机器学习导论
习题五

181220010, 丁豪, 181220010@smail.nju.edu.cn

2020 年 6 月 1 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用**;

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在 LaTeX 模板中**第一页填写个人的姓名、学号、邮箱信息**;

(2) 本次作业需提交该 pdf 文件、问题 4 可直接运行的源码 (学号 _.py)、问题 4 的输出文件 (学号 _ypred.csv)，将以上三个文件压缩成 zip 文件后上传。zip 文件格式为**学号.zip**，例如 170000001.zip；pdf 文件格式为**学号 _ 姓名.pdf**，例如 170000001_张三.pdf。

(3) 未按照要求提交作业，或提交作业格式不正确，将会**被扣除部分作业分数**;

(4) 本次作业提交截止时间为**6 月 5 日 23:59:59**。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

[35 pts] Problem1 1 [PCA]

(1) **[5 pts]** 简要分析为什么主成分分析具有数据降噪能力;

(2) **[10 pts]** 试证明对于 N 个样本（样本维度 D>N）组成的数据集，主成分分析的有效投影子空间不超过 N-1 维;

(3) **[20 pts]** 对以下样本数据进行主成分分析，将其降到一行，要求写出其详细计算过程。

$$X = \begin{bmatrix} 2 & 3 & 3 & 4 & 5 & 7 \\ 2 & 4 & 5 & 5 & 6 & 8 \end{bmatrix} \tag{1}$$

Solution. 此处用于写解答 (中英文均可)
(1) 主成分分析首先对样本数据进行了一次投影，之后按照最近重构性/最大可分性，丢弃了部分特征值最小的对应坐标（投影到了低维空间）。这些特征值较小的坐标在数据受到噪声干扰时，其所对应的特征向量往往和噪声相关，将他们舍弃就可以在一定程度上起到降噪作用。
(2) 主成分分析的新坐标系 $W = (w_1, w_2, \ldots, w_{d'})$ 中的 w_i 为 XX^T 的第 i 大特征值对应特征向量，因此 $d' \leq rank(XX^T)$. 又因为样本进行了中心化, 极大线性无关组的向量数至多为 $N-1$, 因此 $rank(X) \leq min(N-1, D) = N-1, d' \leq rank(XX^T) \leq rank(X) \leq N-1$. 即有效投影子空间不超过 N-1 维。
(3) 首先中心化

$$X' = \begin{bmatrix} -2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3 \end{bmatrix}$$

其对应协方差矩阵

$$X'X'^T = \begin{bmatrix} 16 & 17 \\ 17 & 20 \end{bmatrix}$$

得到其特征值

$$\lambda_1 = 18 + \sqrt{293}, \lambda_2 = 18 - \sqrt{293}$$

λ_1 对应特征向量归一之后为 $\xi = (0.665, 0.747), W = (\xi)$
经过投影降维后的 $X'' = W^T X' = [-3.571 \ -1.412 \ -0.665 \ 0 \ 1.412 \ 4.236]$

[20 pts] Problem 3 [KNN]

已知 $err = 1 - \sum_{c \in Y} P^2(c|x)$ ， $err* = 1 - max_{c \in Y} P(c|x)$ 分别表示最近邻分类器与贝叶斯最优分类器的期望错误率，其中 Y 为类别总数，请证明:

$$err^* \leq err \leq err^*(2 - \frac{|Y|}{|Y|-1} * err^*)$$

Solution. 此处用于写解答 (中英文均可)
先证第一个不等号,

$$\begin{aligned}
\text{即证:} 1 - max_{c\in Y} P(c|x) &\leq 1 - \sum_{c\in Y} P^2(c|x) \\
\text{即证:} \sum_{c\in Y} P^2(c|x) &\leq max_{c\in Y} P(c|x) \\
\sum_{c\in Y} P^2(c|x) &\leq max_{c\in Y} P(c|x) \sum_{c\in Y} P(c|x) \\
&= max_{c\in Y} P(c|x)
\end{aligned}$$

得证.

