《机器学习与模式识别》上机实验指导书

实验4：主元分析（PCA）

**实验目的：**

1. 熟悉函数说明文档的检索，熟悉主要函数的功能与参数含义。
2. 具备合理修改算法参数和基本的改写或编写简单代码的能力。
3. 熟悉sklearn.datasets中部分函数（载入手写数字load\_digits）使用。
4. 熟悉sklearn.decomposition.PCA主要参数、属性与内部函数的含义与使用。
5. 通过噪声去除和图像重建实验，加深对PCA作为通用的特征提取方法的理解。
6. 进一步熟悉Python编程与绘图方法。

**实验内容：**

学习相关材料，熟悉主要工具函数的使用。使用 Anaconda JupyterLab，运行、调试实验用基础代码，通过修改适当参数，增加必要代码，完成以下实验。

（1）熟悉主元分析函数sklearn.decomposition.PCA。

阅读手册文件的函数说明，参照基础代码并进行必要实验以观察效果，熟悉主要参数、属性、内部函数的含义，并简要说明。

（2）手写数字的PCA噪声去除实验。

基于载入的手写数字，利用下面语句生成标准差分布为2,4,8的噪声手写数字：noisy = np.random.normal(digits.data, 2)；并通过 pca = PCA(0.50).fit(noisy) 计算当重建数据的方差分别是原数据的0.5,0.7,0.9时，所需要的主元个数（填表如下）。简要说明结果。

在不同条件下，需要的主元个数

|  |  |  |  |
| --- | --- | --- | --- |
|  | 需要的解释方差百分比 | | |
| 噪声std | 0.5 | 0.7 | 0.9 |
| 2 |  |  |  |
| 4 |  |  |  |
| 8 |  |  |  |

（3）人脸PCA重建实验。

从课程群下载 lfwfunneled.tgz，拷贝到 C:\Users\Administrator\scikit\_learn\_data\中，“解压到当前目录”，确认：在 C:\Users\Administrator\scikit\_learn\_data\lfw\_home\lfw\_funneled 中，有 5760 项目（除11个txt说明文件外，其余为5749个人的一共13233幅人脸图像）。

观察500元PCA绘制的累积解释方差图（所谓碎石图，scree plot），并逐个计算n\_pcs=50,150,300等处的累积解释方差比，简要分析观察结果。

观察使用不同的n\_pcs=50,150,300时重建图像的效果差异，简要分析观察结果。

（4）人脸图像PCA前后，识别率对比实验。

用简单的1-NN，对不同的主元数（30, 50, 100, 150）重建后的识别率进行对比实验，简要分析观察结果。

**数据说明：**

lfwfunneled.tgz(237 MB)已经下载、上传到课程群文件。

**参考资料：**

**scikit-learn user Guide, Release 0.23.1（May 19, 2020）.pdf**

**PCA：sklearn.decomposition.PCA**

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA

**Datasets：sklearn.datasets: Datasets**

<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets>