机器学习导论
习题五

Boardwell, Nanjing University

2017 年 7 月 10 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.
Let $g_i(x) = ln(P(c_i)P(x|c_i))$, in which $y \in c_0, c_1$, $p(x|c_i)$ $N(\mu_i, \Sigma)$ and we achieve that:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - \frac{d}{2}ln2\pi - \frac{1}{2}ln|\Sigma| + lnP(c_i)$$

So that the bayes optimal classifier is $f(x) = g_0(x) - g_1(x)$, which is

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

in which $b = -\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1$. This equation is the same with the equation of LDA.

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.

Search the table directly:

Because of $P(y=1|\mathbf{x}=[1,0])=0.51$ and $P(y=-1|\mathbf{x}=[1,0])=0.49$, according to the probability of two conditions we predict that label is $y=+1$.

Naive Bayes:

$$\begin{aligned}
&P(y=1)=\frac{|D_1|}{|D|}=\frac{271}{400} && P(y=-1)=\frac{|D_{-1}|}{|D|}=\frac{129}{400}\\
&P(x_1=1|y=1)=\frac{|D_{1,x_1=1}|}{|D_1|}=0.3358 && P(x_1=1|y=-1)=\frac{|D_{-1,x_1=1}|}{|D_{-1}|}=0.8450\\
&P(x_2=0|y=1)=\frac{|D_{1,x_2=0}|}{|D_1|}=0.5203 && P(x_2=0|y=-1)=\frac{|D_{-1,x_2=0}|}{|D_{-1}|}=0.4574\\
&P(y=1|\mathbf{x}=[1,0])=0.4735 && P(y=-1|\mathbf{x}=[1,0])=0.4985
\end{aligned} \tag{2.1}$$

So that $P(y=1|\mathbf{x}=[1,0])<P(y=-1|\mathbf{x}=[1,0])$ and we can predict that label is $y=-1$.

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：

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

(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。

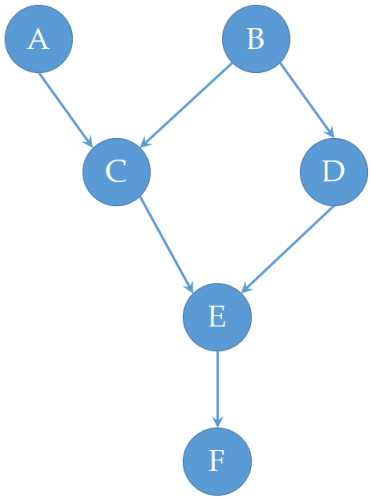

图 1: 题目3-(2)有向图

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

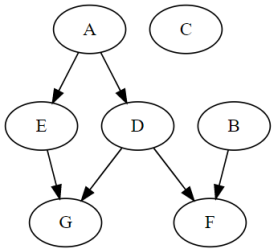

图 2: Bayes Net of problem 3-2

Solution.

(1): Solution is illustrated in Figure 2: Bayes Net of problem 3-2.

(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):

Solution is as above in the origin table.

The moral graph of this problem is as follows in Figure 3.

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`. 同时，请简要谈谈你的感想。实践过程中遇到了什么问题，你是如何解决的？

Solution.

The difficulty I meet is the σ of Gaussian distribution in continuous feature can often be zero because of the bad quality of data itself so that the program can hardly be able to run. My solution is to adjust those σ being a small non-zero constant. It realy works but I still think this solution is too direct and has no support explanation. I have no time to try to come up with a better solution but to use it.

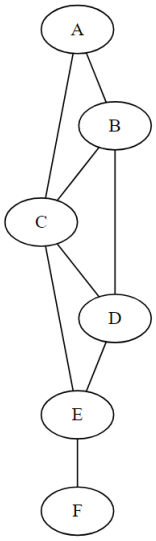

图 3: Moral graph of problem 3-3