机器学习导论
习题五

171860607, 白晋斌, 810594956@qq.com

2020 年 5 月 30 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) 当 PCA 算法将原始数据从 d 维降到 d' 维的时候, 对应于最小的 $d-d'$ 个特征值的特征向量被舍弃了, 而当原始数据收到噪声影响时, 最小的特征值所对应的特征向量往往与噪声有关, 将他们舍弃能在一定程度上起到去噪的效果.

(2) 证明. 首先定义样本数据集 $\mathbf{X}$ 为 $(N \times D)$ 维的矩阵, 对样本数据集进行中心化后的矩阵为

$$Z = \begin{bmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \dots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \dots & -\frac{1}{N} \\ \dots & \dots & \dots & \dots \\ -\frac{1}{N} & -\frac{1}{N} & \dots & 1-\frac{1}{N} \end{bmatrix} X = RX$$

其中 R 是 $N \times N$ 的去均值矩阵,$rank(R) = N-1$.
这样, 协方差矩阵可表示为

$$\mathbf{S} = N^{-1}\mathbf{Z}^T\mathbf{Z}$$

对应的特征向量方程为

$$\frac{1}{N}\mathbf{Z}^T\mathbf{Z}u_i = \lambda_i u_i$$

已知 $rank(\mathbf{Z}^T\mathbf{Z}) = rank(\mathbf{Z}) = min(rank(\mathbf{R}), rank(\mathbf{X})) = min(N-1, rank(\mathbf{Z}))$, 有两个因素决定协方差的秩, 一是数据点个数 N, 二是数据中独立变量的个数 rank(X).
当数据个数很小以至于 N<D 时, 本质上是 $N-1$ 决定了协方差矩阵 $\mathbf{Z}^T\mathbf{Z}$ 的秩的上界, 此时, 大量的特征值 λ 为 0. 是否有更少的非 0 特征值取决于数据中独立变量的个数是否足够多.
设得到的非零特征值所对应的特征向量 u_i 有 k 个, 此时 $k \le N-1$.
将特征向量按对应特征值大小从左到右按列排列成矩阵

$$W = \begin{bmatrix} u_\alpha & u_\beta & \dots & u_\theta \end{bmatrix}$$

其中 $\mathbf{W}$ 是 $D \times k$ 的矩阵, 对 W 做归一化处理后得到 $\mathbf{W}'$.
投影到子空间后的数据为

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}'$$

其中 $\mathbf{Y}$ 是 $N \times k$ 的矩阵. 因为 $k \le N-1$, 这表明对于 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. □

我们也可以举个例子来说明, 确定一条直线 (1 维子空间) 需要至少 2 个点 (样本数量), 确定一个平面 (2 维子空间) 需要至少 3 个点 (样本数量), 确定 n-1 维子空间需要至少 n 个点 (样本数量), 也就是 n 个点 (样本数量) 最多确定 n-1 维子空间.

所以对于 N 个样本数量, 如果投影的子空间 M≥N 维, 这样的 M 维平面就有无穷多个, 因此 PCA 的有效投影子空间不超过 N-1 维.

(3) 将 X 的每一行 (代表一个特征) 进行零均值化 (中心化), 即减去这一行的均值,$x_i = x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$, 得到

$$X = \begin{bmatrix} 2-4 & 3-4 & 3-4 & 4-4 & 5-4 & 7-4 \\ 2-5 & 4-5 & 5-5 & 5-5 & 6-5 & 8-5 \end{bmatrix} = \begin{bmatrix} -2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3 \end{bmatrix}$$

求出协方差矩阵 $C = XX^T$(这里除或不除样本数量 n 或 $n-1$, 对求出的特征向量没有影响, 故为了简化计算我们选择不除)

$$C = \begin{bmatrix} -2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} -2 & -3 \\ -1 & -1 \\ -1 & 0 \\ 0 & 0 \\ 1 & 1 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} 16 & 17 \\ 17 & 20 \end{bmatrix}$$

求出协方差矩阵 C 的特征值及对应的特征向量 $CW = \lambda W$

$$|C - \lambda E| = \begin{vmatrix} 16-\lambda & 17 \\ 17 & 20-\lambda \end{vmatrix} = 0$$

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

把每个特征值代入 $(C - \lambda E)w = 0$, 得

$$w_1 = \begin{bmatrix} \frac{-2+\sqrt{293}}{17} \\ 1 \end{bmatrix}, w_2 = \begin{bmatrix} \frac{-2-\sqrt{293}}{17} \\ 1 \end{bmatrix}$$

将特征向量按对应特征值大小从上到下按行排列成矩阵, 取前 1 行组成矩阵

$$W = \begin{bmatrix} \frac{-2+\sqrt{293}}{17} & 1 \end{bmatrix}$$

对 W 做归一化处理, 得

$$W = \begin{bmatrix} \frac{\sqrt{293}-2}{\sqrt{586-4\sqrt{293}}} & \frac{17}{\sqrt{586-4\sqrt{293}}} \end{bmatrix}$$

