机器学习导论
习题五

171860027, 卫歆, 2523058044@qq.com

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.

1) 当数据受噪声影响时, 最小的特征值所对应的特征向量往往与噪声有关，而这个投影方向不包含任何有用的信息，造成数据维数比实际情况多出很多。使用 PCA 降维时将它们舍弃能在一定程度上起到去噪的效果.

2)PCA 的优化目标为 $\min -tr(W^T X X^T W)$ $s.t.$ $W^T W = I$, 使用拉格朗日乘子法可得 W 为 XX^T 前 d 大的特征值对应的特征向量所组成的投影矩阵。而协方差矩阵的秩不超过 N-1，所以 d 不超过 N-1 维。协方差矩阵的秩不超过 N-1 是因为：对样本集 X，协方差矩阵可以表示为：$C = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$，知道了 n-1 个 $(x_i - \bar{x})$，剩下一个的值就确定了，所以最多只有 n-1 阶。

3)1. 对所有样本进行中心化，

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

2. 计算

$$XX^T = \begin{bmatrix} 16 & 17 \\ 17 & 20 \end{bmatrix}$$

3. 计算协方差矩阵特征值为 35.11724 和 0.88276

4. 取特征值为 35.11724，对应的特征向量为

$$XX^T = \begin{bmatrix} -0.66451 \\ -0.74728 \end{bmatrix}$$

5. 投影矩阵为：

$$XX^T = \begin{bmatrix} 3.57086 & 1.41179 & 0.66451 & 0 & -1.41179 & -4.23537 \end{bmatrix}$$

[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：

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

Solution.

将 $\sum_{c\in Y} P^2(c|x)$ 看做一个加权平均值，$\sum_{c\in Y} P^2(c|x) \leq max_{c\in Y} P(c|x)$，故左边成立。

$$\begin{aligned} err &\leq 1 - max_{c\in Y} P(c|x)^2 \leq (1 - max_{c\in Y} P(c|x))(1 + max_{c\in Y} P(c|x)) \\ &\leq err^*(2 - err^*) \leq err^*(2 - \frac{|Y|}{|Y|-1} * err^*) \end{aligned}$$

右边不等式也成立。

[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.

1)y=0

2)$P(c=0) = \frac{6}{15},\ P(c=1) = \frac{9}{15}$

$P(x_1 = -1|c=0) = \frac{1}{2},\ P(x_1 = -1|c=1) = \frac{2}{9}$

$P(x_1 = 0|c=0) = \frac{1}{3},\ P(x_1 = 0|c=1) = \frac{1}{3}$

$P(x_1 = 1|c=0) = \frac{1}{6},\ P(x_1 = 1|c=1) = \frac{4}{9}$

$P(x_2 = B|c=0) = \frac{1}{2},\ P(x_2 = B|c=1) = \frac{1}{9}$

$P(x_2 = M|c=0) = \frac{1}{3},\ P(x_2 = M|c=1) = \frac{4}{9}$

$P(x_2 = S|c=0) = \frac{1}{6},\ P(x_2 = S|c=1) = \frac{4}{9}$

计算得：$P(x_1 = 0x_2 = B|c=0) = \frac{6}{15} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{15}$

$P(x_1 = 0x_2 = B|c=1) = \frac{9}{15} \cdot \frac{1}{3} \cdot \frac{1}{9} = \frac{1}{45}$

标记为 0.

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

$P(x_1 = 0|c=0) = \frac{2+1}{6+3} = \frac{1}{3},\ P(x_1 = 0|c=1) = \frac{3+1}{9+3} = \frac{1}{3}$

$P(x_2 = B|c=0) = \frac{3+1}{6+3} = \frac{4}{9},\ P(x_2 = B|c=1) = \frac{1+1}{9+3} = \frac{1}{6}$

计算得：$P(x_1 = 0x_2 = B|c=0) = \frac{7}{17} \cdot \frac{1}{3} \cdot \frac{4}{9} = \frac{28}{459}$

$P(x_1 = 0x_2 = B|c=1) = \frac{10}{17} \cdot \frac{1}{3} \cdot \frac{1}{6} = \frac{15}{459}$

标记为 0.

[20 pts] Problem 4 [KNN in Practice]

(1) **[20 pts]** 结合编程题指南，实现 KNN 算法。

Solution.