**2022秋季学期硕士研究生《模式识别技术》**

（任课教师：柏正尧 教授）

期末大作业

截止日期：2022/12/20

**承诺：本次作业完全由本人独立完成，包括代码编写和报告撰写，没有获得或给予不诚实的帮助。**

姓名（学号）： 签名：

**问题1 （25%）**

(1) 图1显示了4种装配零件的彩色图像：螺母(nuts)、螺钉(screws)、支架(brackets)和垫圈(washers)。请你设计一组特征（不超过4种），能很好地把4种零件分开。分析、讨论你的特征和背后的原理。

(2) 编写代码提取每种模式的上述4种特征，绘制每对特征的二维散点图（如, f1 vs. f2, f3 vs. f4, 等），讨论你的结果。

(3) 计算原始空间中每一对样本之间的欧式距离。首先，你得把所有图像都转换为同样大小，比如32×32彩色像素（提示：在MATLAB中，可使用函数imresize）。把距离组成30×30的矩阵进行显示（提示：在MATLAB中，可使用函数imagesc）。在二维特征空间中，重复上述过程。比较两种距离矩阵，在原始图像空间中和特征空间中的类别可分离性有何不同？讨论你的结果。



图1 装备零件数据集

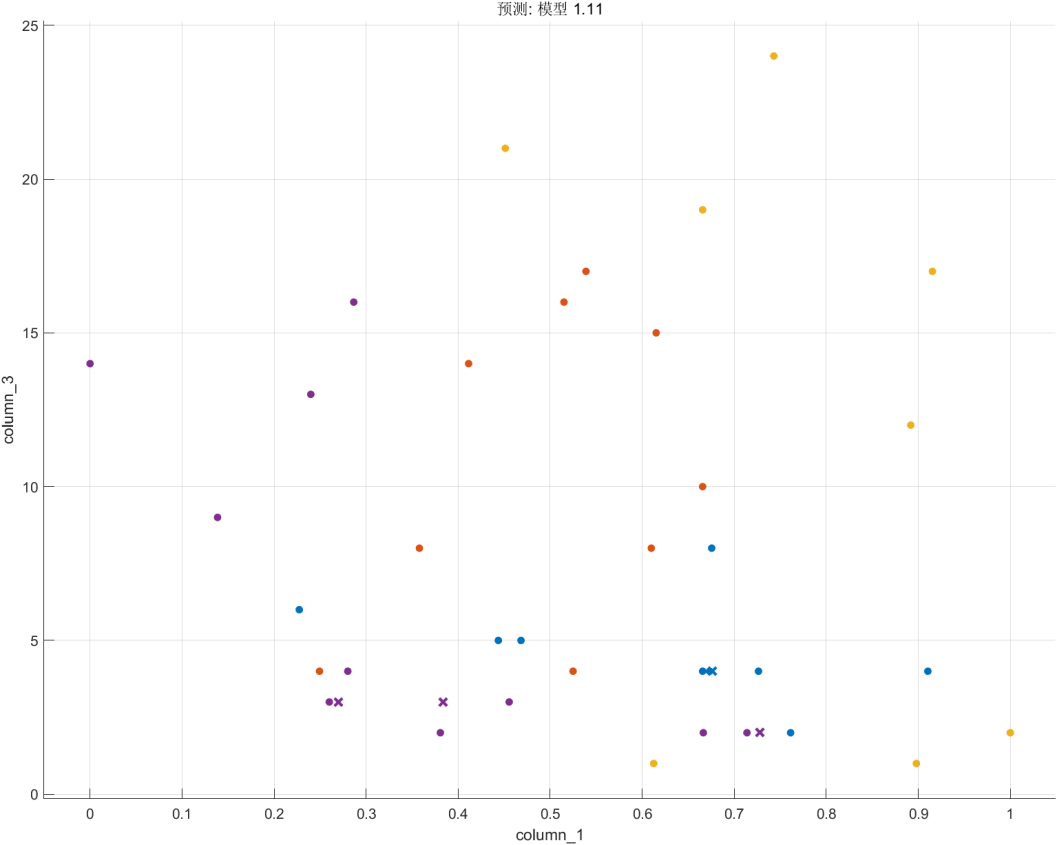
**提示：**1) 使用函数imread读入到一个变量中。为了计算每对32×32图像间的距离，你需要将每个图像转换成一个向量（例如，以光栅扫描方式），然后使用命令dist。2) 为了生成散点图，使用命令text。请使用标记表示三种类别，使用标记表示每个类别的10个样本。可使用命令axes调整两个坐标轴的刻度范围。

