目录

[一、 单因素方差检验的假设 2](#_Toc496539594)

[二、 写出任务的H0和H1假设 2](#_Toc496539595)

[三、 单因素方差进行分析 2](#_Toc496539596)

[四、 数据转换分析 6](#_Toc496539597)

[五、 做单因素方差分析关于非正态分布 10](#_Toc496539598)

[六、 使用logistic regression 进行分类 13](#_Toc496539599)

[七、 附录：补充说明 16](#_Toc496539600)

## 单因素方差检验的假设

**解答：**

若因素A有r个不同水平A1,A2,Ar，在水平Ai下的总体记为Xi,在水平Ai下进行独立实验，对第j次实验，我们记为Aij.则单因素方差分析的数学模型的基本假设为：

1. 总体Xi服从正太分布，也就是各个水平当中的数据都来自正太分布
2. 总体Xi的方差相等
3. 各Xij相互独立，也就是不同实验之间都相互独立。

简单来说就是正态性，方差齐性，独立性。

## 以不同类别分组，对平均年龄做分析，年龄在各组之间是不是有显著差异。写出这个任务的H0和H1假设。

**解答：**

显然对这个问题，两个假设分别为:

H0:不同类别之间的平均值都相等。

H1:至少有两个类别之间的平均值不相等。

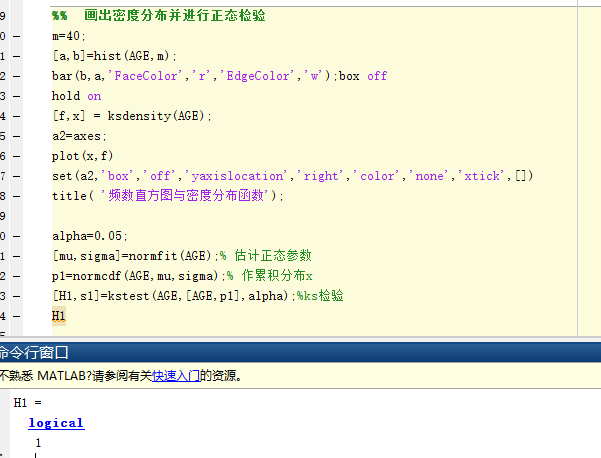
## 使用不同的软件进行分析

1. 画出第七列的密度分布函数，它是不是符合高斯分布，并进行检验

第七列的频率分布直方图如下：

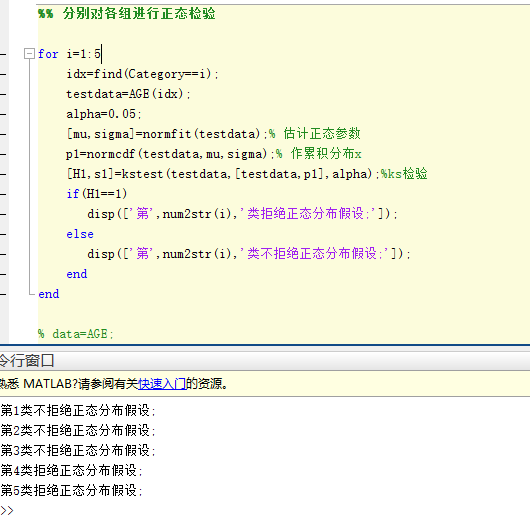


经过Kolmogorov-Smirnov检验，其拒绝正态分布假设。代码和结果截图如下：



1. 检验第七列当中不同类别当中的数据是不是符合正太分布，并检验方差齐性

**解答：**分别对每组内的数据进行正态检验，依然使用Kolmogorov-Smirnov检验，得到结果如下：前三组服从正态分布，后两组拒绝正态分布假设



关于方差齐性检验，使用matlab函数vartest可得到p值接近0，所以不满足方差齐性，中间结果如下：





1. 对第七列数据关于第二列做单因素方差分析，写出结论并提供必要的统计分析，并可视化部分中间结果。

解答：

显然，根据上一问的结果，各组之间的数据不满足方差齐性，并不满足做单因素方差分析的前提，可以预见，如果直接做的话，应该会拒绝H0假设，而接受备择假设。下面尝试做单因素方差分析：



显然这个检验的p值接近于0，显然拒绝原假设。



## 重新选择三列，画出经验分布，并判断哪些列符合第一问当中的假设，他们对应的对数转换之后的结果是不是符合分布。

**解答：**

选择稠密度，性别比和年龄差进行分析。主要进行分布假设检验。

三列数据的经验分布分别如下：







独立性通常满足，对不同类别分别做正态分析：

首先是稠密度：



第1类拒绝正态分布假设;

第2类拒绝正态分布假设;

第3类拒绝正态分布假设;

第4类拒绝正态分布假设;

第5类拒绝正态分布假设;

所以组内不符合正态分布，不再检验方差齐性。

关于性别比：



第1类拒绝正态分布假设;

第2类不拒绝正态分布假设;

第3类不拒绝正态分布假设;

第4类拒绝正态分布假设;

第5类拒绝正态分布假设;

每组不是都符合正态分布假设，不再检验方差齐性。

关于年龄差：



第1类拒绝正态分布假设;

第2类不拒绝正态分布假设;

第3类不拒绝正态分布假设;

第4类拒绝正态分布假设;

第5类拒绝正态分布假设;

不是每组都符合正态分布，正态性假设不成立，不再进行方差齐性检验。

对数据进行对数转换：

稠密度：



第1类不拒绝正态分布假设;

第2类不拒绝正态分布假设;

第3类不拒绝正态分布假设;

第4类拒绝正态分布假设;

第5类不拒绝正态分布假设;

虽然并不是都满足正态，但是已经相比原来有了明显改善。

性别比：由于之前的数据从图形看已经非常接近正态，其中第四类数据转换之后数据本身进行正态估计的参数估计结果很差，且前其他类都拒绝正态分布假设。

年龄差：



第1类拒绝正态分布假设;

第2类拒绝正态分布假设;

第3类不拒绝正态分布假设;

第4类拒绝正态分布假设;

第5类拒绝正态分布假设;

不是都满足正态分布假设，不再进行方差齐性检验。

## 做单因素方差分析关于非正态分布

1. 首先找出可能的方法

**解答：**

1. 进行数据变换，将其转换成正态分布。比如常用的一些对数转换，平方变化，以及倒变换等等。
2. 如果从理论上分析，数据本身可能确实是服从正态分布，可以考虑增加样本量，这样可以改善效果。
3. 可以使用BOX-COX 变换，是统计建模当中一种常见的数据变换方法。
4. 使用一些非参数的方式进行处理。比如Brown and Forsythe 检验法
5. 对你找出的三列做单因素方差分析，这些特征是不是变化的特别显著，可视化结果。

