HW3

asandstar@github

2022 年 10 月 28 日

1 简述估计后验概率的两种策略。

(1) 判别式模型 ($discriminative\ \ models$)：给定数据 $\boldsymbol{x}$, 通过直接拟合 $P(c|\boldsymbol{x})$ 预测 $\boldsymbol{x}$
换言之，利用正负例和分类标签，关注判别模型的边缘分布，不考虑 x 与 y 间的联合分布。
目标函数直接得到分类准确率。
例如：决策树、BP 神经网络、支持向量机
(2) 生成式模型 ($generative\ \ models$)：先对联合概率分布 $P(\boldsymbol{x}, c)$ 建模, 再据此得到 $P(c|\boldsymbol{x})$

对生成式模型，考虑 $P(c|\boldsymbol{x}) = \dfrac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$

由贝叶斯定理，$P(c|\boldsymbol{x}) = \dfrac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}$

$P(c|\boldsymbol{x})$: 后验概率
$P(c)$: 类“先验”概率
$P(\boldsymbol{x}|c)$: 样本 $\boldsymbol{x}$ 相对类标记 c 的类条件概率，即“似然”
$P(\boldsymbol{x})$: 用于归一化的“证据”因子。（对给定样本 x，该因子与类标记无关）
估计 $P(c|\boldsymbol{x})$ 的问题 $\rightarrow$ 基于训练数据估计先验 $P(c)$ 和似然 $P(\boldsymbol{x}|c)$
类先验 $P(c)$：样本空间中各类样本占比。
由大数定律，训练集包含足够独立同分布样本，P(c) 由各类样本出现频率估计
似然 $P(\boldsymbol{x}|c)$：涉及关于 $\boldsymbol{x}$ 所有属性的联合概率，难以按出现频率直接估计
常用策略：假定其由某种确定的概率分布，再基于训练样本估计概率分布的参数 (如极大似然估计 MLE)
D_c：训练中第 c 类样本组合的数据集合
设样本独立，则 θ_c 对 D_c 的似然：

$$P(D_c|\theta_c) = \prod_{x \in D_c} P(\boldsymbol{x}|\theta_c)$$

对数似然：

$$LL(\theta_c) = log\, P(D_c|\theta_c) = \sum_{x \in D_c} P(\boldsymbol{x}|\theta_c)$$

参数 θ_c 的极大似然估计 $\hat{\theta}_c$:

$$\hat{\theta}_c = arg\max_{\theta_c} LL(\theta_c)$$

2 **根据下表数据，使用朴素贝叶斯分类器分别判别（红色、圆形、大苹果）和（青色、非规则形状、小苹果）是否为好果（注：可使用拉普拉斯修正）。**

大小	**颜色**	**形状**	**好果**
小	青	非规则	否
大	红	非规则	是
大	红	圆	是
大	青	圆	否
大	青	非规则	否
小	红	圆	是
大	青	非规则	否
小	红	非规则	否
小	青	圆	否
大	红	圆	是

1. 估计类先验概率 $P(c)$

$P(\text{好果}=\text{是})=\dfrac{2}{5}=0.4$, $P(\text{好果}=\text{否})=\dfrac{3}{5}=0.6$

2. 估计每个属性的条件概率 $P(x_i|c)$

$P_{\text{青}\mid\text{是}}=P(\text{颜色}=\text{青}\mid\text{好果}=\text{是})=\frac{0}{4}=0$

显然不合理，进行“拉普拉斯修正”

3. 重新估计类先验概率 $P(c)$

$P(\text{好果}=\text{是})=\dfrac{4+1}{10+2}=\dfrac{5}{12}$,$P(\text{好果}=\text{否})=\dfrac{6+1}{10+2}=\dfrac{7}{12}$

4. 估计每个属性的条件概率 $P(x_i|c)$

$P_{\text{青}\mid\text{是}}=P(\text{颜色}=\text{青}\mid\text{好果}=\text{是})=\frac{0+1}{4+2}=\frac{1}{6}$

$P_{\text{青}\mid\text{否}}=P(\text{颜色}=\text{青}\mid\text{好果}=\text{否})=\frac{5+1}{6+2}=\frac{6}{8}$

