机器学习导论
习题五

学号, 作者姓名, 邮箱

2020 年 6 月 8 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。

$$X=\left[\begin{array}{cccccc}2 & 3 & 3 & 4 & 5 & 7 \\ 2 & 4 & 5 & 5 & 6 & 8\end{array}\right] \tag{1}$$

Solution. （解答用中英文皆可）

（1）从最大可分性解释，主成分分析（PCA）先将中心化后的 D 维数据投影到另一个同维度的特征空间中（该空间的一组正交基底为协方差矩阵的全部特征向量），并选择出前 d 个最大特征值所对应的特征向量组成基底张成子空间。通过方差分析，我们知道 λ_i=Var(y_i), 其中 Var(y_i) 是特征空间第 i 维的方差，λ_i 是第 i 个特征空间对应的特征值。因此，提取最大特征值对应特征向量的操作，就是在保存特征空间中方差最大的这些维度。通常认为，某一维拥有较大方差表征了该维具有较丰富的信息，而方差较小往往代表该维有效信息较少，更可能是噪声引起的微小波动。在这种考虑下，PCA 在变换后剔除掉了方差较小的特征空间维度，在降维过程中抑制了数据的噪声，因此具有数据降噪能力。

（2）证明：将数据集记为 X，则 $rank(X) \leq N$，因此协方差矩阵的秩

$$\operatorname{rank}\left(X X^{\top}\right) \leq \min \left\{\operatorname{rank}(X), \operatorname{rank}\left(X^{\top}\right)\right\} \leq N$$

则非零特征值不超过 N。将非零特征值以及最小的特征值删去不影响重构阈值，因此有效投影子空间不超过 N-*1* 维。

由于 PCA 首先会将样本中心化，因此，数据集中的任意 N-1 个样本都可以还原出完整的数据集，即只有 N-1 个样本是线性无关的。所以，中心化后的数据矩阵的秩至多为 N-1，这导致矩阵的行秩也至多 N-1, 即 D 个维度中至多有 N-1 个维度是线性无关的。为了使投影后的数据各个维度均线性无关，我们只能将数据投影到维度至多为 N-1 的子空间中，即有效投影子空间维数不超过 N-1。综上，主成分分析法有效投影子空间不超过 N-1 维，得证。

（3）主成分分析的一般步骤：

1）去平均值（即去中心化），即每一位特征减去各自的平均值;

2）计算协方差矩阵（注：这里除或不除样本数量 n 或 n-1，其实对求出的特征向量没有影响）;

3）用特征值分解方法求协方差矩阵的特征值与特征向量;

4）对特征值从大到小排序，选择其中最大的 k 个。然后将其对应的 k 个特征向量分别作为行向量组成特征向量矩阵 P;

5）将矩阵转换到 k 个特征向量构建的新空间中，即 Y=PY。

对该题而言：本矩阵为六个样本，每个样本 2 维特征：

（a）去中心化：$\mu=\binom{4}{5}$, 有 $X^*_{\cdot j}=X_{\cdot j}$ - μ,$X^*=\begin{pmatrix}-2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3\end{pmatrix}$;

（b）计算协方差矩阵，$M=X^*(X^*)^T=\begin{pmatrix}16 & 17 \\ 17 & 20\end{pmatrix}$;

（c）对 M 进行特征值分解：令 M 特征多项式 $f_M(\lambda)$=$\begin{vmatrix} \lambda\text{-}16 & 17 \\ 17 & \lambda\text{-}20 \end{vmatrix}$=$\lambda^2-36\lambda$+31=0, 解得 M 两个特征值 $\lambda_1 = 18+\sqrt{293}\approx 35.1172$,$\lambda_2$=18-$\sqrt{293}\approx$0.8828；分别求得它们所对应的一个单位特征向量为 ω_1=$\begin{pmatrix} 0.6645 \\ 0.7473 \end{pmatrix}$,$\omega_2$=$\begin{pmatrix} -0.7473 \\ 0.6645 \end{pmatrix}$；
（d）由于需要降为一维，只需取最大的特征值即可，即取 λ_1 对应的特征值，其相对应的特征向量为 ω_1=$\begin{pmatrix} 0.6645 \\ 0.7473 \end{pmatrix}$ 或 $\begin{pmatrix} -0.6645 \\ -0.7473 \end{pmatrix}$，取特征值 λ_1 所对应特征向量 ω_1 构造变换矩阵 W=（ω_1）;
（e）进行降维线性变换

$$Z{=}W^TX^*{=}\begin{pmatrix} -3.5709 & -1.4118 & -0.6645 & 0 & 1.4118 & 4.2354 \end{pmatrix}$$

或

$$Z{=}W^TX^*{=}\begin{pmatrix} 3.5709 & 1.4118 & 0.6645 & 0 & -1.4118 & -4.2354 \end{pmatrix}。$$

助教反馈:
P1.1 和 P1.2 比较简单，大家总体得分较高，不过部分同学依然没有答，或者敷衍了事，比较可惜；P1.3 是基本的计算题，不少同学在计算特征值和特征向量的时候出错；大部分是**计算出错，步骤过于简单**，虽然有无理数，可是没有规定不能用编程工具算数值。

[20 pts] Problem 2 [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. （解答用中英文皆可）
证明：记 c^*=$\underset{c\in Y}{\arg\max}$ P(c | x), 则 err^*=1-P(c*|x)。
先证明 $err \ge err^*$, 证明如下:

$$\begin{aligned} err &= 1 - \sum_{c\in Y} P^2(c|x) &(2)\\ &\ge 1 - \sum_{c\in Y}[P(c|x)P(c^*|x)] &(3)\\ &= 1 - P(c^*|x)\sum_{c\in Y} P(c|x) &(4)\\ &= 1 - P(c^*|x) &(5)\\ &= err^* &(6) \end{aligned}$$

