第五次作业

分类问题：

汽车评价数据库一共有6个变量，即buying， maint, doors, persons, lug boot, safety (**在本次作业中，不考虑第6个变量safety**,否则协方差求逆有bug,我们简化之，即只要前五个变量).每个变量经过了等级划分。数据的类别为4类，即unacc, acc, good, vgood。问：

1）假设4个类别总体方差都相等，请根据训练数据(train.txt)，用马氏距离预测出测试集的类别(test.txt)。

2) 假设4个类别总体方差不一致，请根据训练数据(train.txt)，用马氏距离预测出测试集的类别(test.txt)。

提示：

1） 本次作业的测试集类别不公开，作为改作业的分数判断。马氏距离有一定的局限性，是算法的问题，能做到多高准确率就做多少准确率。

2） 作为自我验证，可以自己从训练集中划分一小部分数据，作为自我评测（通常称之为评测集,evaluation set）。

3）这些等级，可以依次取0,1,2，...值。

数据相关：

分成4类：

unacc, acc, good, vgood

变量等级：

buying: vhigh, high, med, low.

maint: vhigh, high, med, low.

doors: 2, 3, 4, 5more.

persons: 2, 4, more.

lug_boot: small, med, big.

safety: low, med, high.

```
   CAR                      car acceptability
   . PRICE                  overall price
   . . buying               buying price
   . . maint                price of the maintenance
   . TECH                   technical characteristics
   . . COMFORT              comfort
   . . . doors              number of doors
   . . . persons            capacity in terms of persons
to carry
   . . . lug_boot           the size of luggage boot
   . . safety               estimated safety of the car
```

数据类别比例描述：

```
   class      N          N[%]
   ------------------------------
   unacc     1210     (70.023 %)
   acc        384     (22.222 %)
   good        69     ( 3.993 %)
   v-good      65     ( 3.762 %)
```