$Y = WX$ 即为降维到 1 维后的数据

$$\begin{aligned} Y &= \begin{bmatrix} \frac{\sqrt{293}-2}{\sqrt{586-4\sqrt{293}}} & \frac{17}{\sqrt{586-4\sqrt{293}}} \end{bmatrix} \begin{bmatrix} -2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} \frac{-2\sqrt{293}-47}{\sqrt{586-4\sqrt{293}}} & -\frac{\sqrt{293}+15}{\sqrt{586-4\sqrt{293}}} & -\frac{\sqrt{293}-2}{\sqrt{586-4\sqrt{293}}} & 0 & \frac{\sqrt{293}+15}{\sqrt{586-4\sqrt{293}}} & \frac{3\left(\sqrt{293}+15\right)}{\sqrt{586-4\sqrt{293}}} \end{bmatrix} \\ &\approx \begin{bmatrix} -3.57086 & -1.41179 & -0.664514 & 0 & 1.41179 & 4.23537 \end{bmatrix} \end{aligned}$$

[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. 此处用于写解答 (中英文均可)

证明. 设

$$c^* = \underset{c\in Y}{\operatorname{argmax}} P(c|x)$$

则

$$P(c^*|x) = max_{c\in Y} P(c|x)$$

$$err^* = 1 - P(c^*|x)$$

先证 $err^* \le err$:
$\sum_{c\in Y} P^2(c|x)$ 可以看成是 $P(c|x)$ 的带权线性组合, 总权值为 1. 此时

$$\sum_{c\in Y} P^2(c|x) \le \left[\sum_{c\in Y} P(c|x)\right] P(c^*|x) = P(c^*|x)$$

$$1 - P(c^*|x) \le 1 - \sum_{c\in Y} P^2(c|x)$$

$$err^* \le err$$

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

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

当除了 $P(c^*|x)$ 以外的 $P(c|x)$ 全部相等时,$\sum_{c\in Y-c^*} P^2(c|x)$ 取最小值, 即

$$P_{c\in Y-c^*}(c|x) = \frac{1 - P(c^*|x)}{|Y|-1} = \frac{err^*}{|Y|-1}$$

$$\sum_{c\in Y-c^*} P^2(c|x) \ge (|Y|-1)\left(\frac{err^*}{|Y|-1}\right)^2 = \frac{(err^*)^2}{|Y|-1}$$

所以

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

综上,

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

□

[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) 由编号 6 的数据知,$x = \{0, B\}$ 的类别为 $y = 0$.

(2) 首先估计类先验概率 $P(y)$:

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

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

然后为每个属性估计条件概率 $P(x_i|y)$:

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

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

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

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

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

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

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

$$P(x_1 = 0|y = 1) = \frac{3}{9} = \frac{1}{3}$$
$$P(x_1 = 1|y = 1) = \frac{4}{9}$$
$$P(x_2 = B|y = 1) = \frac{1}{9}$$
$$P(x_2 = M|y = 1) = \frac{4}{9}$$
$$P(x_2 = S|y = 1) = \frac{4}{9}$$

于是有

$$P(y = 0) \times P(x_1 = 0|y = 0) \times P(x_2 = B|y = 0) = \frac{1}{15}$$
$$P(y = 1) \times P(x_1 = 0|y = 1) \times P(x_2 = B|y = 1) = \frac{1}{45}$$

由于 $\frac{1}{15} > \frac{1}{45}$, 因此朴素贝叶斯分类器将测试样本 $x = \{0, B\}$ 判别为 $y = 0$.

(3) 首先估计类先验概率 $P(y)$:

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

然后为每个属性估计条件概率 $P(x_i|y)$:

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

$$P(x_1 = -1|y = 1) = \frac{2+1}{9+3} = \frac{1}{4}$$
$$P(x_1 = 0|y = 1) = \frac{3+1}{9+3} = \frac{1}{3}$$
$$P(x_1 = 1|y = 1) = \frac{4+1}{9+3} = \frac{5}{12}$$
$$P(x_2 = B|y = 1) = \frac{1+1}{9+3} = \frac{1}{6}$$
$$P(x_2 = M|y = 1) = \frac{4+1}{9+3} = \frac{5}{12}$$

$$P(x_2 = S|y = 1) = \frac{4+1}{9+3} = \frac{5}{12}$$

于是有

$$P(y = 0) \times P(x_1 = 0|y = 0) \times P(x_2 = B|y = 0) = \frac{28}{459} \approx 0.061$$

$$P(y = 1) \times P(x_1 = 0|y = 1) \times P(x_2 = B|y = 1) = \frac{10}{306} \approx 0.033$$

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 $y = 0$.

[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)

(1) 代码详见 171860607.*py*. 关于 k 的选取, 我们在训练集上做十折交叉验证, 最终发现参数 $k = 7$ 时效果最好, 我们的 171860607*_ypred.csv* 以参数 $k = 7$ 预测结果并输出.