人工智能的数学基础

华东师范大学 数学科学学院 黎芳(教授) 2019年9月23日

Chapter 4 朴素贝叶斯法（Naive Bayes）

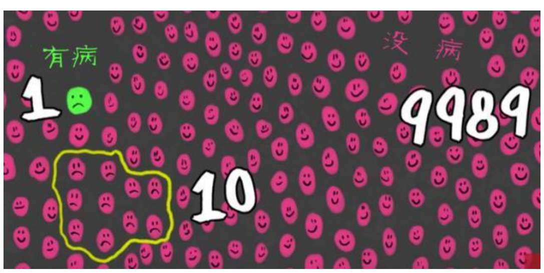
Table of Contents

4.0 贝叶斯公式直观理解

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

假设，有一种叫做「叶贝死」的病，人群中得病概率是万分之一，即 0.0001。然后，有一种测试可以检测你是否患有「叶贝死」病，准确率为 99.9%。你做了一次测试，结果被告知得病了！ 问真正的病的可能性是多少?



4.1 朴素贝叶斯法的学习与分类

4.1.1 基本方法

- 输入空间： $\mathcal{X}\subseteq\mathbf{R}^n$ 为n维向量的集合
- 输出空间： $\mathcal{Y}=\{c_1,c_2,\cdots,c_K\}$ 为类标集合
- 输入特征向量： $x\in\mathcal{X}$
- 输出类标记(class label)： $y\in\mathcal{Y}$

- X, Y 是定义在输入和输出空间上的随机变量
- $P(X, Y)$：X, Y 的联合概率分布
- 训练数据集：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ 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 $P(X, Y)$，即学习

- 先验概率分布：$P(Y = c_k), \quad k = 1, 2, \cdots, K$
- 条件概率分布：$P(X = x|Y = c_k) = P\left(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)}|Y = c_k\right), \quad k = 1, 2, \cdots, K$

注意：条件概率含指数级别的参数 $K\prod_{j=1}^{n} S_j$，S_j 表示 $x^{(j)}$ 取值的个数.

朴素贝叶斯法对条件概率分布做了独立性假设：

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

朴素贝叶斯法实际上学到的是数据生成的机制，所以属于生成模型.

运用贝叶斯方法对 x 进行分类（后验概率最大化）：

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

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

$$\Leftrightarrow y = \arg\max_{c_i} P(Y = c_k)\prod_j P\left(X^{(j)} = x^{(j)}|Y = c_k\right)$$

4.1.2 后验概率最大化的含义

后验概率最大化⇔期望风险最小化

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

期望风险函数：$R_{\exp}(f) = E[L(Y, f(X))] = E_X\sum_{k=1}^{K}[L(c_k, f(X))]P(c_k|X)$（条件期望）

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

期望风险最小化推出后验概率最大化

4.2 朴素贝叶斯法的参数估计

4.2.1 极大似然估计

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1, 2, \cdots, K \qquad (4.8)$$

$$P\left(X^{(j)} = a_{jl} | Y = c_k\right) = \frac{\sum_{i=1}^{N} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right)}{\sum_{i=1}^{N} I(y_i = c_k)} \qquad (4.9)$$

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

$x_i^{(j)}$-第i个样本的第j个特征， a_{jl}-第j个特征可能的第l个取值， I-指示函数

(4.8)的证明:

设$\theta_k = P(Y = c_k), k = 1, 2, \ldots, K$， $I_k = \sum_{i=1}^{N} I(y_i = c_k)$

$$L(\theta_1, \theta_2, \cdots, \theta_K) = \prod_{i=1}^{N} P(y_i) = \prod_{k=1}^{K} \theta_k^{I_k}$$

其中$\sum_{k=1}^{K} \theta_k = 1, \sum_{k=1}^{K} I_k = N$.

$$l(\theta) = \log L(\theta) = \sum_{k=1}^{K} I_k \log \theta_k$$

对它求导，利用约束条件，求使导数为0的θ值。

拉格朗日函数$\mathscr{L} = \sum_{k=1}^{K}\{I_k \log \theta_k\} + \gamma(\sum_{k=1}^{K}\theta_k - 1)$

