**模式识别实验**

基于IRIS数据集的贝叶斯分类决策

3115315017

高榕

# 一：实验要求

## 实验数据：IRIS数据集

IRIS以鸢尾花的特征作为数据来源，常用在分类操作中。该数据及由3种不同类型的鸢尾花的50个样本数据构成。其中的一个种类与另外两个种类是线性可分的，后两种是非线性可分的。

该数据集包含了5个属性：

1. Sepal Length（花萼长度），单位是cm；
2. Sepal Width（花萼宽度），单位是cm；
3. Petal Length（花瓣长度），单位是cm；
4. Petal Width（花瓣宽度），单位是cm；
5. Iris class（鸢尾花种类），Iris Setosa（山鸢尾）、Iris Versicolour（杂色鸢尾）、Iris Virginica（维吉尼亚鸢尾）。

## 实验模型：两两分类

IRIS数据集中有三种类型的数据，对应三种鸢尾花类型。每种类型中包含50个4维向量，向量即每种鸢尾花的属性，4个维度即鸢尾花的花萼长度，花萼宽度，花瓣长度和花瓣宽度。

两两分类要求分别组合两种类型的鸢尾花数据，然后通过训练，并正确的将组合打乱的两种鸢尾花分离出来。即如果鸢尾花种类记为IRIS-S、IRIS-Ver和IRIS-Vir，则两两分类有以下三种情况：IRIS-S和IRIS-Ver、IRIS-S和IRIS-Vir、IRIS-Ver和IRIS-Vir，即将非训练数据融合打乱，根据之前的数据训练情况，分类打乱的数据。

## 数据类型：正态分布

IRIS数据集数据服从正态分布。

## 实验目的：

在每种鸢尾花类型中，选择部分数据作为训练样本，作为当前类型的均值和协方差估计，然后以此训练数据作为分类器分类其余未训练数据。

# 二：实验原理

对于具有多个特征参数的样本（如本实验的iris数据样本有个参数），其正态分布的概率密度函数可定义为



式中，是维行向量，是维行向量，是维协方差矩阵，是的逆矩阵，是的行列式。

本实验我们采用最小错误率的贝叶斯决策，使用如下的函数作为判别函数

（3个类别）

其中为类别发生的先验概率，为类别的类条件概率密度函数。

由其判决规则，如果使对一切成立，则将归为类。

我们根据假设：类别，i=1,2,……,N的类条件概率密度函数，i=1,2,……,N服从正态分布，即有~，那么上式就可以写为



对上式右端取对数，可得



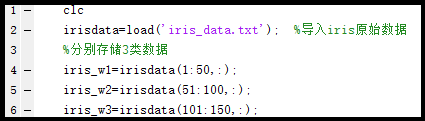
上式中的第二项与样本所属类别无关，将其从判别函数中消去，不会改变分类结果。则判别函数可简化为以下形式



