝叶斯法

1. 朴素贝叶斯法的学习与分类
2. 朴素贝叶斯法的参数估计

*《机器学习方法》第4章

第8章 朴素贝叶斯法

1. 朴素贝叶斯法的学习与分类

2. 朴素贝叶斯法的参数估计

*《机器学习方法》第4章

朴素贝叶斯法

朴素贝叶斯（naïve Bayes）法是基于特征条件独立假设与贝叶斯定理的分类方法。

◻ 基本思想

- 对于给定的训练数据集，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；
- 然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。

朴素贝叶斯法实现简单，学习与预测的效率都很高，是一种常用的方法。

基本方法

- 输入空间$\mathcal{X} \subseteq \boldsymbol{R}^n$为$n$维向量的集合，输入为特征向量$x \in \mathcal{X}$
- 输出空间为类别标记集合$\mathcal{Y} = \{c_1, c_2, \cdots, c_K\}$，输出为类标记$y \in \mathcal{Y}$

将输入与输出看作是定义在输入（特征）空间与输出空间上的随机变量的取值

- 输入变量：X
- 输出变量：Y
- 输入变量取值：x
- 输出变量取值：y

- 假设输入与输出的随机变量X和Y遵循联合概率分布$P(X,Y)$
- 训练数据$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$被看作是依联合概率分布$P(X,Y)$独立同分布产生的

基本方法

基本思想

- 对于给定的训练数据集，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$P(X,Y)$；

具体地，学习以下先验概率分布及条件概率分布。

先验概率分布：$P(Y=c_k), k=1,2,\cdots,K$

条件概率分布：

$$P(X=x|Y=c_k)$$

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

$$P(X=x,Y=c_k)=P(Y=c_k)P(X=x|Y=c_k), k=1,2,\cdots,K$$

基本方法

- 基本思想

> 对于给定的训练数据集，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$P(X,Y)$；

- 具体地，学习以下先验概率分布及条件概率分布。

先验概率分布：$P(Y=c_k), k=1,2,\cdots,K$

特征条件独立假设

条件概率分布：

$$P(X=x|Y=c_k)$$

$$=P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)$$

$$=\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\,|Y=c_k\right), k=1,2,\cdots,K$$

基本方法

基本思想

- 对于给定的训练数据集，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$P(X,Y)$；

具体地，学习以下先验概率分布及条件概率分布。

先验概率分布：$P(Y=c_k),k=1,2,\cdots,K$

特征条件独立假设

条件概率分布：

$$P(X=x|Y=c_k)$$

$$=\prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}|Y=c_k\right),k=1,2,\cdots,K$$

用于分类的特征在类确定的条件下都是条件独立的

基本方法

基本思想

- 然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。

具体地，对给定的输入x，通过学习得到的联合概率分布，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x的类输出。后验概率计算根据贝叶斯定理进行：

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

$$=\frac{P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)},k=1,2,\cdots,K$$

基本方法

朴素贝叶斯法分类的基本公式

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)},$$

$$k=1,2,\cdots,K$$

朴素贝叶斯分类器

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

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

后验概率最大化的含义

朴素贝叶斯分类器

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

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

- 朴素贝叶斯法将实例分到后验概率最大的类中，这等价于期望风险最小化。

第4章 朴素贝叶斯法

1. 朴素贝叶斯法的学习与分类

2. 朴素贝叶斯法的参数估计

*《机器学习方法》第4章

极大似然估计

在朴素贝叶斯法中，学习意味着估计$P(Y=c_k)$和$P\left(X^{(j)}=x^{(j)}\,|Y=c_k\right)$，可以采用极大似然估计法来对其进行估计。

- 先验概率分布$P(Y=c_k)$的极大似然估计

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

极大似然估计

在朴素贝叶斯法中，学习意味着估计$P(Y=c_k)$和$P\left(X^{(j)}=x^{(j)}|Y=c_k\right)$，可以采用极大似然估计法来对其进行估计。

