机器学习导论
习题五

171840708, 张逸凯, zykhelloha@gmail.com

2020 年 6 月 5 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;

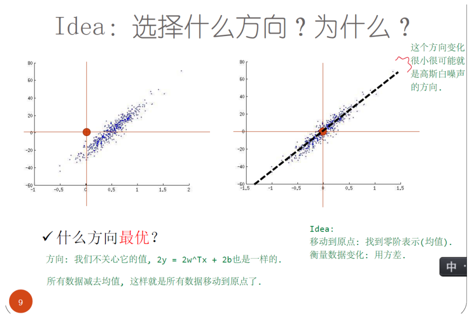
(3) **[20 pts]** 对以下样本数据进行主成分分析，将其降到一行，要求写出其详细计算过程。

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

Solution. 解:

(1) 需要数据降噪时, 在方差最大(较大特征值)的方向往往是数据本身的信息使他们分开, 但是在较小特征值的方向往往是数据的噪声, 同数据(信息)是不相关的, PCA舍弃较小特征值方向即可以在一定程度上降噪.

如下是一个简单的例子:



(2)

主成分分析的有效投影子空间维度与协方差矩阵非零特征值个数有关, 不妨令协方差矩阵为(课本10.17):

$$Cov(x) = \sum_i \boldsymbol{x_i}\boldsymbol{x_i^T} = \boldsymbol{X}\boldsymbol{X^T} \in \mathbb{R}^{D\times D}$$

以下证明一个引理:

- *lemma.* $\boldsymbol{X}\boldsymbol{X}^T$与$\boldsymbol{X}^T\boldsymbol{X}$有相同的非零特征值

  首先证明$r(XX^T) = r(X^TX) = r(X) = r(X^T)$:

  设线性方程分别为 $Xu = 0$ 与 $X^TXu = 0$.

  - $X^T(Xu) = 0$, 所以$Xu = 0$的解为$X^TXu = 0$的解.
  - $X^TXu = 0 \Rightarrow u^TX^TXu = (Xu)^TXu = u^T \cdot 0 = 0$, 所以$Xu$的模为零 $\Rightarrow$ $Xu = 0$, 所以$X^TXu = 0$的解为$Xu = 0$的解.

  综上, 我们证明了$Xu = 0$ 与 $X^TXu = 0$有相同的解空间, 同理可以证明其他, 所以$r(XX^T) = r(X^TX) = r(X) = r(X^T)$.

不妨令X^TX的特征值为$\boldsymbol{\xi}$, 我们有:

$$X^TX\boldsymbol{\xi}=\lambda\boldsymbol{\xi} \ \Rightarrow \ XX^T(X\boldsymbol{\xi})=\lambda(X\boldsymbol{\xi})$$

综上所述, $\boldsymbol{X}^T\boldsymbol{X}$与$\boldsymbol{X}\boldsymbol{X}^T$有相同非零特征值.

所以$Cov(x)=\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}\in\mathbb{R}^{D\times D}$ 与 $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\in\mathbb{R}^{N\times N}$ 有相同非零特征值, 秩不大于矩阵维度N维, 又$\boldsymbol{X}$经过中心化处理, 所以所有列向量相加为 $\mathbf{0} \ \Rightarrow \ r(\boldsymbol{X})$不大于$N-1$, 所以有效投影子空间不超过 $N-1$.

(3)

$(x_i\in\mathbb{R}^2,\ N=6)$, 计算均值:

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{n}x_i=[\ 4,\ 6\ ]$$

计算协方差矩阵:

$$Cov(X)=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T=\begin{pmatrix} 8 & 8.5 \\ 8.5 & 10 \end{pmatrix}$$

特征值分解(谱分解), 特征向量为$\xi_1,\ \xi_2$, 对应的特征值为$\lambda_1,\ \lambda_2$:

$$\xi_1=\begin{pmatrix} -0.7473 \\ -0.6645 \end{pmatrix}\ \lambda_1=0.4414\ ; \quad \xi_2=\begin{pmatrix} 0.6645 \\ -0.7473 \end{pmatrix}\ \lambda_2=17.5586\ ;$$

降维至一维表示:

$$\begin{aligned} X' &= \left(\ \xi_1^T(x_1-\bar{x}),\xi_1^T(x_2-\bar{x}),\ \cdots\ ,\xi_1^T(x_6-\bar{x})\ \right) \\ &= (\ 3.5709, 1.4118, 0.6645, 0, -1.4118, -4.2354\ ) \end{aligned}$$

[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：

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

Solution. 由书本10.1:

