

机器学习

第5次实验：

对PCA降维后的手写体数字图片数据分类

实验目的

01

会用Python对数据进行PCA降维

02

使用多层感知机对PCA降维后的手写体数字图数据片训练分类模型，对测试数据分类

03

会通过调整主成分贡献率来提高分类效果

实验内容

使用sklearn.decomposition的PCA类对手写体数字图片数据进行降维；
使用sklearn.neural_network的MLPClassifier类对降维后的数据训练分类模型，对测试数据进行分类，要求分类准确率达到80%以上

- ✓ 训练数据：digits_training.csv
- ✓ 测试数据：digits_testing.csv
- ✓ 第1列是类别，其他列是特征属性

实验内容

1

载入训练数据、分出特征属性和类别，对特征属性标准化，
显示读入数据的行数

#标准化函数

```
def normalizeData(X):
```

```
    return X - np.mean(X, axis=0)
```

2

使用PCA对数据降维，显示主成分个数

```
for i in range(len(pca.explained_variance_ratio_)):  
    total_ratio += pca.explained_variance_ratio_[i]
```

实验内容

3

使用多层感知机对降维后的数据训练分类模型，保存分类模型为
mlpNN_pca.m

```
MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(48, 24),  
random_state=1)
```

4

载入测试数据、分出特征属性和类别，对特征属性标准化，显示读
入数据的行数

```
xTest = pca.transform(xTest)
```

5

使用分类模型对测试数据分类，显示分类准确率

执行结果图例

```
1.载入训练数据,对自变量进行标准化,输出训练数据的样本数.....
装载训练数据: 5000 条
2.使用PCA对数据降维,取累积贡献率超过95%的前K个元素.....
选取 149 个主成分
训练数据维度: (5000, 149)
3.使用神经网络(多层感知)机训练分类模型.....
4.保存分类模型...
5.载入测试数据,对自变量进行标准化,输出测试数据的样本数.....
测试数据维度: (500, 149)
6.使用模型对测试数据分类...
7.输出模型准确率...
错误: 49 条
测试数据准确率: 0.902
```