$P_{\text{红}\mid\text{是}}=P(\text{颜色}=\text{红}\mid\text{好果}=\text{是})=\frac{4+1}{4+2}=\frac{5}{6}$

$P_{\text{红}\mid\text{否}}=P(\text{颜色}=\text{红}\mid\text{好果}=\text{否})=\frac{1+1}{6+2}=\frac{2}{8}$

$P_{\text{圆}\mid\text{是}}=P(\text{形状}=\text{圆}\mid\text{好果}=\text{是})=\frac{3+1}{4+2}=\frac{4}{6}$

$P_{\text{圆}\mid\text{否}}=P(\text{形状}=\text{圆}\mid\text{好果}=\text{否})=\frac{2+1}{6+2}=\frac{3}{8}$

$P_{\text{非规则}\mid\text{是}}=P(\text{形状}=\text{非规则}\mid\text{好果}=\text{是})=\frac{1+1}{4+2}=\frac{2}{6}$

$P_{\text{非规则}\mid\text{否}}=P(\text{形状}=\text{非规则}\mid\text{好果}=\text{否})=\frac{4+1}{6+2}=\frac{5}{8}$

$P_{\text{大}\mid\text{是}}=P(\text{大小}=\text{大}\mid\text{好果}=\text{是})=\frac{3+1}{4+2}=\frac{4}{6}$

$P_{\text{大}\mid\text{否}}=P(\text{大小}=\text{大}\mid\text{好果}=\text{否})=\frac{3+1}{6+2}=\frac{4}{8}$

$P_{\text{小}\mid\text{是}}=P(\text{大小}=\text{小}\mid\text{好果}=\text{是})=\frac{1+1}{4+2}=\frac{2}{6}$

$P_{小\mid否} = P(大小 = 小 \mid 好果 = 否) = \frac{3+1}{6+2} = \frac{4}{8}$

5. 判断是否为好果

对（红色、圆形、大苹果）

$P(好果 = 是) \times P_{红\mid是} \times P_{圆\mid是} \times P_{大\mid是} = \frac{5}{6} \times \frac{4}{6} \times \frac{4}{6} \approx 0.370$

$P(好果 = 否) \times P_{红\mid否} \times P_{圆\mid否} \times P_{大\mid否} = \frac{2}{8} \times \frac{3}{8} \times \frac{4}{8} \approx 0.047$

因为 0.563>0.028, 所以（红色、圆形、大苹果）是好果

对（青色、非规则形状、小苹果）

$P(好果 = 是) \times P_{青\mid是} \times P_{非规则\mid是} \times P_{小\mid是} = \frac{1}{6} \times \frac{2}{6} \times \frac{2}{6} \approx 0.019$

$P(好果 = 否) \times P_{青\mid否} \times P_{非规则\mid否} \times P_{小\mid否} = \frac{6}{8} \times \frac{5}{8} \times \frac{4}{8} \approx 0.234$

因为 0.019<0.234，所以（青色、非规则形状、小苹果）不是好果

3 使用半朴素贝叶斯分类器中的 SPODE 方法, 对于下表数据，假定 x_2 为超父，试预测 $x_1 = 1, x_2 = 1, x_3 = 0$ 时 $y = 1$ 的概率。

x_1	x_2	x_3	y
1	1	1	1
1	0	0	1
1	1	1	1
1	0	0	0
1	1	1	0
0	0	0	0
0	1	1	0
0	1	0	1
0	1	1	0
0	0	0	0

半朴素贝叶斯的基本想法：考虑一部分属性间的相互依赖信息，从而既不需进行完全联合概率计算，又不至于彻底忽略了比较强的属性依赖关系

独依赖估计 ($One-Dependent\,Estimator$, 简称 ODE): 半朴素贝叶斯分类器最常用的一种策略

独依赖: 假设每个属性在类别之外最多仅依赖于一个其他属性

$$P(c|\boldsymbol{x}) \propto P(c) \prod_{i=1}^{d} P(x_i|c, pa_i)$$

pa_i 是 x_i 的父属性，即属性 x_i 所依赖的属性

问题的关键 $\rightarrow$ 如何确定每个属性的父属性

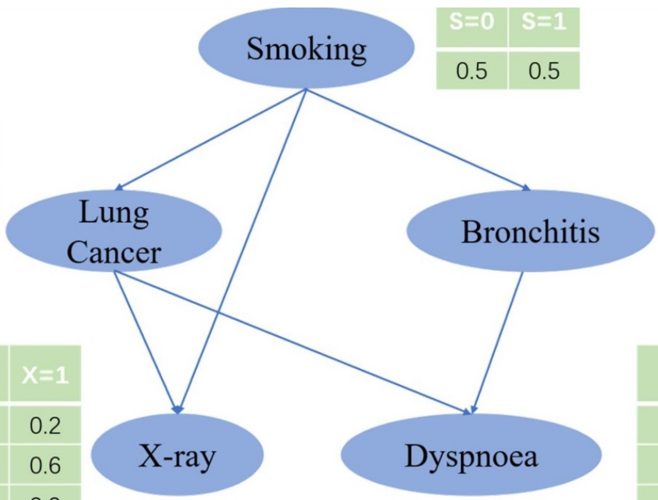
$SPODE(Super-Parent\,ODE)$ 方法: 假设所有属性都依赖于同一个属性“超父”(super-parent), 然后通过交叉验证等模型选择方法来确定超父属性

此时 $y = arg\,max\, P(c)P(x_j|c) \prod_{i=1, i\neq j}^{d} P(x_i|c, x_j)$, 其中 x_j 是超父

$P(y = 1) = \frac{4}{10}$

假定超父是 x_1，那么，对于 $y=1$ 的可能性有：
$P(x_1=1|y=1)=\frac{3}{4}$
$P(x_2=1|y=1,x_1=1)=\frac{2}{3}$
$P(x_3=0|y=1,x_1=1)=\frac{1}{3}$
$P_1=P(y=1)P(x_1|y=1)\prod_{i=1,i\neq j}^{d}P(x_i|y=1,x_1)=\frac{4}{10}\times\frac{3}{4}\times\frac{2}{3}\times\frac{1}{3}=\frac{1}{15}$
假定超父是 x_2，那么，对于 $y=1$ 的可能性有：
$P(x_2=1|y=1)=\frac{3}{4}$
$P(x_1=1|y=1,x_2=1)=\frac{2}{3}$
$P(x_3=0|y=1,x_2=1)=\frac{1}{3}$
$P_2=P(y=1)P(x_2|y=1)\prod_{i=1,i\neq j}^{d}P(x_i|y=1,x_2)=\frac{4}{10}\times\frac{3}{4}\times\frac{2}{3}\times\frac{1}{3}=\frac{1}{15}$
假定超父是 x_3，那么，对于 $y=1$ 的可能性有：
$P(x_3=0|y=1)=\frac{2}{4}$
$P(x_1=1|y=1,x_3=1)=\frac{1}{2}$
$P(x_2=1|y=1,x_3=1)=\frac{1}{2}$
$P_3=P(y=1)P(x_3|y=1)\prod_{i=1,i\neq j}^{d}P(x_i|y=1,x_3)=\frac{4}{10}\times\frac{2}{4}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{20}$
因此 x_1 和 x_2 均可作为超父，且 $x_1=1,x_2=1,x_3=0$ 时 $y=1$ 的概率为 $\frac{1}{15}$

4 给出如下贝叶斯网络，在一个人呼吸困难 (Dyspnoea) 的情况下，其抽烟 (Smoking) 的概率是多少。



S=0	S=1
0.5	0.5

S	L=0	L=1
0	0.45	0.55
1	0.2	0.8

S	B=0	B=1
0	0.5	0.5
1	0.3	0.7

S	L	X=0	X=1
0	0	0.8	0.2
0	1	0.4	0.6
1	0	0.1	0.9
1	1	0.2	0.8

C	B	D=0	D=1
0	0	0.9	0.1
0	1	0.3	0.7
1	0	0.2	0.8
1	1	0.1	0.9

由图可以得出 $P(S,L,B,X,D)=P(S)P(L|S)P(B|S)P(X|S,L)P(D|L,B)$
在一个人呼吸困难 (Dyspnoea) 的情况下，其抽烟 (Smoking) 的概率为

$$P(S=1|D=1)=\frac{P(S=1,D=1)}{P(D=1)}$$

$P(S=1,D=1)=P(S=1)\sum_{D=1}\sum_{B}P(B|S=1)\sum_{X}\sum_{L}P(L|S=1)P(X|S=1,L)P(D|L,B)$
$=0.5\times\sum_{D=1}\sum_{B}P(B|S=1)\sum_{X}\sum_{L}P(L|S=1)P(X|S=1,L)P(D|L,B)$

$= 0.5\times \sum_{D=1}\sum_{B} P(B|S=1)[P(L=0|S=1)P(X=0|S=1,L=0)P(D|L=1,B)$
$+P(L=0|S=1)P(X=1|S=1,L=0)P(D|L=1,B)$
$+P(L=1|S=1)P(X=0|S=1,L=1)P(D|L=1,B)$
$+P(L=1|S=1)P(X=1|S=1,L=1)P(D|L=1,B)]$
$= 0.5\times [P(B=0|S=1)P(L=0|S=1)P(X=0|S=1,L=0)P(D=1|L=1,B=0)$
$+P(B=1|S=1)P(L=0|S=1)P(X=0|S=1,L=0)P(D=1|L=1,B=1)$
$+P(B=0|S=1)P(L=0|S=1)P(X=1|S=1,L=0)P(D=1|L=1,B=0)$
$+P(B=1|S=1)P(L=0|S=1)P(X=1|S=1,L=0)P(D=1|L=1,B=1)$
$+P(B=0|S=1)P(L=1|S=1)P(X=0|S=1,L=1)P(D=1|L=1,B=0)$
$+P(B=1|S=1)P(L=1|S=1)P(X=0|S=1,L=1)P(D=1|L=1,B=1)$
$+P(B=0|S=1)P(L=1|S=1)P(X=1|S=1,L=1)P(D=1|L=1,B=0)$
$+P(B=1|S=1)P(L=1|S=1)P(X=1|S=1,L=1)P(D=1|L=1,B=1)]$
$= 0.5\times [0.3\times 0.2\times 0.1\times 0.8 + 0.7\times 0.2\times 0.1\times 0.9 + 0.3\times 0.2\times 0.9\times 0.8 + 0.7\times 0.2\times 0.9\times 0.9$
$+ 0.3\times 0.8\times 0.2\times 0.8 + 0.7\times 0.8\times 0.2\times 0.9 + 0.3\times 0.8\times 0.8\times 0.8 + 0.7\times 0.8\times 0.8\times 0.9]$
$= 0.435$

$P(B=1) = P(B=1|S)\times P(S) = P(B=1|S=0)\times P(S=0) + P(B=1|S=1)\times P(S=1)$
$= 0.5\times 0.5 + 0.7\times 0.5 = 0.6$
$P(B=0) = P(B=0|S)\times P(S) = P(B=0|S=0)\times P(S=0) + P(B=0|S=1)\times P(S=1)$
$= 0.5\times 0.5 + 0.3\times 0.5 = 0.4$
$P(L=1) = P(L=1|S)\times P(S) = P(L=1|S=0)\times P(S=0) + P(L=1|S=1)\times P(S=1)$
$= 0.55\times 0.5 + 0.8\times 0.5 = 0.675$
$P(L=0) = P(L=0|S)\times P(S) = P(L=0|S=0)\times P(S=0) + P(L=0|S=1)\times P(S=1)$
$= 0.45\times 0.5 + 0.2\times 0.5 = 0.325$

$P(D=1) = P(D=1|L,B)\times P(L,B) = P(D=1|L,B)\times P(L)\times P(B)$
$= P(D=1|L=0,B=0)\times P(L=0)\times P(B=0) + P(D=1|L=0,B=1)\times P(L=0)\times P(B=1)$
$+P(D=1|L=1,B=0)\times P(L=1)\times P(B=0) + P(D=1|L=1,B=1)\times P(L=1)\times P(B=1)$
$= 0.1\times 0.325\times 0.4 + 0.7\times 0.325\times 0.6$
$+0.8\times 0.675\times 0.4 + 0.9\times 0.675\times 0.6$
$= 0.73$

$P(S=1|D=1) = \frac{P(S=1,D=1)}{P(D=1)} = \frac{0.435}{0.73} \approx 0.596$
综上，一个人呼吸困难的情况下，其抽烟的概率是 0.596