再证第二个不等号, 记 $c^* = argmax_{c\in Y} P(c|x)$

$$\begin{aligned}
&\text{由柯西不等式}(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2) \geq (\sum_{i=1}^{n} a_i b_i)^2 \\
&\text{取} a_i = P(c_i|x), b_i = 1, n = |Y| - 1 \\
&\text{可得}(|Y|-1) \sum_{c\in Y-c^*} P^2(c|x) \geq (\sum_{c\in Y-c*} P(c|x))^2 = err^{*2} \\
\therefore err &= 1 - \sum_{c\in Y} P^2(c|x) \\
&\leq 1 - P^2(c^*|x) - \frac{1}{|Y|-1} err^{*2} \\
&= (1 - P(c^*|x))(1 + P(c^*|x)) - \frac{1}{|Y|-1} err^{*2} \\
&= err^*(1 + P(c^*|x) - \frac{1}{|Y|-1} err^*) \\
&= err^*(2 - err^* - \frac{1}{|Y|-1} err^*) \\
&= err^*(2 - \frac{|Y|}{|Y|-1} * err^*)
\end{aligned}$$

综上, $err^* \leq err \leq err^*(2 - \frac{|Y|}{|Y|-1} * err^*)$

[25 pts] Problem 2 [Naive Bayes Classifier]

通过对课本的学习，我们了解了采用“属性条件独立性假设”的朴素贝叶斯分类器。现在我们有如下表所示的一个数据集，其中 x_1 与 x_2 为特征，其取值集合分别为 $x_1 = \{-1, 0, 1\}$，$x_2 = \{B, M, S\}$，y 为类别标记，其取值集合为 $y = \{0, 1\}$：

表 1: 数据集

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	-1	-1	-1	-1	-1	0	0	0	0	0	1	1	1	1	1
x_2	B	M	M	B	B	B	M	M	S	S	S	M	M	S	S
y	0	0	1	1	0	0	0	1	1	1	1	1	1	1	0

(1) **[5pts]** 通过查表直接给出的 $x = \{0, B\}$ 的类别；

(2) **[10pts]** 使用所给训练数据，学习一个朴素贝叶斯试分类器，并确定 $x = \{0, B\}$ 的标记，要求写出详细计算过程；

(3) **[10pts]** 使用 “拉普拉斯修正”，即取 λ=1，再重新计算 $x = \{0, B\}$ 的标记，要求写出详细计算过程。

Solution. 此处用于写解答 (中英文均可)

(1) 查表可得 $x = \{0, B\}$ 的类别为 0.

(2) 首先估计类先验概率 $P(c)$, 显然有

$$P(y = 0) = \frac{6}{15} = 0.4$$
$$P(y = 1) = \frac{9}{15} = 0.6$$

然后为每个属性估计条件概率 $P(x_i|c)$:

$$P(x_1 = 0|y = 0) = \frac{2}{6} = \frac{1}{3}$$
$$P(x_1 = 0|y = 1) = \frac{3}{9} = \frac{1}{3}$$
$$P(x_2 = B|y = 0) = \frac{3}{6} = 0.5$$
$$P(x_2 = B|y = 1) = \frac{1}{9}$$

于是有:

$$P(y = 0) * P(x_1 = 0|y = 0) * P(x_2 = B|y = 0) \approx 0.067$$
$$P(y = 1) * P(x_1 = 0|y = 1) * P(x_2 = B|y = 1) \approx 0.022$$

由于 0.067>0.022，因此朴素贝叶斯分类器将测试样本 x 盘别为 0 类.

(3) 首先估计类先验概率 $P(c)$, 显然有

$$P(y = 0) = \frac{6 + 1}{15 + 2} = \frac{7}{17}$$
$$P(y = 1) = \frac{9 + 1}{15 + 2} = \frac{10}{17}$$

然后为每个属性估计条件概率 $P(x_i|c)$:

$$P(x_1 = 0|y = 0) = \frac{2+1}{6+3} = \frac{1}{3}$$
$$P(x_1 = 0|y = 1) = \frac{3+1}{9+3} = \frac{1}{3}$$
$$P(x_2 = B|y = 0) = \frac{3+1}{6+3} = \frac{4}{9}$$
$$P(x_2 = B|y = 1) = \frac{1+1}{9+3} = \frac{1}{6}$$

于是有:

$$P(y = 0) * P(x_1 = 0|y = 0) * P(x_2 = B|y = 0) \approx 0.061$$
$$P(y = 1) * P(x_1 = 0|y = 1) * P(x_2 = B|y = 1) \approx 0.033$$

由于 0.061>0.033，因此朴素贝叶斯分类器将测试样本 x 盘别为 0 类.

[20 pts] Problem 4 [KNN in Practice]

(1) **[20 pts]** 结合编程题指南，实现 KNN 算法。

Solution. 此处用于写解答 (中英文均可)