第 7 章：特征提取与特征选择

第一部分：简述题

1. 简述 PCA 的原理、学习模型和算法步骤。
2. 简述 LAD 的原理和学习模型。
3. 作为一类非线性降维方法，简述流形学习的基本思想。
4. 根据特征选择与分类器的结合程度，简述特征选择的主要方法，指出各类方法的特点。
5. 简述最优特征选择的基本思想。

第二部分：编程题

编程实现 1：PCA+KNN：即首先 PCA 进行降维，然后采用最近邻分类器（1 近邻分类器）作为分类器进行分类。

编程实现 2：LDA +KNN，即首先 LDA 进行降维，然后采用最近邻分类器（1 近邻分类器）作为分类器进行分类。

任务：采用 80%作样本作训练集，20%样本做测试集，报告降至不同维数时的分类性能。

(a) 所用数据集 1：AT&T 40 个人脸数据集（即著名的 ORL 数据集）。样本个数：400，样本维数：256，类别总数：40
提示：降维时可以以 5 为间隔，比如，10,15,20，…

读取数据和类标签信息的 Matlab 代码如下：

```
load ORLData_25;
X = ORLData';
X = double(X);
[n, d] = size(X);

labels = X(:, dim);                %获取各样本的类别标签
labels = floor(double(labels));
c = max(labels);                   % c = 40

X(:, dim) = [];                    %  获取样本数据
clear ORLData;
```

(b) 所用数据集 2：Vehicle 数据集。样本个数：846，样本维数：18，类别总数：4

读取数据和类标签信息的 Matlab 代码如下：

```
load vehicle;
out = UCI_entropy_data.train_data;

X = out';
X = double(X);
[n, d] = size(X);
labels = X(:, dim);
labels = floor(double(labels));    % 获取各样本的类别标签
c = max(labels);                   % c = 4
X(:, dim) = [];                    % 获取样本数据
clear UCI_entropy_data;
clear out;
```