即 $err \ge err^*$。

其次证明 $err \leq err^*(2\text{-}\frac{|Y|}{|Y|-1}\text{err}^*)$，证明如下:

$$\begin{aligned}
err &= 1 - \sum_{c \in Y} P^2(c|x) && (7)\\
&= 1 - P^2(c^*|x) - \sum_{c \in Y, c \neq c^*} P^2(c|x) && (8)\\
&= (2-1) - 2P(c^*|x) + 2P(c^*|x) - P^2(c^*|x) - \sum_{c \in Y, c \neq c^*} P^2(c|x) && (9)\\
&= 2(1 - P(c^*|x)) - (1 - P(c^*|x))^2 - \sum_{c \in Y, c \neq c^*} P^2(c|x) && (10)\\
&= 2err^* - (err^*)^2 - \sum_{c \in Y, c \neq c^*} P^2(c|x) && (11)
\end{aligned}$$

而由于 $f(x)=x^2$, 则由琴生不等式可知:

$$\begin{aligned}
\frac{\sum\limits_{c \in Y, c \neq c^*} P^2(c|x)}{|Y|-1} &\geq \left[\frac{\sum\limits_{c \in Y, c \neq c^*} P(c|x)}{|Y|-1}\right]^2 && (12)\\
&= \left[\frac{1 - P(c^*|x)}{|Y|-1}\right]^2 && (13)\\
&= \frac{(err^*)^2}{(|Y|-1)^2} && (14)
\end{aligned}$$

则可知

$$\sum_{c \in Y, c \neq c^*} P^2(c|x) \geq \frac{(err^*)^2}{|Y|-1}$$

因此

$$\begin{aligned}
err &= 2err^* - (err^*)^2 - \sum_{c \in Y, c \neq c^*} P^2(c|x) && (15)\\
&\leq (2 - err^*)err^* - \frac{(err^*)^2}{|Y|-1} && (16)\\
&= err^*(2 - err^* - \frac{err^*}{|Y|-1}) && (17)\\
&= err^*(2 - \frac{|Y|}{|Y|-1}err^*) && (18)
\end{aligned}$$

综上

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

得证。

助教反馈:

P2 是教材上的习题，大家总体做得比较好，被扣分的大部分是答题的**不够规范，公式排版混乱，引用格式不够严谨**等，希望大家引以为戒。

[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. （解答用中英文皆可）

（1）找到表中编号为 6 的的样本满足 x=$\left\{0,B\right\}$, 故 x=$\left\{0,B\right\}$ 的类别为 0。

（2）首先估计类先验概率 *P(c)*，有 P(y=0)=$\frac{6}{15}$, P(y=1)=$\frac{9}{15}$

然后，为数据集每个属性的属性值估计条件概率 P(x$_i$|c):

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

接下来对样本 x 查找训练好的条件概率表，得

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

故

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

由于 $\frac{1}{15} > \frac{1}{45}$，因此朴素贝叶斯分类器将样本 $x = \{0, B\}$ 类别标记判定为 0。

(3) 加入拉普拉斯修正后，因为 λ=1, 则有

$$\widehat{P}(c)=\frac{|D_c|+1}{D+N}$$
$$\widehat{P}(x_i|\mathrm{c})=\frac{|D_{c_i x_i}|+1}{|D_c|+N_i}$$

估计类先验概率 $P(c)$ 为：$\widehat{P}(\mathrm{y}=0)=\frac{7}{17}$，$\widehat{P}(\mathrm{y}=1)=\frac{10}{17}$。
为数据集每个属性的属性值估计修正后条件概率$\widehat{P}(\mathrm{x}_i|\mathrm{c})$：

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

接下来对样本 x 查找训练好的条件概率表，得

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

故

$$\widehat{P}(y=0|x_1=0,x_2=B)=\widehat{P}(y=0)\widehat{P}(x_1=0|y=0)\widehat{P}(x_2=B|y=0)=\frac{28}{459}\approx 0.0610$$
$$\widehat{P}(y=1|x_1=0,x_2=B)=\widehat{P}(y=1)\widehat{P}(x_1=0|y=1)\widehat{P}(x_2=B|y=1)=\frac{5}{153}\approx 0.0327$$

由于 $0.0610 > 0.0327$，因此朴素贝叶斯分类器将样本 $x = \{0, B\}$ 类别标记判定为 0。

助教反馈：

P3.1 直接查表即可；P3.2 和 p3.3，出错的同学很多，大部分体现在**计算**出错，对稍微复杂一些的计算**算错、步骤比较混乱**；另外，部分同学存在浮点数和分数混用的情况，不知道是出于何种想法，计算过程中用分数、最终答案用浮点数或者分数无可厚非，但是不建议计算过程中混用。

[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)

助教反馈：

P4 编程题比较简单，部分同学尝试了部分新的方法，效果比较好，给大佬们点赞。

总体而言，不少同学**没有仔细看文件提交要求**，在文件命名格式，latex 中个人信息填写方面没有按照要求进行，被额外扣了一些分，希望大家引以为戒；有些同学答题中用**拍照**的形式，非

常不建议以上行为，并且给予一定的惩罚；另外，部分同学的**排版和格式问题**比较大，希望引以为戒，之后能有所改进，不重视格式，内容也会很大程度上被打折扣或者被质疑，希望大家引起重视！