$\Rightarrow \frac{\partial \ \mathscr{L}}{\partial \ \theta_k} = \frac{I_k}{\theta_k} + \gamma = 0,\ \theta_k = -\frac{I_k}{\gamma}$

$\Rightarrow \sum_{k=1}^{K}\theta_k = -\frac{1}{\gamma}\sum_{k=1}^{K} I_k = 1, \gamma = -N$

$\Rightarrow \theta_k^{MLE} = \frac{I_k}{N}$

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

(4.9)的证明:

$$P\left(X^{(j)} = a_{jl} | Y = c_k\right) = \frac{P\left(X^{(j)} = a_{jl}, Y = c_k\right)}{P(Y = c_k)}$$

同(4.8)的证明可得分子的估计

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

再利用(4.8)，可得(4.9).

算法**4.1** 朴素贝叶斯算法（naive **Bayes** algorithm）

输入：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})$，$x_j \in \{a_{j1}, a_{j2}, \cdots, a_{jS_j}\}, y_i \in \{c_1, c_2, \cdots, c_K\}$.

输出：x的分类

- (1)计算先验概率及条件概率
- (2)对输入实例x，计算$P(Y=c_k)\prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} | Y = c_k\right)$， $k = 1, 2, \cdots, K$
- (3)确定实例X的分类$y = \arg\max_{c_k} P(Y = c_k)\prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} | Y = c_k\right)$

```
% Example 4.1
X1 = [1 1 1 1 1 2 2 2 2 2 3 3 3 3 3]';
X2 = ['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]';
A = unique(X1);
B = unique(X2);
C = unique(Y);
```

```
T = table(X1,X2,Y)
```

T = 15×3 table

	X1	X2	Y
1	1	S	-1
2	1	M	-1
3	1	M	1
4	1	S	1
5	1	S	-1
6	2	S	-1
7	2	M	-1
8	2	M	1
9	2	L	1
10	2	L	1

⋮

```
% prior probability
p1 = mean(T{:,3}==1); % p(Y=1)
p2 = mean(T{:,3}==-1); % p(Y=-1)

% conditional probability
p11 =  sum((T{:,1}==1).*(T{:,3}==1))./sum(T{:,3}==1);% p(X1=1|Y=1)
p21 = sum((T{:,1}==2).*(T{:,3}==1))./sum(T{:,3}==1);% p(X1=2|Y=1)
p31 = sum((T{:,1}==3).*(T{:,3}==1))./sum(T{:,3}==1);% p(X1=3|Y=1)
ps1 = sum((T{:,2}=='S').*(T{:,3}==1))./sum(T{:,3}==1);%p(X2=S|Y=1)
pm1 = sum((T{:,2}=='M').*(T{:,3}==1))./sum(T{:,3}==1);%p(X2=M|Y=1)
pl1 = sum((T{:,2}=='L').*(T{:,3}==1))./sum(T{:,3}==1);%p(X2=L|Y=1)

p12 = sum((T{:,1}==1).*(T{:,3}==-1))./sum(T{:,3}==-1);% p(X1=1|Y=-1)
p22 = sum((T{:,1}==2).*(T{:,3}==-1))./sum(T{:,3}==-1);% p(X1=2|Y=-1)
p32 = sum((T{:,1}==3).*(T{:,3}==-1))./sum(T{:,3}==-1);% p(X1=3|Y=-1)
ps2 = sum((T{:,2}=='S').*(T{:,3}==-1))./sum(T{:,3}==-1);%p(X2=S|Y=-1)
pm2 = sum((T{:,2}=='M').*(T{:,3}==-1))./sum(T{:,3}==-1);%p(X2=M|Y=-1)
pl2 = sum((T{:,2}=='L').*(T{:,3}==-1))./sum(T{:,3}==-1);%p(X2=L|Y=-1)

% x = (2,S)^T
px1 = p1*p21*ps1
```

