1. 数据描述性分析实验题目

1.1 通过模拟方法生成二元正态分布向量（样本容量为500），其均值设定为(0，0）’，而协方差矩阵分别为如下情形：

1),; 2),; 3),;

4),; 5),;

6),; 7),; 8),，

请画出图像，并给出评价。

* 1. 对data\_1的数据剔除ID为841的观测值，
  2. 对四个变量EXPE、QUAL、LOYA和SATI画出散点图矩阵；
  3. 画出各变量的箱线图；
  4. 计算观测数据的pearson相关系数矩阵，并做相关性的显著性检验。

参考资料：描述性分析参考文档

1.3 已知 8 个乳房肿瘤病灶组织的样本，其中前3 个为良性肿瘤，后5 个为恶性肿瘤。数据为细胞核显微图像的5 个量化特征：细胞核直径，质地，周长，面积，光滑度。已知样本的数据如下：

13.54,14.36,87.46,566.3,0.09779

13.08,15.71,85.63,520,0.1075

9.504,12.44,60.34,273.9,0.1024

17.99,10.38,122.8,1001,0.1184

20.57,17.77,132.9,1326,0.08474

19.69,21.25,130,1203,0.1096

11.42,20.38,77.58,386.1,0.1425

20.29,14.34,135.1,1297,0.1003

试根据已知样本利用距离判别（分别用协方差矩阵相等、协方差矩阵不等）对下面未知种类的三个样本进行分类：

16.6,28.08,108.3,858.1,0.08455

20.6,29.33,140.1,1265,0.1178

7.76,24.54,47.92,181,0.05263

二、主成分分析部分实验题目

2.1 附表（data21.xls）中列出了2007年我国31个省、市、自治区和直辖市的农村居民家庭平均每人全年消费性支出的8个主要变量数据.

1. 试根据这8个主要变量的观测数据进行主成分分析，给出各主成分的贡献率;
2. 试分析前两个主成分的意义，并按第一主成分得分将31个省、市、自治区和直辖市排序；
3. 画出前两个主成分得分的散点图，并在散点图上标注上各个地区的名称.

参考资料：Matlab 帮助文档help中的princomp, 这里提供一个简单的注释pca.docx及相应程序pcatest.m.

2.2 （该题建议采用matlab） 参考资料中TrainDatabase图像库作为训练图像进行训练，该图像库中有10个人，每人有两幅图像，每幅图像大小为的。在TestDatabase中有10幅测试图像，其大小也为的。

1. 请利用TrainDatabase中的图像进行主成分分析，并将TestDatabase中的测试图像进行分类，选前三个最大的特征值。请显示特征脸，并给出分类依据；
2. 若选前两个和前四个最大的特征值，结果如何，请列出相应的结果。

参考资料：pcaadd.doc 主成分分析人脸识别说明文档,示例程序等文件：

三、 C均值聚类部分实验题目

* 1. 课本220页习题6.5.

3.2 （该题建议采用matlab）自己拍一张照片，利用C均值算法进行图像的分割和向量量化（分类数自选）.

参考资料：C均值图像分割（examp\_seg.m， examp\_quantity.m）



(a) 原图 (b) 分割图

图1，利用C均值算法进行图像分割



图2，利用C均值算法进行图像向量量化