

# 机器学习导论

## 习题五

参考解答

2017 年 6 月 4 日

### 1 [25pts] Bayes Optimal Classifier

试证明在二分类问题中，但两类数据同先验、满足高斯分布且协方差相等时，LDA可产生贝叶斯最优分类器。

**Solution.** 令  $g_i(\mathbf{x}) = \ln(P(c_i)P(\mathbf{x}|c_i))$ , 其中  $y \in \{c_0, c_1\}$ ,  $p(x|c_i) \sim \mathcal{N}(\mu_i, \Sigma)$ . 可得,

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma^{-1}(\mathbf{x} - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| + \ln P(c_i). \quad (1.1)$$

因此, 贝叶斯最优分类器为  $f(\mathbf{x}) = g_0(\mathbf{x}) - g_1(\mathbf{x})$ , 即

$$f(\mathbf{x}) = (\Sigma^{-1}(\mu_0 - \mu_1))^T \mathbf{x} + b. \quad (1.2)$$

其中,  $b = -\frac{1}{2}\mu_0^T \Sigma^{-1} \mu_0 + \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1$ . 式1.2与书中3.39一致, 证毕.

## 2 [25pts] Naive Bayes

考虑下面的400个训练数据的数据统计情况，其中特征维度为2 ( $\mathbf{x} = [x_1, x_2]$ )，每种特征取值0或1，类别标记 $y \in \{-1, +1\}$ 。详细信息如表1所示。

根据该数据统计情况，请分别利用直接查表的方式和朴素贝叶斯分类器给出 $\mathbf{x} = [1, 0]$ 的测试样本的类别预测，并写出具体的推导过程。

表 1: 数据统计信息

$x_1$	$x_2$	$y = +1$	$y = -1$
0	0	90	10
0	1	90	10
1	0	51	49
1	1	40	60

**Solution.**

(1) 根据表1可知 $\mathbf{x} = [1, 0]$ ，预测类别为+1.

(2) 首先估计出类先验概率 $P(c)$ 和每个属性的条件概率 $P(x_i|c)$ :

$$P(y = +1) = \frac{90 + 90 + 51 + 40}{400} \approx 0.678 ,$$

$$P(y = -1) = \frac{10 + 10 + 49 + 60}{400} \approx 0.322 ,$$

$$P(x_1 = 1|y = +1) = \frac{51 + 40}{90 + 90 + 51 + 40} \approx 0.336 ,$$

$$P(x_1 = 0|y = -1) = \frac{49 + 60}{10 + 10 + 49 + 60} \approx 0.845 ,$$

$$P(x_2 = 0|y = +1) = \frac{90 + 51}{90 + 90 + 51 + 40} \approx 0.520 ,$$

$$P(x_2 = 0|y = -1) = \frac{10 + 49}{10 + 10 + 49 + 60} \approx 0.457 .$$

于是，有

$$P(y = +1) \times P(x_1 = 1|y = +1) \times P(x_2 = 0|y = +1) \approx 0.118 ,$$

$$P(y = -1) \times P(x_1 = 1|y = -1) \times P(x_2 = 0|y = -1) \approx 0.124 .$$

由于 $0.118 < 0.124$ ，因此，朴素贝叶斯分类器将测试样本判别为-1.

### 3 [25pts] Bayesian Network

贝叶斯网(Bayesian Network)是一种经典的概率图模型，请学习书本7.5节内容回答下面的问题：

- (1) [5pts] 请画出下面的联合概率分布的分解式对应的贝叶斯网结构：

$$\Pr(A, B, C, D, E, F) = \Pr(A) \Pr(B) \Pr(C) \Pr(D|A) \Pr(E|A) \Pr(F|B, D) \Pr(G|D, E)$$

- (2) [5pts] 请写出图1中贝叶斯网结构的联合概率分布的分解表达式。

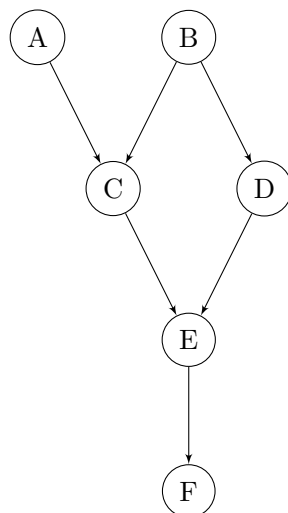


图 1: 题目3-(2)有向图

- (3) [15pts] 基于第(2)问中的图1, 请判断表格2中的论断是否正确，只需将下面的表格填完整即可。

表 2: 判断表格中的论断是否正确

序号	关系	True/False	序号	关系	True/False
1	$A \perp\!\!\!\perp B$		7	$F \perp\!\!\!\perp B C$	
2	$A \perp\!\!\!\perp B C$		8	$F \perp\!\!\!\perp B C, D$	
3	$C \perp\!\!\!\perp D$		9	$F \perp\!\!\!\perp B E$	
4	$C \perp\!\!\!\perp D E$		10	$A \perp\!\!\!\perp F$	
5	$C \perp\!\!\!\perp D B, F$		11	$A \perp\!\!\!\perp F C$	
6	$F \perp\!\!\!\perp B$		12	$A \perp\!\!\!\perp F D$	

**Solution.**

- (1) 对应的贝叶斯网结构如图2所示。

(2)  $\Pr(A, B, C, D, E, F) = \Pr(A) \Pr(B) \Pr(C|A, B) \Pr(D|B) \Pr(E|C, D) \Pr(F|E)$

- (3) 注：每个子问题2分。

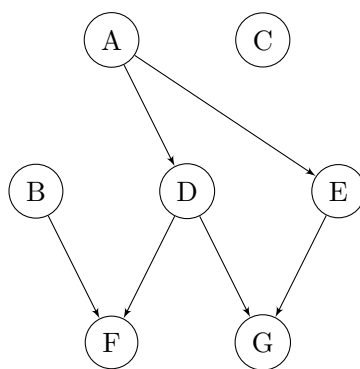


图 2: 贝叶斯网络

表 3: 判断表格中的论断是否正确

序号	关系	True/False	序号	关系	True/False
1	$A \perp\!\!\!\perp B$	True	7	$F \perp B C$	False
2	$A \perp B C$	False	8	$F \perp B C, D$	True
3	$C \perp\!\!\!\perp D$	False	9	$F \perp B E$	True
4	$C \perp D E$	False	10	$A \perp\!\!\!\perp F$	False
5	$C \perp D B, F$	False	11	$A \perp F C$	False
6	$F \perp\!\!\!\perp B$	False	12	$A \perp F D$	False

## 4 [25pts] Naive Bayes in Practice

请实现朴素贝叶斯分类器，同时支持离散属性和连续属性。详细编程题指南请参见链接：[http://lamda.nju.edu.cn/ml2017/PS5/ML5\\_programming.html](http://lamda.nju.edu.cn/ml2017/PS5/ML5_programming.html)。

### Solution.

main代码能够运行得10分，评测代码运行后，根据大家普遍的情形，0.5以上精度得25分，0.4~0.5之间得20分，0.4以下得15分。

编程题的反馈和参考解答见页面 [http://lamda.nju.edu.cn/ml2017/ml\\_faq.html](http://lamda.nju.edu.cn/ml2017/ml_faq.html) 中的更新内容。