作业3 PCA、线性回归、统计学习理论

1. 给定两类样本的特征，其中

类别1：

类别2：

1. 用PCA降维，分别将特征降至1维和2维，并给出降维后每个样本的位置。

答：PCA为无监督学习方法，降维过程中与标签y无关。

PCA投影过程：

1. 计算所有样本的均值：

2）计算中心化后的样本：

类别1：

类别2：

3）计算协方差矩阵：

4）对进行特征值分解，

得到其3个特征值均为2，

特征向量分别为：

5）将降至2维：**，**

代码如下：

import numpy as np

X = np.array([[0,0,0],

[1,0,0],

[1,0,1],

[1,1,0],

[0,0,1],

[0,1,0],

[0,1,1],

[1,1,1]])

mean = np.mean(X, axis=0)

X = np.subtract(X, mean)

X = X.T #PCA中X为D\*N的矩阵

S = np.dot(X, X.T)

Lambdas, V = np.linalg.eig(S)

# 前两大特征值对应的特征向量：

W = V[:, 0:2]

# 输出降维后的结果（中心化）

Z = np.dot(W.T, X)

print(Z)

print(Z.T)

1. 给出每个样本点降至2维后，再重构回3维空间的坐标。

答：重构

[ 0. , 0. , 0.5],

[ 1. , 0. , 0.5],

[ 1. , 0. , 0.5],

[ 1. , 1. , 0.5],

[ 0. , 0. , 0.5],

[ 0. , 1. , 0.5],

[ 0. , 1. , 0.5],

[ 1. , 1. , 0.5]

原来3维空间中分开的点重构后相同（如第2个样本和第3个样本），损失一部分信息。

代码如下：

X\_hat = np.dot(W, Z)

print(X\_hat)

X\_hat = X\_hat.T + mean

1. 假设我们采用带如下正则的线性回归，目标函数为：

当从0增加到时，描述以下各项如何变化：

1. 模型的过拟合行为；

答：为正则参数。当较小时，目标函数中第1项占主导地位，模型与训练数据拟合好，此时模型容易过拟合；当时，目标函数中第2项占主导地位，模型倾向于欠拟合。

1. 的值和模的大小；

答：当较小时，取值可以较大，取值较大；当时，取值小，取值较小。

（3）模型的偏差和方差。（注意：这里的偏差不是指）

答：当较小时，取值可以较大，模型复杂，偏差小方差大；当时，取值小，模型简单，偏差小方差大。

1. 假设我们想要训练线性模型。我们将使用梯度下降来最小化个训练样本上的误差平方和：

计算目标函数损失相对于和的偏导数。

答：

1. 假设我们采用不同的验证集划分方式： 2折交叉验证、10折交叉验证、留一交叉验证、单次70%/30%的训练集/验证集划分。
2. 不同的验证集划分方式会对模型性能产生什么影响？（如训练误差、泛化误差）

答：从2折交叉验证、10折交叉验证、留一交叉验证，训练样本数目越来越多，得到的训练误差和泛化误差也更接近采用所有数据训练模型时的性能。

1. 哪种方式得到的验证误差会提供 在“未见过的测试集”上误差的最佳近似？

答：一般而言，留一交叉验证的验证误差是测试误差的好估计，但留一交叉验证需要重复次的模型训练，每次训练的训练数据多，运行成本高。（为数据集中的样本数目）

1. 原始数据集有多大有影响吗？对于一个非常大或非常小的数据集，你会得到不同的结论吗？

答：原始数据集大小有影响。对非常大的数据集，上述方法会得到的相似的结果。但一般选择单次留出验证集的方式，只需训练一次模型，运行成本低，此时训练数据充分，验证集也足够多，验证误差为测试误差的很好估计；对非常小的数据集，留一交叉验证才能得到足够多的训练数据，此时验证误差是测试误差的较好估计。

1. 就计算而言，哪种方式最快？

答：单次的70%/30%的训练集/验证集划分最快。