机器学习导论
习题五

18122028, 李佩然, 181220028@smail.nju.edu.cn

2020 年 5 月 24 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] Problem 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) 主成分分析相当于将数据投影到了一个超平面上，对于数据噪声，其对于所有测量数据都应该是均匀的，且与主要相关的特征无关，相当于在超平面的法向量上，因为进行主成分分析后，可以有效的将数据的维度中与噪声相关的去除，保留主要特征，达到数据降噪的目的。

(2) 对于题目所描述的数据集，其构成的数据集$X = (x_1, x_2, \cdots, x_N)$，在中心化后，我们有$\sum_{i=1}^{N} x_i = \mathbf{0}$，即矩阵的列向量组是一个线性相关的向量组，所以矩阵的列秩小于N，而由线性代数知识我们知道矩阵的秩小于等于列秩，而秩一定是个整数，所以$rank(X^T) = rank(X) \leqslant N-1$，而矩阵乘法是不会增加秩的，所以$rank(XX^T) \leqslant N-1$，所以其非零特征值不多于$N-1$个，对应的特征向量不超过$N-1$个，因而有效投影子空间不超过$N-1$维。得证。

(3) 首先我们进行中心化，得到

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

接着我们开始计算样本的协方差矩阵XX^T。

则协方差矩阵XX^T为

$$\begin{bmatrix} 16 & 17 \\ 17 & 20 \end{bmatrix} \tag{3}$$

然后我们对协方差矩阵XX^T做特征值分解。令特征多项式 $f(\lambda) = \begin{vmatrix} \lambda - 16 & -17 \\ -17 & \lambda - 20 \end{vmatrix} = \lambda^2 - 36\lambda + 31 = 0$

解得$\lambda_{1,2} = 18 \pm \sqrt{293} \approx 0.883, 35.117$显然，35.117是较大的那个特征值。最后，我们取最大的特征值对应的特征向量。设35.117对应的特征向量为$(x_1, x_2)^\top$，则有

$$\begin{bmatrix} 19.117 & -17 \\ -17 & 15.117 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{0}$$

令$x_1 = -0.665$，则$x_2 = -0.747$，$w_1 = (-0.665, -0.747)^\top$，$W = (w_1)$ 我们得到的投影矩阵W，接下来只需计算$Z = W^T X$即可

$$Z = X = \begin{bmatrix} -0.665 & -0.747 \end{bmatrix} \begin{bmatrix} -2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 3.571 & 1.412 & 0.665 & 0 & -1.412 & -4.236 \end{bmatrix}$$

[20 pts] Problem 2 [KNN]

已知$err = 1 - \sum_{c\in\mathcal{Y}} P^2(c|x)$ ，$err* = 1 - \max_{c\in\mathcal{Y}} P(c|x)$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$\mathcal{Y}$为类别总数，请证明：

$$err^* \leqslant err \leqslant err^*(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} \times err^*)$$

Solution.

a) 先证$err^* \leqslant err$

要证$err^* \leqslant err$

只要证$1 - \max_{c\in\mathcal{Y}} P(c|x) \leqslant 1 - \sum_{c\in\mathcal{Y}} P^2(c|x)$

只要证$\sum_{c\in\mathcal{Y}} P^2(c|x) \leqslant \max_{c\in\mathcal{Y}} P(c|x)$

$$\begin{aligned} \text{左边} &= \sum_{c\in\mathcal{Y}} P^2(c|x) = \sum_{c\in\mathcal{Y}} P(c|x)P(c|x) \leqslant \sum_{c\in\mathcal{Y}} (P(c|x) \max_{c\in\mathcal{Y}} P(c|x)) \\ &= \max_{c\in\mathcal{Y}} P(c|x) \sum_{c\in\mathcal{Y}} P(c|x) \quad \text{区域内的最大值是个常数} \\ &= \max_{c\in\mathcal{Y}} P(c|x) \quad \text{条件概率也是概率，由概率的规范性可得和为1} \end{aligned} \tag{4}$$

所以$\sum_{c\in\mathcal{Y}} P^2(c|x) \leqslant \max_{c\in\mathcal{Y}} P(c|x)$，$err^* \leqslant err$得证。

b) 下面证明$err \leqslant err^*(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \times err^*)$