解：基于以下两个特征使用多种算法进行分类，分别为

1. 图像中圆形的面积占比(包含同心圆)
2. 图像中4连域名的个数
3. 二值化图像的和(对分类的贡献不大，放弃)

训练策略：不做任何防止过拟合的操作(全部参与训练)

预处理：先二值化，然后缩放到64X64



|  |  |
| --- | --- |
| 线性判别 | 52.5 |
| 二次判别 | 65.0 |
| 朴素贝叶斯 | 65.0 |
| 线性SVM | 57.5 |
| 二次SVM | 75.0 |
| 三次SVM | 77.5 |
| 精细高斯SVM | 90.0 |
| 加权KNN | 100 |

**问题2 （20%）**

考虑一个医学诊断问题，用快速生化检验筛查病人。根据下列似然函数，健康者的检验返回结果接近0，受感染者的返回结果接近1：



假设平均1万个患者中有1人受感染，且误诊的代价如下：(1) 将健康者诊断为“感染者”：预计病人综合医疗费用为2万人民币。(2) 将感染者诊断为“健康者”：预计由于误诊导致的医疗费用为100万人民币。

根据下列准则，分析并确定决策规则：1) 最大似然，2) 最大后验概率，3) 最小贝叶斯风险。讨论你的结果。

**提示：**可以使用MATLAB中的solve命令求解多项式的用符号表示的根，用命令subs把符号表达式转换为数值。

解：

1. 最大似然估计

将患者的检验结果分别带入上面的两个正态分布公式中，哪个类别概率大就判定为

**问题3 （15%）**

文件夹“cs\_fac”包含53幅人脸正面图像，对这些人脸进行PCA分解。(1) 生成平均脸的图像，(2) 生成前六个特征向量的图像（即“特征脸”），(3) 绘制相应主成分的二维PCA散点图。讨论您的结果。

**问题4 （15%）**

数据集hw1p4\_dat中包含下述三维问题的数据。

(1) 绘制数据集中每一对特征的二维散点图，并对数据的结构进行分析。

(2) 估计数据的平均向量和协方差矩阵。协方差矩阵中非对角线项与(1)中的散点图一致吗？为什么一致或不一致？

(3) 使用你在(2)中估计的均值向量和协方差矩阵生成高斯分布的数据集。（提示：可使用命令mvnrnd）

(4) 使用你在(3)中生成的数据集，重复做(1)。这里的散点图与(1)中的一致吗？为什么一致或不一致？

讨论你的结果。

**问题5 （25%）**

数据集hw1p5\_data中包含由非线性函数合成的数据，其中*n*为加性噪声。请你研究多项式函数能在多大程度上可以用来表示这个关系。

(1) 随机选择个数据点作为训练数据，其余数据点用作测试样本。建

立一阶多项式模型（例如，，**提示：**在 MATLAB中，使用命令polyfit）。

以测试样本对测试样本的形式，绘制模型输出。计算模型的均方误差（MSE，模型预测值与正确输出值之间的平方误差平均值）。

(2) 对于2-10阶多项式，重复做(1)。

(3) 重复做(1)和(2)各100次，估计每一阶多项式重复100次的平均MSE。绘制（即对数尺度的MSE）与多项式阶数的关系图。

(4) 对训练集大小为，重复做(1)-(3)。

(5) 讨论模型的随多项式阶数和用于训练模型的样本数如何变化。

讨论你的结果。