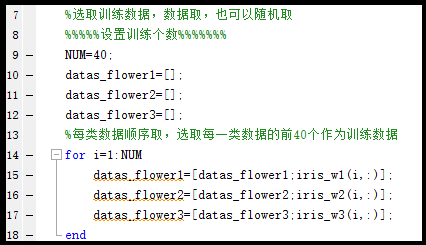
# 三：实验过程

## 提取数据

1. 从iris\_data.txt文本文件中读取IRIS数据集，代码如下：



1. 每一类鸢尾花数据中取前40个作为训练样本，代码如下：

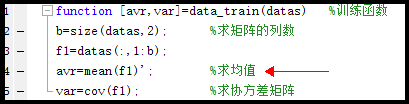


## 训练数据

1. 求取鸢尾花训练数据平均值，公式为

，i=1,2,3,4

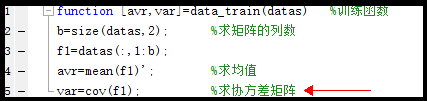
代码如下：



1. 求取鸢尾花训练数据的协方差，公式为



代码如下：



1. 注：

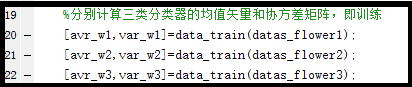
代表类的第个样本，第j个特征值；

代表类的个样品第j个特征的平均值；

代表类的第l个样品，第k个特征值；

代表类的个样品第k个特征的平均值。

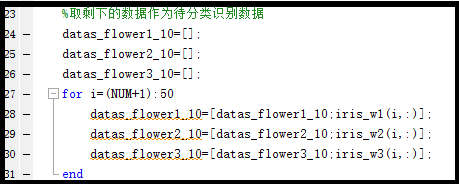
1. 鸢尾花三类数据训练，代码如下：



## 非训练数据整合

鸢尾花每一类型的数据有50个，之前选取了每一类40个，一共120个作为训练数据，剩下的非训练数据两两整合，然后测试分类结果。

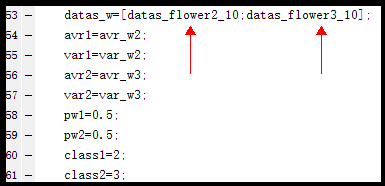
非训练数据整合代码如下：



## 数据训练

利用每一类型的前40个数据训练好的分类器，即均值和协方差数据，对剩下的2组10个，一共20个数据，进行分类。

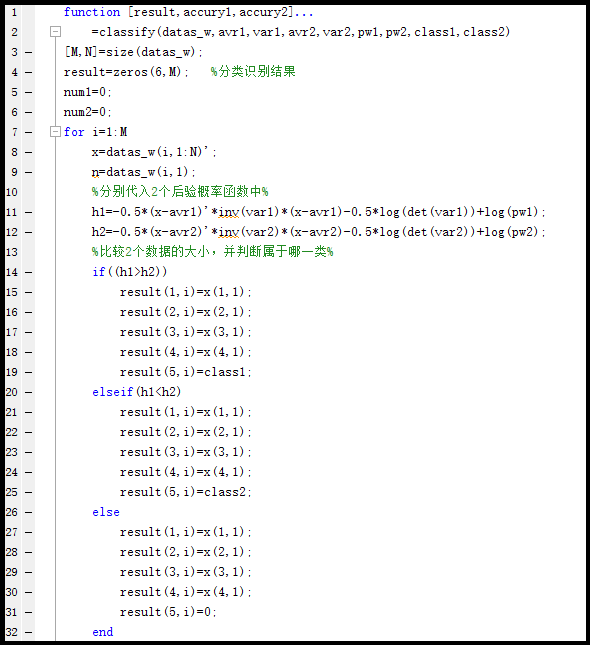
代码如下：



其中图示是将第二种和第三种鸢尾花的未训练数据进行整合，然后将第二种和第三种的均值和协方差作为传入参数，pw1和pw2作为先验概率，class1和class2作为数据分类标识。分类函数代码如下：

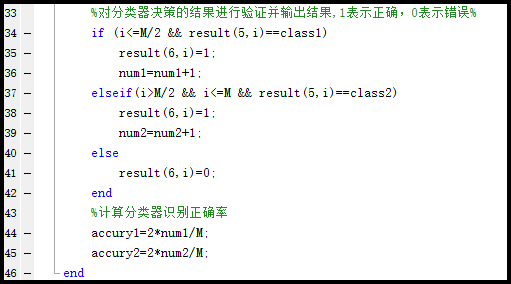


具体实现代码如下：



根据原理公式，在图中标示出贝叶斯后验概率计算步骤，选取非训练数据的每一行，计算其均值和协方差，然后比较两个数据的大小，判别函数结果大的即属于当前那一类，否则，哪一类都不属于（可能遇到两贝叶斯后验概率计算结果大小相同的情况，此时仅根据这一分类器无法判断），标记当前结果为0.

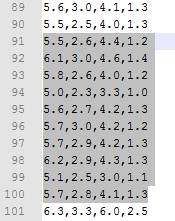
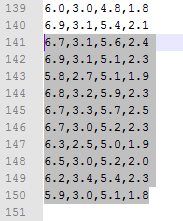
result矩阵存放分类结果。result是一个6\*M维矩阵，M是当前待分类的数据个数，第二种类型和第三种类型的待分类数据一共为20个，所以这里M=20；6是result存放结果的维数，前4维表示当前数据属性，即数据的拷贝，第5维表示分类结果，1表示第一类型的数据，2表示第二类型的数据，3表示第三类型的数据，第6维表示分类判断，根据鸢尾花数据的确定性，将分类结果与原本的结果进行比较，1为分类正确，0为分类错误。



最后将正确率输出，即分类情况。

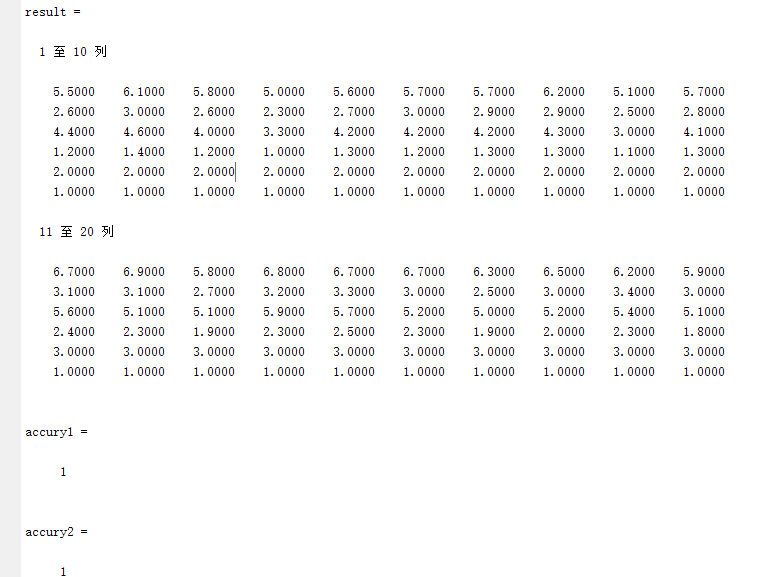
# 实验分析

## 实验数据

91到100即第二种类型的鸢尾花的待分类数据，141到150即第三种类型的鸢尾花的待分类数据。实验以第二种和第三种分类。

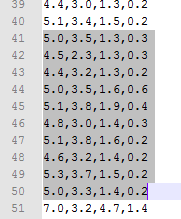
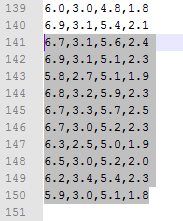
## 实验结果