px1 = 0.0222

```
px2 = p2*p22*ps2
```

px2 = 0.0667

```
% y = -1
```

4.2.3 贝叶斯估计

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。解决办法：使用贝叶斯估计

$$P_\lambda\left(X^{(j)}=a_{jl}|Y=c_k\right)=\frac{\sum_{i=1}^{N} I\left(x_i^{(j)}=a_{jl}, y_i=c_k\right)+\lambda}{\sum_{i=1}^{N} I\left(y_i=c_k\right)+S_j\lambda}，\lambda \geq 0 \quad (4.10)$$

$\lambda=1$称为拉普拉斯平滑（Laplacian smoothing）

显然有

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

先验概率的贝叶斯估计为

$$P_\lambda\left(Y=c_k\right)=\frac{\sum_{i=1}^{N} I\left(y_i=c_k\right)+\lambda}{N+K\lambda} \quad (4.11)$$

```
% Example 4.2
lambda = 1;
N = length(Y); % number of samples
K = 2; % number of classes
Sj = 3;
% prior probability
p1 = (sum(T{:,3}==1)+lambda)./(N+K*lambda); % p(Y=1)
p2 = (sum(T{:,3}==-1)+lambda)./(N+K*lambda); % p(Y=-1)
% conditional probability
p11 = (sum((T{:,1}==1).*(T{:,3}==1))+lambda)./(sum(T{:,3}==1)+Sj*lambda); % p(X1=1|Y=1)
p21 = (sum((T{:,1}==2).*(T{:,3}==1))+lambda)./(sum(T{:,3}==1)+Sj*lambda);% p(X1=2|Y=1)
p31 = (sum((T{:,1}==3).*(T{:,3}==1))+lambda)./(sum(T{:,3}==1)+Sj*lambda);% p(X1=3|Y=1)
ps1 = (sum((T{:,2}=='S').*(T{:,3}==1))+lambda)./(sum(T{:,3}==1)+Sj*lambda);%p(X2=S|Y=1)
pm1 = (sum((T{:,2}=='M').*(T{:,3}==1))+lambda)./(sum(T{:,3}==1)+Sj*lambda);%p(X2=M|Y=1)
pl1 = (sum((T{:,2}=='L').*(T{:,3}==1))+lambda)./(sum(T{:,3}==1)+Sj*lambda);%p(X2=L|Y=1)

p12 = (sum((T{:,1}==1).*(T{:,3}==-1))+lambda)./(sum(T{:,3}==-1)+Sj*lambda);% p(X1=1|Y=-1)
p22 = (sum((T{:,1}==2).*(T{:,3}==-1))+lambda)./(sum(T{:,3}==-1)+Sj*lambda);% p(X1=2|Y=-1)
p32 = (sum((T{:,1}==3).*(T{:,3}==-1))+lambda)./(sum(T{:,3}==-1)+Sj*lambda);% p(X1=3|Y=-1)
ps2 = (sum((T{:,2}=='S').*(T{:,3}==-1))+lambda)./(sum(T{:,3}==-1)+Sj*lambda);%p(X2=S|Y=-1)
pm2 = (sum((T{:,2}=='M').*(T{:,3}==-1))+lambda)./(sum(T{:,3}==-1)+Sj*lambda);%p(X2=M|Y=-1)
pl2 = (sum((T{:,2}=='L').*(T{:,3}==-1))+lambda)./(sum(T{:,3}==-1)+Sj*lambda);%p(X2=L|Y=-1)

% x = (2,S)^T
px1 = p1*p21*ps1
```

```
px1 = 0.0327
```

```
px2 = p2*p22*ps2
```

```
px2 = 0.0610
```

```
% y = -1
```

作业

习题**1** 用贝叶斯估计法推出朴素贝叶斯法中的概率估计公式 (4.10)及公式 (4.11).

习题2 用朴素贝叶斯法推测一辆Red Domestic SUV是否会被盗.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes