

# 机器学习编程作业2

南京农业大学

## 1 分类编程实验

### 1.1 数据集介绍

该数据集是一个mat文件，你可以使用如下的代码读取。

```
import numpy as np
from scipy.io import loadmat
data = loadmat("ORL_32x32.mat")
fea = data['fea'].astype(np.float32)
gnd= data['gnd'].astype(np.float32)
```

其中fea表示特征矩阵，是一个大小为 $400 \times 1024$ 的矩阵，每一行是一个样本。gnd是一个 $400 \times 1$ 的列向量，每一行的数字表示当前这个样本的类别，一共10个类别。

### 1.2 任务

请完成以下任务

- (10) 请编写函数将数据**随机划分**为训练集和测试集，其中训练集中每个类别的样本为6个，该类别中剩下的样本作为测试集。将函数的名称定义如下：

```
[trainData ,trainGnd ,testData ,testGnd]=dataSplit ( fea ,gnd , ntr )
```

其中ntr表示每个类别选取的训练样本数目，本实验中ntr=6。

- (30) 请从训练集中选取类别为3和8的样本训练KNN(k=1),逻辑斯蒂回归分类器，并使用类别为3和8测试集评估模型性能。请用图或者表给出ntr=5,6,7,8的测试准确率。kNN和逻辑斯蒂回归可以使用sklearn中的函数完成。
- (50) 请训练一个能对10个类别进行分类的逻辑斯蒂回归模型，说出你的具体步骤，并在测试集上验证所提方案的性能。
- (10) 请简单总结这次编程作业的体会或者发现，没有此项不用写。