□ 条件概率分布$P\left(X^{(j)}=x^{(j)}|Y=c_k\right)$的极大似然估计

设第j个特征$x^{(j)}$可能取值的集合为$\left\{a_{j1},a_{j2},\cdots,a_{jS_j}\right\}$，条件概率$P\left(X^{(j)}=a_{jl}|Y=c_k\right)$的极大似然估计是

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

$$j=1,2,\cdots,n,l=1,2,\cdots,S_j,k=1,2,\cdots,K$$

$x_i^{(j)}$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；I为指示函数。

学习与分类算法（算法4.1）

算法4.1（朴素贝叶斯算法（naïve Bayes algorithm））

输入： 训练数据集$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，实例$x$

其中$x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$, $x_i^{(j)}$是第i个样本的第j个特征，

$x_i^{(j)}\in\left\{a_{j1},a_{j2},\cdots,a_{jS_j}\right\}$, a_{jl}是第j个特征可能取的第l个值，$j=1,2,\cdots,n,l=1,2,\cdots,S_j,y_i\in\{c_1,c_2,\dots c_K\},i=1,\cdots,N$。

输出： 实例x所属的类y。

学习与分类算法（算法4.1）

算法4.1（朴素贝叶斯算法（naïve Bayes algorithm））

（1）计算先验概率及条件概率

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

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

$$j=1,2,\cdots,n,l=1,2,\cdots,S_j,k=1,2,\cdots,K$$

$x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^{\mathrm{T}}$，$x_i^{(j)}$是第$i$个样本的第$j$个特征，

$x_i^{(j)}\in\left\{a_{j1},a_{j2},\cdots,a_{jS_j}\right\}$，$a_{jl}$是第$j$个特征可能取的第$l$个值，

$j=1,2,\cdots,n,l=1,2,\cdots,S_j,y_i\in\{c_1,c_2,\dots c_K\},i=1,\cdots,N$

学习与分类算法（算法4.1）

算法4.1（朴素贝叶斯算法（naïve Bayes algorithm））

（2）对于给定的实例$x = \left(x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right)^{\mathrm{T}}$，计算

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

（3）确定实例x的类

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

例4.1

试由表4.1的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^{\mathrm{T}}$ 的类标记 y。表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，Y 为类标记，$Y\in C=\{1,-1\}$。

表 4.1 训练数据

	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

例4.1

解：根据算法4. 1，由表4. 1容易计算下列概率：

先验概率 $P(Y=1)=\frac{9}{15}, P(Y=-1)=\frac{6}{15}$

条件概率

$$P\left(X^{(1)}=1 \middle| Y=1\right)=\frac{2}{9}, P\left(X^{(1)}=2 \middle| Y=1\right)=\frac{3}{9}, P\left(X^{(1)}=3 \middle| Y=1\right)=\frac{4}{9}$$

$$P\left(X^{(2)}=S \middle| Y=1\right)=\frac{1}{9}, P\left(X^{(2)}=M \middle| Y=1\right)=\frac{4}{9}, P\left(X^{(2)}=L \middle| Y=1\right)=\frac{4}{9}$$

$$P\left(X^{(1)}=1 \middle| Y=-1\right)=\frac{3}{6}, P\left(X^{(1)}=2 \middle| Y=-1\right)=\frac{2}{6}, P\left(X^{(1)}=3 \middle| Y=-1\right)=\frac{1}{6}$$

$$P\left(X^{(2)}=S \middle| Y=-1\right)=\frac{4}{9}, P\left(X^{(2)}=M \middle| Y=-1\right)=\frac{2}{6}, P\left(X^{(2)}=L \middle| Y=-1\right)=\frac{1}{6}$$

对于给定的$x=(2,S)^{\mathrm{T}}$，计算

后验概率

$$P(Y=1)P\left(X^{(1)}=2 \middle| Y=1\right)P\left(X^{(2)}=S \middle| Y=1\right)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P\left(X^{(1)}=2 \middle| Y=-1\right)P\left(X^{(2)}=S \middle| Y=-1\right)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$$

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

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，这时会使分类产生偏差，解决这一问题可以采用贝叶斯估计。

- 先验概率分布的贝叶斯估计

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

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，这时会使分类产生偏差，解决这一问题可以采用贝叶斯估计。

