实验报告一

学号：20201060287

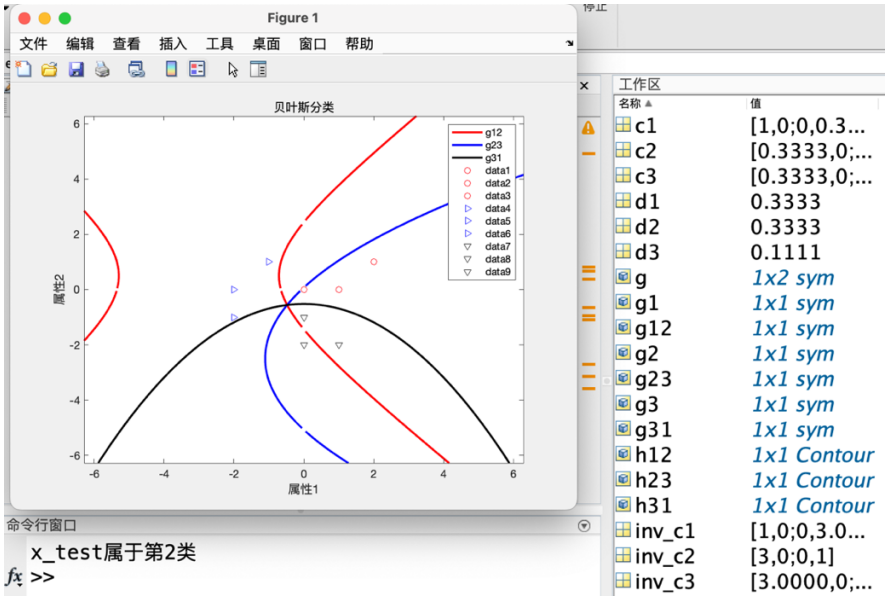
姓名：李昂

实验名称：基于最小错误率的贝叶斯分类器

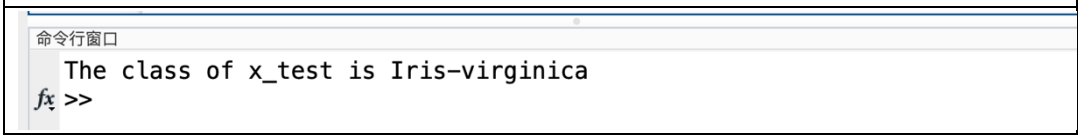
实验内容：使用 MATLAB 编程环境，设计基于最小错误率的贝叶斯分类器，并进行实验

实验要求及结果：

1. 运行demo_1.m文件，给出测试样本x_test=[-2, 2]的分类结果，用不同颜色画出三类模式的分界线，用不同颜色和形状画出三类模式的训练样本，请分别贴出分类结果的截图、绘制的分界线和训练样本示意图

分类结果、分界线、训练样本示意图


2. 改写上述文件中给出的示例程序，采用iris数据集进行贝叶斯分类实验，给出测试样本x_test =[6, 3.5, 4.5, 2.5]的分类结果（请给出分类结果的截图）

分类结果


问题

请简述基于最小错误率的贝叶斯分类器的工作原理：

答：分类时希望分类错误率可以降到最低，因此从这个目标出发，得到的分类决策就被称作最小错误率贝叶斯决策。

具体来说, 可以概括为一下步骤：

- 收集训练数据集：首先需要收集一组已知分类的数据集，这些数据集应该尽可能地代表了我们要分类的整个数据集
- 计算先验概率：对于每个类别，需要计算其在整个数据集中出现的先验概率。这个先验概率可以通过计算每个类别在训练数据集中出现的频率来得到
- 计算条件概率：对于每个特征和每个类别，需要计算其在给定类别的条件下出现的概率。这个条件概率可以通过计算每个特征在给定类别下出现的频率来得到
- 计算后验概率：对于每个测试样本，需要计算其在每个类别下的后验概率。这个后验概率可以通过将测试样本的每个特征在给定类别下的条件概率相乘，再乘以该类别的先验概率来得到
- 分类：将测试样本分配给具有最高后验概率的类别，即分类器认为最有可能的类别

附采用 iris 数据集进行贝叶斯分类实验的代码：

代码如下：

```
close all;
clear all;
clc;

% 加载鸢尾花数据集
[attrib1,attrib2,attrib3,attrib4,class]=textread('/Users/wallanceleon/
Desktop/模式识别/实验/实验一/Iris.data', '%f%f%f%f%s', 'delimiter', ',');
attrib = [attrib1, attrib2, attrib3, attrib4];
label_set = char('Iris-setosa','Iris-versicolor','Iris-virginica');
label = zeros(150, 1);
label(strcmp(class, 'Iris-setosa')) = 1;
label(strcmp(class, 'Iris-versicolor')) = 2;
label(strcmp(class, 'Iris-virginica')) = 3;

% 将鸢尾花数据集按类别分组
setosa = attrib(label == 1, :);
versicolor = attrib(label == 2, :);
virginica = attrib(label == 3, :);
```

```
% 求取各类的均值，协方差矩阵及其逆矩阵
mean_setosa = mean(setosa);
mean_versicolor = mean(versicolor);
mean_virginica = mean(virginica);

cov_setosa = cov(setosa);
cov_versicolor = cov(versicolor);
cov_virginica = cov(virginica);

inv_cov_setosa = inv(cov_setosa);
inv_cov_versicolor = inv(cov_versicolor);
inv_cov_virginica = inv(cov_virginica);

det_cov_setosa = det(cov_setosa);
det_cov_versicolor = det(cov_versicolor);
det_cov_virginica = det(cov_virginica);

% 给定一个测试样本 x_test,根据公式(2-39)判断 x_test 的类别归属
x_test =[6, 3.5, 4.5, 2.5];

p1 = 1/3 * exp(-1/2 * (x_test - mean_setosa) * inv_cov_setosa * (x_test
- mean_setosa)') / sqrt(det_cov_setosa);
p2 = 1/3 * exp(-1/2 * (x_test - mean_versicolor) * inv_cov_versicolor *
(x_test - mean_versicolor)') / sqrt(det_cov_versicolor);
p3 = 1/3 * exp(-1/2 * (x_test - mean_virginica) * inv_cov_virginica *
(x_test - mean_virginica)') / sqrt(det_cov_virginica);

p = [p1, p2, p3];
[max_p, index] = max(p);
disp(['The class of x_test is ', label_set(index, :)]);
```

实验总结：

加深了 Matlab 相关语法的熟练程度，在完成鸢尾花的实验时，由于一开始没有注意到条件概率密度属于正态分布，期望通过求解先验概率，后验概率和条件概率密度等条件完成求解，错误的使用基于最小错误率的贝叶斯分析，因此没能得出满意的结果，找到问题原因后基于公式（2-19）完成了相关实验