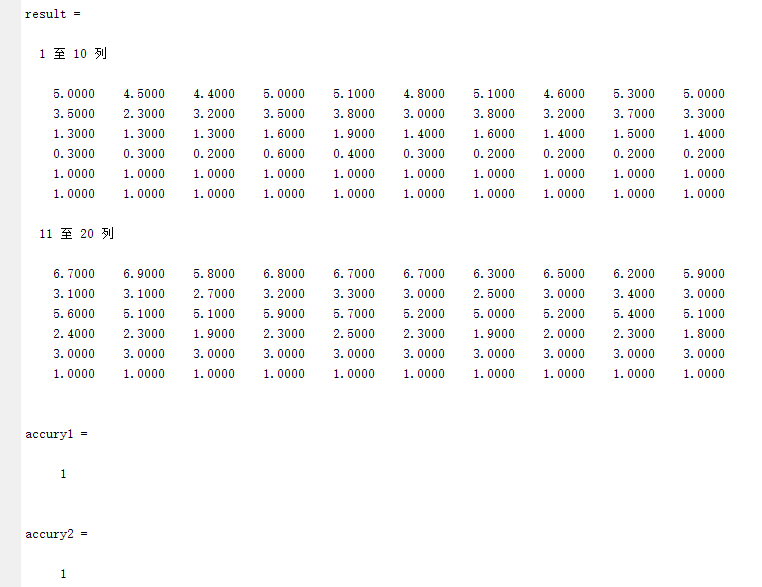


一共20列，为第二种和第三种类型待分类数据的组合，根据上图，第1列到第10列的第5维均为2，通过贝叶斯后验概率计算结果标明前10组数据属于鸢尾花第二种，检测准确率100%，第11列到第20列的第6维数据均为3，同样标明后10组数据属于鸢尾花第三种，检测准确率100%。

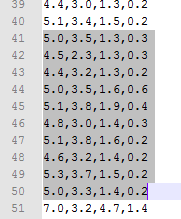
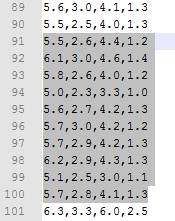
对照实验数据可得，每一列前4维均为当前数据属性的拷贝，由此可以确认分类结果完全符合IRIS数据本身。

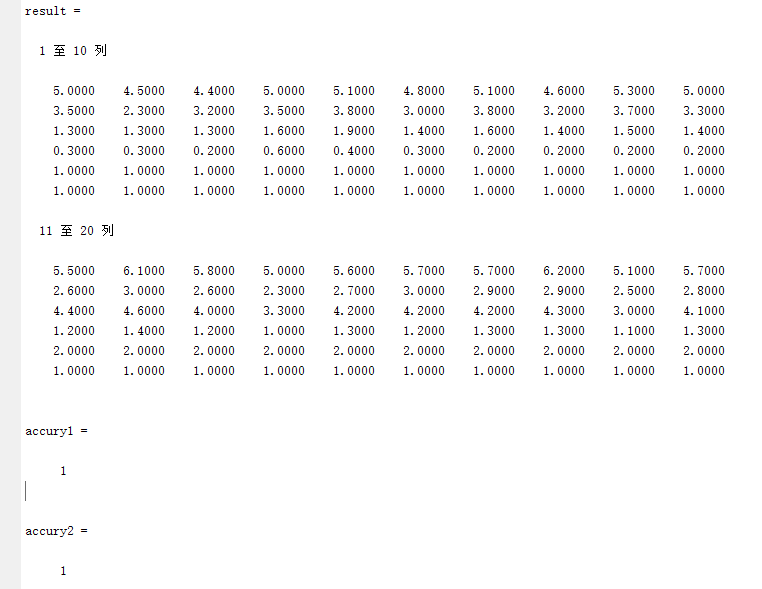
## 其他分类结果



第一种类型的数据和第三种类型的数据分类结果正确。



第一种类型的数据和第二种类型的数据分类结果正确。

## 实验总结

通过本次模式识别实验，采用最小错误率的贝叶斯决策方案，利用训练数据计算当前鸢尾花类型的均值和协方差，以此进行分类器训练，能够正确的对待分类数据进行分类，并经过与原数据对比，确定了分类的正确性。

通过实验，一方面加深了自己对Matlab编程的认识和理解，积累了有价值的经验和技能；另一方面更使自己对模式识别分类的基本情况有了概要性的了解，对分类器的构造以及它在分类过程中的作用有比较清晰的认识。总之，第一个实验收获很大。