

机器学习导论

习题五

181220031, 李惟康, liwk@smail.nju.edu.cn

2020 年 5 月 26 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。

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

Solution.

- (1) 当数据受到噪声影响时, 较小特征值对应的特征向量相对更可能与噪声有关, 因而将它们舍弃能在一定程度上起到降噪的效果。
- (2) 显然, 主成分分析当中主成分 (即特征向量) 的最大值与 N 个样本 (样本维度 $D > N$) 组成的数据集对应的协方差矩阵的秩相等。而矩阵的秩被样本数所限制, 对于协方差矩阵 (N 维方阵), 由于所有样本都进行了中心化处理, 至多有 $N-1$ 个非零特征值所对应的特征向量, 因此主成分分析的有效影子空间不超过 $N-1$ 维。
- (3) 首先对所有样本进行中心化, 有:

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

上述样本数据对应的协方差矩阵为:

$$S = \frac{1}{6} X X^T = \begin{bmatrix} \frac{8}{3} & \frac{17}{6} \\ \frac{17}{6} & \frac{10}{3} \end{bmatrix}$$

接下来对协方差矩阵进行特征值分解, 求解得特征值:

$$\lambda_1 = \frac{18 + \sqrt{293}}{6}, \lambda_2 = \frac{18 - \sqrt{293}}{6}.$$

分别代入, 因此对应归一化后特征向量为:

$$\xi_1 = \begin{bmatrix} 0.66451 \\ 0.74728 \end{bmatrix} \quad \xi_2 = \begin{bmatrix} 0.74728 \\ -0.66451 \end{bmatrix}$$

降维后所得的结果为:

$$\begin{aligned} X' = \xi_1^T X &= \begin{bmatrix} 0.66451 & 0.74728 \end{bmatrix} \begin{bmatrix} -2 & -1 & -1 & 0 & 1 & 3 \\ -3 & -1 & 0 & 0 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} -3.57086 & -1.41179 & -0.66451 & 0 & 1.41179 & 4.23537 \end{bmatrix} \end{aligned}$$

[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:

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

Solution.

该题证明部分参考 [1] 中的相关内容。

令 $c^* = \arg \max_{c \in Y} P(c|x)$. 注意到: $\sum_{c \in Y \setminus \{c^*\}} P(c|x) = 1 - P(c^*|x) = err^*$.

首先考虑上述不等式的下界, 容易得到:

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

不等式两边同时加 $P^2(c^*|x)$ 有:

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

由 err 和 err^* 的定义, 故有:

$$err^* \leq err.$$

接下来考虑上界, 根据 Cauchy-Schwarz 不等式有:

$$(|Y| - 1) \sum_{c \in Y \setminus \{c^*\}} P^2(c|x) \geq \left[\sum_{c \in Y \setminus \{c^*\}} P(c|x) \right]^2 = (err^*)^2$$

不等式两边同时加 $(|Y| - 1)P^2(c^*|x)$ 有:

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

由于 $err = 1 - \sum_{c \in Y} P^2(c|x)$, 将(2)代入可得:

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

因此可得:

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

[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 $y = \{0, 1\}$:

(1) [5pts] 通过查表直接给出的 $x = \{0, B\}$ 的类别;

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

(2) [10pts] 使用所给训练数据, 学习一个朴素贝叶斯分类器, 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 $\lambda=1$, 再重新计算 $x = \{0, B\}$ 的标记, 要求写出详细计算过程。

Solution.

(1) 查表知 $x = \{0, B\}$ 的类别为 0.

(2) 由表 1 可知类先验概率:

$$P(y=0) = \frac{2}{5}, P(y=1) = \frac{3}{5}$$

以及相应的条件概率:

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

由此对样本 $x = \{0, B\}$ 进行预测:

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

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

由于 $\frac{1}{15} > \frac{1}{45}$, 因此该样本被预测为 $y=0$ 类。

(3) 由表 1 可知类先验概率:

$$\hat{P}(y=0) = \frac{7}{17}, \hat{P}(y=1) = \frac{10}{17}$$

以及相应的条件概率:

$$\begin{aligned} \hat{P}(x_1 = -1|y=0) &= \frac{4}{9}, \hat{P}(x_1 = 0|y=0) = \frac{1}{3}, \hat{P}(x_1 = 1|y=0) = \frac{2}{9} \\ \hat{P}(x_1 = -1|y=1) &= \frac{1}{4}, \hat{P}(x_1 = 0|y=1) = \frac{1}{3}, \hat{P}(x_1 = 1|y=1) = \frac{5}{12} \\ \hat{P}(x_2 = B|y=0) &= \frac{4}{9}, \hat{P}(x_2 = M|y=0) = \frac{1}{3}, \hat{P}(x_2 = S|y=0) = \frac{2}{9} \\ \hat{P}(x_2 = B|y=1) &= \frac{1}{6}, \hat{P}(x_2 = M|y=1) = \frac{5}{12}, \hat{P}(x_2 = S|y=1) = \frac{5}{12} \end{aligned}$$

由此对样本 $x = \{0, B\}$ 进行预测：

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

$$\hat{P}(y = 1) \times \hat{P}(x_1 = 0|y = 1) \times \hat{P}(x_2 = B|y = 1) = \frac{5}{153}$$

由于 $\frac{28}{459} > \frac{5}{153}$ ，因此该样本仍被预测为 $y = 0$ 类。

[20 pts] Problem 4 [KNN in Practice]

结合编程题指南，实现 KNN 算法。

Solution.

函数实现及相应预测结果见附件。

参考文献

- [1] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.