$err = 1 - \sum_{c\in\mathcal{Y}} P^2(c|x)$

将满足$\max_{c\in\mathcal{Y}} P(c|x)$的$c$记作$c^*$，则有：

$$
\begin{aligned}
\text{左边} &= 1 - P^2(c^*|x) - \sum_{c\in\mathcal{Y}-\{c^*\}} P^2(c|x) = (1-P(c^*|x))(1+P(c^*|x)) - \sum_{c\in\mathcal{Y}-\{c^*\}} P^2(c|x) \\
&= err^*(2-err^*) - \sum_{c\in|\mathcal{Y}|-\{c^*\}} P^2(c|x) \sum_{i=1}^{|\mathcal{Y}|-1} (\frac{1}{\sqrt{|\mathcal{Y}|-1}})^2 \quad \text{由柯西不等式} \\
&\leqslant err^*(2-err^*) - (\sum_{c\in\mathcal{Y}-\{c^*\}} P(c|x)\frac{1}{\sqrt{|\mathcal{Y}|-1}})^2 \\
&\leqslant err^*(2-err^*) - \frac{1}{|\mathcal{Y}|-1}(\sum_{c\in\mathcal{Y}-\{c^*\}} P(c|x))^2 \quad \text{由概率的规范性} \\
&= err^*(2-err^*) - \frac{1}{|\mathcal{Y}|-1}(1-P(c^*|x))^2 \\
&= err^*(2-err^*) - \frac{1}{|\mathcal{Y}|-1}(err^*)^2 \\
&= 2err^* - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}(err^*)^2 \\
&= err^*(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \times err^*) = \text{右边}
\end{aligned}
\tag{5}
$$

到此，$err \leqslant err^*(2-\frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}\times err^*)$得证。

综合(a)(b)的结论，题目所要求的$err^* \leqslant err \leqslant err^*(2-\frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}\times err^*)$得证。

[25 pts] Problem 3 [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$x = \{0, B\}$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) 由表中编号6，我们知道$x = \{0, B\}$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$P(x_i|c)$:

$$P(x_1 = -1|y = 0) = \frac{3}{6} = 0.5$$

$$P(x_1 = 0|y = 0) = \frac{2}{6} \approx 0.333$$

$$P(x_1 = 1|y = 0) = \frac{1}{6} = \approx 0.167$$

$$P(x_1 = -1|y = 1) = \frac{2}{9} \approx 0.222$$

$$P(x_1 = 0|y = 1) = \frac{3}{9} \approx 0.333$$

$$P(x_1 = 1|y = 1) = \frac{4}{9} \approx 0.444$$

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

$$P(x_2 = M|y = 0) = \frac{2}{6} \approx 0.333$$

$$P(x_2 = S|y = 0) = \frac{1}{6} \approx 0.167$$

$$P(x_2 = B|y = 1) = \frac{1}{9} \approx 0.111$$

$$P(x_2 = M|y = 1) = \frac{4}{9} \approx 0.444$$

$$P(x_2 = S|y = 1) = \frac{4}{9} \approx 0.444$$

下面我们用训练所得的朴素贝叶斯分类器，对测试样例$x = \{0, B\}$进行分类：

$$P(y = 0) \times P(x_1 = 0|y = 0) \times P(x_2 = B|y = 0) \approx 0.067$$

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

由于$0.067 > 0.022$，因此，朴素贝叶斯分类器将测试样本$x = \{0, B\}$判别为类0。

(3) 首先估计类先验概率$P(c)$，显然有

$$P(y = 0) = \frac{6 + 1}{15 + 2} = 0.412$$

$$P(y = 1) = \frac{9 + 1}{15 + 2} = 0.588$$

然后，为每个属性估计条件概率$P(x_i|c)$:

$$P(x_1 = -1|y = 0) = \frac{3 + 1}{6 + 3} = 0.444$$

$$P(x_1 = 0|y = 0) = \frac{2 + 1}{6 + 3} \approx 0.333$$

$$P(x_1 = 1|y = 0) = \frac{1 + 1}{6 + 3} = \approx 0.222$$

$$P(x_1 = -1|y = 1) = \frac{2 + 1}{9 + 3} \approx 0.25$$

$$P(x_1 = 0|y = 1) = \frac{3 + 1}{9 + 3} \approx 0.333$$

$$P(x_1 = 1|y = 1) = \frac{4 + 1}{9 + 3} \approx 0.417$$

$$P(x_2 = B|y = 0) = \frac{3 + 1}{6 + 3} = 0.444$$

$$P(x_2 = M|y = 0) = \frac{2 + 1}{6 + 3} \approx 0.333$$

$$P(x_2 = S|y = 0) = \frac{1 + 1}{6 + 3} \approx 0.222$$

$$P(x_2 = B|y = 1) = \frac{1 + 1}{9 + 3} \approx 0.167$$

$$P(x_2 = M|y = 1) = \frac{4 + 1}{9 + 3} \approx 0.417$$

$$P(x_2 = S|y = 1) = \frac{4+1}{9+3} \approx 0.417$$

下面我们用训练所得经拉普拉斯修正后的朴素贝叶斯分类器，对测试样例$x = \{0, B\}$进行分类：

$$P(y = 0) \times P(x_1 = 0|y = 0) \times P(x_2 = B|y = 0) \approx 0.061$$

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

由于$0.061 > 0.033$，因此，朴素贝叶斯分类器将测试样本$x = \{0, B\}$判别为类0。

[20 pts] Problem 4 [KNN in Practice]

(1) **[20 pts]** 结合编程题指南，实现KNN算法。

Solution.

见文件夹其余文件