机器学习导论
习题五

171830635, 俞星凯, yuxk@smail.nju.edu.cn

2020 年 6 月 1 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) 在较小特征值对应的特征向量方向上，数据的变化较小，可以认为这些轻微变化是由噪声引起，PCA 舍弃这些方向可以完成数据降噪。

(2) 考虑 $X = (x_1, x_2, \cdots, x_N) \in \mathbb{R}^{D\times N}$，中心化得到 $\hat{X} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_N - \bar{x})$，显然 $\sum_{i=1}^{N}(x_i - \bar{x}) = \mathbf{0}$，即 $\hat{X}$ 的列线性相关，$rank(\hat{X}) \leq \min(D, N) - 1 = N - 1$，因此 $rank(\hat{X}\hat{X}^T) \leq N - 1$，PCA 有效投影子空间不超过 $N - 1$ 维。

(3) 中心化

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

计算协方差矩阵

$$\hat{X}\hat{X}^T = \begin{bmatrix} 16 & 17 \\ 17 & 20 \end{bmatrix}$$

求解特征方程

$$0 = |\lambda E - \hat{X}\hat{X}^T| = \lambda^2 - 36\lambda + 31$$

得到特征值

$$\lambda = 35.1172,\ 0.8828$$

第一个特征向量

$$w = \begin{bmatrix} 0.6645 \\ 0.7473 \end{bmatrix}$$

降维之后数据

$$w^T\hat{X} = \begin{bmatrix} -3.5709 & -1.4118 & -0.6645 & 0 & 1.4118 & 4.2354 \end{bmatrix}$$

[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.

令 $c^* = \arg\max_{c \in Y} P(c|x)$

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

不等式左边得证。

利用柯西不等式

$$\begin{aligned} err &= 1 - \sum_{c \in Y} P^2(c|x) \\ &= 1 - P^2(c^*|x) - \sum_{c \in Y, c \neq c^*} P^2(c|x) \\ &\leq 1 - P^2(c^*|x) - \frac{[\sum_{c \in Y, c \neq c^*} P(c|x)]^2}{|Y| - 1} \\ &= 1 - P^2(c^*|x) - \frac{[1 - P(c^*|x)]^2}{|Y| - 1} \\ &= (1 - P(c^*|x)) \left(1 + P(c^*|x) - \frac{1 - P(c^*|x)}{|Y| - 1}\right) \\ &= err^*(2 - err^* - \frac{err^*}{|Y| - 1}) \\ &= err^*(2 - \frac{|Y|}{|Y| - 1} \times err^*) \end{aligned}$$

不等式右边得证。

[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) 表中 6 号数据 $x = \{0, B\}$，类别为 0。

(2)

$$p(y=0) = \frac{6}{15},\ p(y=1) = \frac{9}{15}$$
$$p(x_1=-1|y=0) = \frac{3}{6},\ p(x_1=0|y=0) = \frac{2}{6},\ p(x_1=1|y=0) = \frac{1}{6}$$
$$p(x_2=B|y=0) = \frac{3}{6},\ p(x_2=M|y=0) = \frac{2}{6},\ p(x_2=S|y=0) = \frac{1}{6}$$
$$p(x_1=-1|y=1) = \frac{2}{9},\ p(x_1=0|y=1) = \frac{3}{9},\ p(x_1=1|y=1) = \frac{4}{9}$$
$$p(x_2=B|y=1) = \frac{1}{9},\ p(x_2=M|y=1) = \frac{4}{9},\ p(x_2=S|y=1) = \frac{4}{9}$$

由此得出

$$p(y=0)p(x_1=0, x_2=B|y=0) = p(y=0)p(x_1=0|y=0)p(x_2=B|y=0) = \frac{6}{15} \times \frac{2}{6} \times \frac{3}{6} = \frac{1}{15}$$
$$p(y=1)p(x_1=0, x_2=B|y=1) = p(y=1)p(x_1=0|y=1)p(x_2=B|y=1) = \frac{9}{15} \times \frac{3}{9} \times \frac{1}{9} = \frac{1}{45}$$

前者大于后者，标记为 0。

(3)

$$p(y=0)=\frac{7}{17},\ p(y=1)=\frac{10}{17}$$
$$p(x_1=-1|y=0)=\frac{4}{9},\ p(x_1=0|y=0)=\frac{3}{9},\ p(x_1=1|y=0)=\frac{2}{9}$$
$$p(x_2=B|y=0)=\frac{4}{9},\ p(x_2=M|y=0)=\frac{3}{9},\ p(x_2=S|y=0)=\frac{2}{9}$$
$$p(x_1=-1|y=1)=\frac{3}{12},\ p(x_1=0|y=1)=\frac{4}{12},\ p(x_1=1|y=1)=\frac{5}{12}$$
$$p(x_2=B|y=1)=\frac{2}{12},\ p(x_2=M|y=1)=\frac{5}{12},\ p(x_2=S|y=1)=\frac{5}{12}$$

由此得出

$$p(y=0)p(x_1=0|y=0)p(x_2=B|y=0)=\frac{7}{17}\times\frac{3}{9}\times\frac{4}{9}=\frac{28}{459}=0.0610$$
$$p(y=1)p(x_1=0|y=1)p(x_2=B|y=1)=\frac{10}{17}\times\frac{4}{12}\times\frac{2}{12}=\frac{5}{153}=0.0327$$

前者大于后者，标记为 0。

[20 pts] Problem 4 [KNN in Practice]

(1) **[20 pts]** 结合编程题指南，实现 KNN 算法。

Solution.

KNN 是机器学习中最简单的模型之一，这里使用 Python 和 Numpy 库实现一个 KNN 类，用来解决一个分类问题。

首先 KNN 类的构造函数需要指定 k。

```
class KNN():
  def __init__(self, k=5):
    self.k = k
```

其次定义函数计算测试样例到训练集的距离。

```
def distance(self, one_sample, X_train):
  return np.sum(np.square(X_train - one_sample), 1)
```

然后定义函数获取 k 个近邻的标签。

```
def get_k_neighbor_labels(self, distances, y_train):
  return y_train[np.argsort(distances)[:self.k]]
```

接着完成单个测试样例的预测，调用上面二者并选择票数最多的标签。

```
def vote(self, one_sample, X_train, y_train):
  distances = self.distance(one_sample, X_train)
  labels = self.get_k_neighbor_labels(distances, y_train)
  labels = list(labels)
  return max(set(labels), key=labels.count)
```

最后对测试集的预测可以通过循环调用上者得到。

```
def predict(self, X_test, X_train, y_train):
  n = X_test.shape[0]
  y_pred = np.zeros(n)
  for i in range(n):
    y_pred[i] = self.vote(X_test[i], X_train, y_train)
  return y_pred
```