机器学习导论
习题五

141242006, 袁帅, 141242006@smail.nju.edu.cn

2017 年 5 月 30 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. Suppose classes $c = \{1, 2\}$, and the corresponding data distributions are $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. In Linear Discriminant Analysis(LDA), we find the vector $\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ to map data points on one single line. Since

$$\boldsymbol{S}_w = \sum_{\boldsymbol{x} \in X_1} (\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_2} (\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^{\mathrm{T}} = (m-2)\Sigma \tag{1.1}$$

is the unbiased estimation, we obtain $\boldsymbol{w} = \frac{1}{m-2}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Therefore, the prediction for $\boldsymbol{x}$ by LDA would be based on whether $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ is closer to $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$ or $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_2$, which is equivalent to judge the sign of $\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})$, so the resulting classifier would be

$$\text{prediction}_{\text{LDA}}(\boldsymbol{x}) = \begin{cases} 1, & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}) > 0, \\ 2, & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}) < 0. \end{cases} \tag{1.2}$$

Now, consider the Bayes Optimal Classifier: both classes have the same prior distribution, so by deciding whether $p(y = 1|\boldsymbol{x}) > p(y = 2|\boldsymbol{x})$ or vice versa, we can focus on whether

$$\begin{aligned} \frac{p(y=1|\boldsymbol{x})}{p(y=2|\boldsymbol{x})} &= \frac{\frac{p(y=1)\cdot p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}}{\frac{p(y=2)\cdot p(\boldsymbol{x}|y=2)}{p(\boldsymbol{x})}} = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=2)} = \frac{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)]}{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)]} \\ &= \exp[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)] \end{aligned} \tag{1.3}$$

is greater or less than 1. That is equivalent as whether

$$\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)] = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}) \tag{1.4}$$

is positive or negative. Consequently, since both method's discriminant criteria are the same, LDA successfully generates a Bayes Optimal Classifier.

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. According the table above, the prior distribution $P(y = +1) = \frac{90+90+51+40}{400} = 0.6775$ and $P(y = -1) = \frac{10+10+49+60}{400} = 0.3225$. The idea of Bayes classifier is to compute the posterior distribution $P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x})}$, and since the denominator $P(\boldsymbol{x})$ is identical to both classes, we obtain the discriminant criterion as

$$h(\boldsymbol{x}) = \underset{c\in\{+1,-1\}}{\arg\min}\ P(y = c)P(\boldsymbol{x}|y = c) \tag{2.1}$$

Bayes Optimal Classifier If we directly refer to the table above, we find that more positive labels (51 vs. 49) are present in the $\boldsymbol{x} = [1, 0]$ case. The result is equivalent as the Bayes Optimal Classifier. The table indicates that the class-conditional distribution is given by $P(\boldsymbol{x} = [1, 0]|y = +1) = \frac{51}{90+90+51+40} = 0.1882$ and $P(\boldsymbol{x} = [1, 0]|y = -1) = \frac{49}{10+10+49+60} = 0.3798$, so we have

$$P(y = +1)P(\boldsymbol{x} = [1, 0]|y = +1) = 0.6775 \times 0.1882 = 0.1275, \tag{2.2}$$

$$P(y = -1)P(\boldsymbol{x} = [1, 0]|y = -1) = 0.3225 \times 0.3798 = 0.1225. \tag{2.3}$$

The prediction for $\boldsymbol{x} = [1, 0]$ would be $+1$.

Naïve Bayes Classifier Naïve Bayes Classifiers assume attributes are conditionally independent, i.e. $P(\boldsymbol{x} = [1, 0]|y = +1) = P(x_1 = 1|y = +1)P(x_2 = 0|y = +1) = \frac{51+40}{90+90+51+40} \cdot \frac{90+51}{90+90+51+40} = 0.1747$, $P(\boldsymbol{x} = [1, 0]|y = -1) = P(x_1 = 1|y = -1)P(x_2 = 0|y = -1) = \frac{49+60}{10+10+49+60} \cdot \frac{10+49}{10+10+49+60} = 0.3865$. Therefore,

$$P(y = +1)P(\boldsymbol{x} = [1, 0]|y = +1) = 0.6775 \times 0.1747 = 0.1184, \tag{2.4}$$

$$P(y = -1)P(\boldsymbol{x} = [1, 0]|y = -1) = 0.3225 \times 0.3865 = 0.1246. \tag{2.5}$$

The prediction for $\boldsymbol{x} = [1, 0]$ would be -1[1].

[1]If we do Laplacian correction, the result would be 0.1186 vs. 0.1244, which still yields a -1 prediction

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：

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

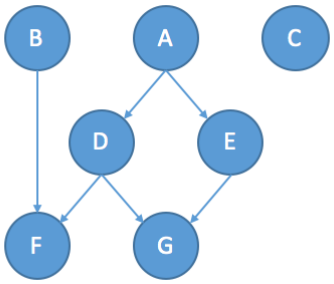

图 1: 题目 3-(1) 有向图

(2) **[5pts]** 请写出图3中贝叶斯网结构的联合概率分布的分解表达式。

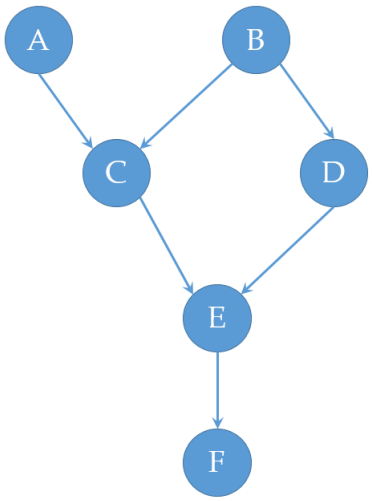

图 2: 题目 3-(2) 有向图

$$P(A, B, C, D, E, F) = P(A)P(B)P(C|A, B)P(D|B)P(E|C, D)P(F|E)$$

(3) **[15pts]** 基于第 (2) 问中的图3, 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	$F \perp\!\!\!\perp B \mid C$	False
2	$A \perp\!\!\!\perp B \mid C$	False	8	$F \perp\!\!\!\perp B \mid C, D$	True
3	$C \perp\!\!\!\perp D$	False	9	$F \perp\!\!\!\perp B \mid E$	True
4	$C \perp\!\!\!\perp D \mid E$	False	10	$A \perp\!\!\!\perp F$	False
5	$C \perp\!\!\!\perp D \mid B, F$	False	11	$A \perp\!\!\!\perp F \mid C$	False
6	$F \perp\!\!\!\perp B$	False	12	$A \perp\!\!\!\perp F \mid D$	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`.

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 training process is just computing various distributions(prior, class-conditional, etc.), so no random factors are introduced. In training the model, numerical issues are fatal: the product of probabilities would vanish; the standard deviation could be zero, making Gaussian distribution doesn't work. Therefore, we take logarithms and add a small portion ($\epsilon \cdot \max(\sigma)$, $\epsilon = 0.001$) to all σ.

Since I did vectorization in MATLAB coding, the program terminates in **0.5 second** (training + predicting) on my MacBook Pro. The consequent accuracy is about **68.85%**.