- 条件概率分布的贝叶斯估计

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

$$j=1,2,\cdots,n,l=1,2,\cdots,S_j,k=1,2,\cdots,K$$

$x_i^{(j)}$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；I为指示函数。

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

当$\lambda=0$时就是极大似然估计；

常取$\lambda=1$，这时称为拉普拉斯平滑（Laplacian smoothing）。

学习与分类算法（算法4.1）

算法4.1（朴素贝叶斯算法（naïve Bayes algorithm））

（2）对于给定的实例$x=\left(x^{(1)},x^{(2)},\cdots,x^{(n)}\right)^{\mathrm{T}}$，计算

$$P_\lambda(Y=c_k)\prod_{j=1}^{n}P_\lambda\left(X^{(j)}=x^{(j)}\,|Y=c_k\right),k=1,2,\cdots,K$$

（3）确定实例x的类

$$y=f(x)=\arg\max_{c_k}P_\lambda(Y=c_k)\prod_{j=1}^{n}P_\lambda\left(X^{(j)}=x^{(j)}\,|Y=c_k\right)$$

例4.2

问题同例 4.1，按照拉普拉斯平滑估计概率，即取$\lambda = 1$。

解：$A_1 = \{1,2,3\}$，$A_2 = \{S, M, L\}$，$C = \{1, -1\}$。按照式(4.10)和式(4.11)计算下列概率：

先验概率

$$P(Y=1) = \frac{10}{17},\quad P(Y=-1) = \frac{7}{17}$$

$$P\left(X^{(1)} = 1 \middle| Y = 1\right) = \frac{3}{12},\quad P\left(X^{(1)} = 2 \middle| Y = 1\right) = \frac{4}{12},\quad P\left(X^{(1)} = 3 \middle| Y = 1\right) = \frac{5}{12}$$

$$P\left(X^{(2)} = S \middle| Y = 1\right) = \frac{2}{12},\quad P\left(X^{(2)} = M \middle| Y = 1\right) = \frac{5}{12},\quad P\left(X^{(2)} = L \middle| Y = 1\right) = \frac{5}{12}$$

$$P\left(X^{(1)} = 1 \middle| Y = -1\right) = \frac{4}{9}, P\left(X^{(1)} = 2 \middle| Y = -1\right) = \frac{3}{9},\ P\left(X^{(1)} = 3 \middle| Y = -1\right) = \frac{2}{9}$$

$$P\left(X^{(2)} = S \middle| Y = -1\right) = \frac{4}{9}, P\left(X^{(2)} = M \middle| Y = -1\right) = \frac{3}{9},\ P\left(X^{(2)} = L \middle| Y = -1\right) = \frac{2}{9}$$

条件概率

例4.2

对于给定的$x = (2, S)^{\mathrm{T}}$，计算

$$
\begin{aligned}
P(Y=1)P\left(X^{(1)}=2 \middle| Y=1\right)P\left(X^{(2)}=S \middle| Y=1\right) &= \frac{10}{17}\cdot\frac{4}{12}\cdot\frac{2}{12} \\
&= \frac{5}{153} \\
&= 0.0327
\end{aligned}
$$

$$
\begin{aligned}
P(Y=-1)P\left(X^{(1)}=2 \middle| Y=-1\right)P\left(X^{(2)}=S \middle| Y=-1\right) &= \frac{7}{17}\cdot\frac{3}{9}\cdot\frac{4}{9} \\
&= \frac{28}{459} \\
&= 0.0610
\end{aligned}
$$

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

后验概率

第8章 朴素贝叶斯法

1. 朴素贝叶斯法的学习与分类
（数据、朴素贝叶斯法的基本公式、朴素贝叶斯分类器）
2. 朴素贝叶斯法的参数估计
（极大似然估计、学习与分类算法、贝叶斯估计）

*《机器学习方法》第4章

第8章 朴素贝叶斯法

1. 朴素贝叶斯法的学习与分类

（数据、朴素贝叶斯法的基本公式、朴素贝叶斯分类器）

2. 朴素贝叶斯法的参数估计

（极大似然估计、学习与分类算法、贝叶斯估计）

*《机器学习方法》第4章