北京邮電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

DSSC
Data Science and Service Center

朴素贝叶斯

分享人：pcc
2019/6/27

朴素贝叶斯

- 任务：分类
- 输入：x（点的特征向量）
- 输出：y（类别）

- 朴素贝叶斯是基于**贝叶斯定理**和特征**条件独立假设**的分类方法。

基本方法

- 训练数据集：

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

贝叶斯公式：P（Y|X）$= \frac{P(X,Y)}{P(X)} = \frac{P(Y)P(X|Y)}{P(X)}$

- 朴素贝叶斯通过训练数据集学习联合概率分布P(X,Y)，
  - 即先验概率分布：$P(Y=c_k)$，$k=1,2,\cdots,K$
  - 及条件概率分布：

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

贝叶斯定理

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

- 分子：**P(X=x, Y=Ck)**　　联合概率
- 分母：**P(X=x)**　　根据全概率公式
- 分母*左式： **P(X=x, Y=Ck)**

全概率公式

$$P(B)=\sum_{i=1}^{n} P(A_i)P(B|A_i)$$

公式描述：　公式表示若事件A1，A2，…，An构成一个完备事件组且都有正概率，则对任意一个事件B都有公式成立。

基本方法

- 条件独立性假设：

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

- “朴素”贝叶斯名字由来，牺牲分类准确性。
- 贝叶斯定理：$P(Y=c_k\mid X=x)=\dfrac{P(X=x\mid Y=c_k)P(Y=c_k)}{\sum_k P(X=x\mid Y=c_k)P(Y=c_k)}$
- 代入上式：

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

基本方法

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

$P(Y=c_k \mid X=x)$ 表示把具有x特征的输入节点，分为Ck这个类别的概率

- 贝叶斯分类器： 找到使式子值最大的类别Ck （概率最大）

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

- 分母对所有c_k都相同： **P(X=x)**

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

朴素贝叶斯的参数估计

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

- 应用极大似然估计法估计相应的概率：
- 先验概率$P(Y=c_k)$的极大似然估计是：

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

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

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

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

梳理算法：输入输出

- 学习与分类算法Naïve Bayes Algorithm：
- 输入：
  - 训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

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

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

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

- 输出：
  - x的分类

北京邮电大学 BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

梳理算法步骤

- 步骤
  - 1、计算先验概率和条件概率

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

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

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

梳理算法步骤

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

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

  - 3、确定x的类别

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

算法实例

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 $x=(2,S)^T$ 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

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

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

$$P(X^{(1)}=1\mid Y=1)=\frac{2}{9},\quad P(X^{(1)}=2\mid Y=1)=\frac{3}{9},\quad P(X^{(1)}=3\mid Y=1)=\frac{4}{9}$$

$$P(X^{(2)}=S\mid Y=1)=\frac{1}{9},\quad P(X^{(2)}=M\mid Y=1)=\frac{4}{9},\quad P(X^{(2)}=L\mid Y=1)=\frac{4}{9}$$

$$P(X^{(1)}=1\mid Y=-1)=\frac{3}{6},\quad P(X^{(1)}=2\mid Y=-1)=\frac{2}{6},\quad P(X^{(1)}=3\mid Y=-1)=\frac{1}{6}$$

$$P(X^{(2)}=S\mid Y=-1)=\frac{3}{6},\quad P(X^{(2)}=M\mid Y=-1)=\frac{2}{6},\quad P(X^{(2)}=L\mid Y=-1)=\frac{1}{6}$$

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

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

对于给定的 $x=(2,S)^T$ 计算：

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

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

因为 $P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)$ 最大，所以 $y=-1$.

贝叶斯估计

- **零概率问题**

在计算实例的概率时，如果某个实例x，在观察样本库（训练集）中没有出现过，会导致整个实例的概率结果是0。不能因为一个事件没有观察到就武断的认为该事件的概率是0。

ᘐ考虑：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。

ᘐ条件概率的贝叶斯估计：

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

ᘐ先验概率的贝叶斯估计：

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

- $\lambda=0$：极大似然估计
- $\lambda=1$：加法平滑也叫做拉普拉斯平滑。

假定训练样本很大时，每个分量x的计数加1造成的估计概率变化可以忽略不计，但可以方便有效的避免零概率问题。

- Sj为X的第j个特征的取值集合大小
- k为Y的取值集合大小

算法实例

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 $x=(2,S)^T$ 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

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

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：

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

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

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

$$P(X^{(1)}=1\mid Y=-1)=\frac{4}{9},\quad P(X^{(1)}=2\mid Y=-1)=\frac{3}{9},\quad P(X^{(1)}=3\mid Y=-1)=\frac{2}{9}$$

$$P(X^{(2)}=S\mid Y=-1)=\frac{4}{9},\quad P(X^{(2)}=M\mid Y=-1)=\frac{3}{9},\quad P(X^{(2)}=L\mid Y=-1)=\frac{2}{9}$$

对于给定的 $x=(2,S)^T$ 计算：

$$P(Y=1)P(X^{(1)}=2\mid Y=1)P(X^{(2)}=S\mid Y=1)=\frac{10}{17}\cdot\frac{4}{12}\cdot\frac{2}{12}=\frac{5}{153}=0.0327$$

$$P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)=\frac{7}{17}\cdot\frac{3}{9}\cdot\frac{4}{9}=\frac{28}{459}=0.0610$$

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

条件概率的贝叶斯估计：

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

先验概率的贝叶斯估计：

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

- Sj为X的第j个特征的取值集合大小
- k为Y的取值集合大小

总结

- 任务：分类
- 假设：条件独立性
- 策略：朴素贝叶斯法利用贝叶斯定理与学到的联合概率模型进行分类预测。

$$P(Y|X)=\frac{P(X,Y)}{P(X)}=\frac{P(Y)P(X|Y)}{\sum_Y P(Y)P(X|Y)}$$

将输入x分到后验概率最大的类y。

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

- 零概率问题：平滑策略

DSSC
Data Science and Service Center

谢谢！