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

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

$$\text{不妨令:}\ \ c^\star=argmax_{c\in\mathcal{Y}}P(c|x)\Rightarrow err^\star=1-P(c^\star|x)$$

由10.2, 并进行一些简单的数学推导:

$$err = 1 - \sum_{c\in\mathcal{Y}} P(c|x)P(c|z) \approx 1 - \sum_{c\in\mathcal{Y}} P^2(c|x) \geq 1 - \left(\sum_{c\in\mathcal{Y}} P(c|x)\right) P(c^\star|x) = err^\star$$

这样我们就证明了左边的小于等于. 下面来证右边:

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

取定$err^\star$时:

$\sum_{c\in\mathcal{Y}-c^*} P(c|x)$为定值, 由拉格朗日乘子法得出最小值为$P(c|x)$都相等时, 其中$c\in\mathcal{Y}-c^*$.

$$\therefore err \leq \left(2 - err^\star\right) err^\star - \frac{\left(err^\star\right)^2}{|Y|-1} = err^\star\left(2 - \frac{|Y|}{|Y|-1} * err^\star\right)$$

综上所述, 原不等式得证.

[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) $x = \{0, B\}$的类别为0, $y = 0$.

(2) 估计类先验概率:

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

为每个属性估计条件概率$P(x_i|c)$:

$$P(x_1=-1|y=0)=\frac{3}{6}=0.5, P(x_1=0|y=0)=\frac{2}{6}\approx 0.33, P(x_1=1|y=0)=\frac{1}{6}\approx 0.17$$

$$P(x_1=-1|y=1)=\frac{2}{9}\approx 0.22, P(x_1=0|y=1)=\frac{3}{9}\approx 0.33, P(x_1=1|y=1)=\frac{4}{9}\approx 0.44$$

$$P(x_2=B|y=0)=\frac{3}{6}=0.5, P(x_2=M|y=0)=\frac{2}{6}\approx 0.33, P(x_2=S|y=0)=\frac{1}{6}\approx 0.17$$

$$P(x_2=B|y=1)=\frac{1}{9}\approx 0.11, P(x_2=M|y=1)=\frac{4}{9}\approx 0.44, P(x_2=S|y=1)=\frac{4}{9}\approx 0.44$$

$\therefore$ 确定$x=\{\ 0,B\ \}$的标记:

$$P(y=0)\times P(x_1=0|y=0)\times P(x_2=B|y=0)\approx 0.4\times 0.33\times 0.5=0.066$$

$$P(y=1)\times P(x_1=0|y=1)\times P(x_2=B|y=1)\approx 0.6\times 0.33\times 0.11=0.02178$$

(这里最后的计算是近似, 因为之前类先验概率和条件概率有近似)综上所述, 朴素贝叶斯分类器将$x=\{\ 0,B\ \}$分类为0.

(3)

类先验概率修正为:

$$P(y=0)=\frac{6+1}{15+2}\approx 0.412, P(y=1)=\frac{9+1}{15+2}\approx 0.588$$

类似地,

$$P(x_1=0|y=0)=\frac{2+1}{6+3}\approx 0.33, P(x_2=B|y=0)=\frac{3+1}{6+3}\approx 0.44$$

$$P(x_1=0|y=1)=\frac{3+1}{9+3}\approx 0.33, P(x_2=B|y=1)=\frac{1+1}{9+3}=0.17$$

$\therefore$ 重新计算 $x=\{\ 0,B\ \}$的标记:

$$P(y=0)\times P(x_1=0|y=0)\times P(x_2=B|y=0)\approx 0.412\times 0.33\times 0.5=0.06798$$

$$P(y=1)\times P(x_1=0|y=1)\times P(x_2=B|y=1)\approx 0.588\times 0.33\times 0.11=0.02134$$

(这里最后的计算是近似, 因为之前类先验概率和条件概率有近似)综上所述, ”拉普拉斯修正”的朴素贝叶斯分类器将$x=\{\ 0,B\ \}$分类为0.

[20 pts] Problem 4 [KNN in Practice]

(1) **[20 pts]** 结合编程题指南，实现KNN算法。

Solution. 这里虽然给的是double型的label, 但是只有0/1, 所以认为是手误, 我预测的结果就是0/1非double型的.

按照KNN算法思路和给定的框架, 较容易可以实现, 主要就是利用好Python一些写好的函数, 这次的框架给的真的很赞!

准确率大概是0.7+, 稍微调了一下超参数近邻个数.

参考文献

[1] 周志华, 机器学习.