**解答:**

考虑使用对不符合正态的数据进行变换。然后进行单因素方差分析。

稠密度分析：





由结果可以知道，对于稠密度，不同组之间具有显著差异。

对于地域集中度：





显然对于地域集中度，不同类别之间也显著区分。

对于年龄差：





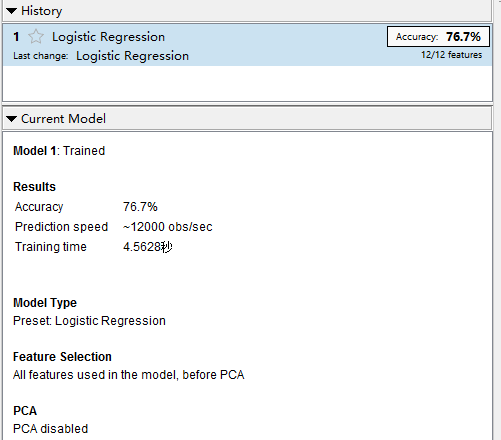
对于年龄差，显然也是具有显著区分。

## 选择两类，用logistics回归进行分类，或者你可以尝试一下多分类。

**解答：**

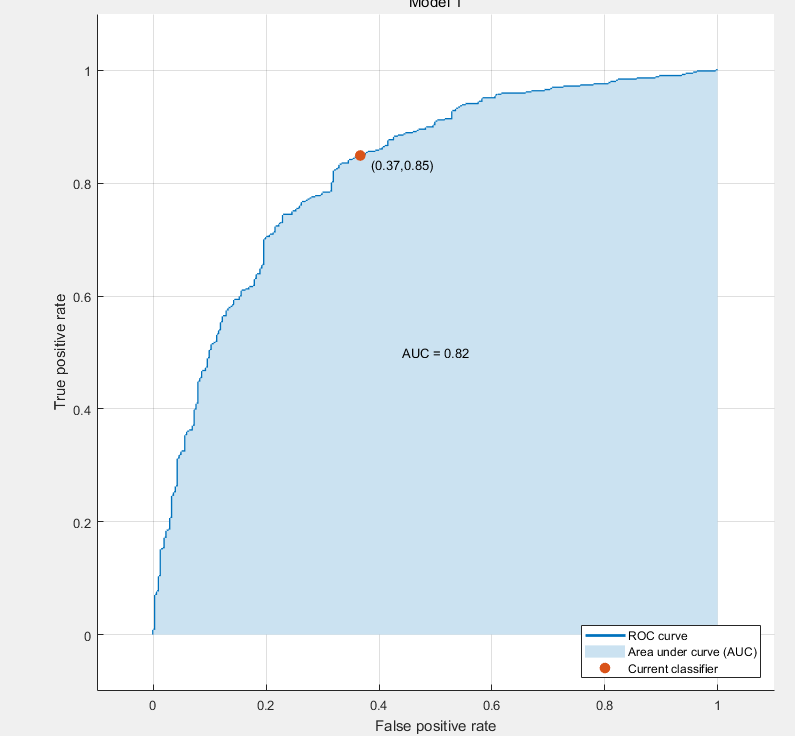
适用matlab 2017a 自带的classification learning app对标号为1,2类进行分类。

参数模型如下：

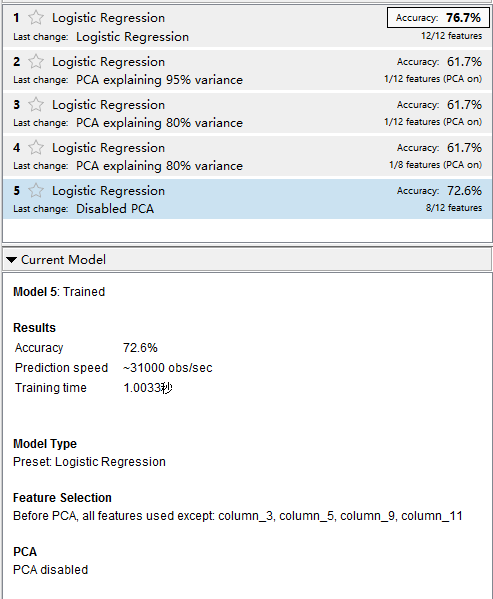


所有12列特征都使用，使用5折交叉验证。一共784个样例，标号为1的有484个。

分类器的正确率为76.7%。其ROC曲线如下：



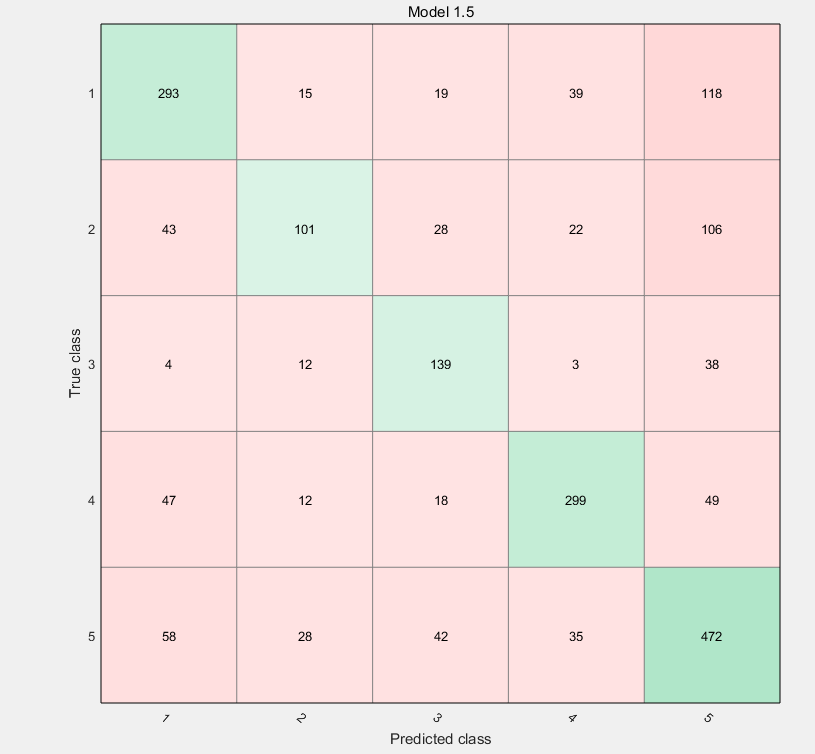
尝试修改model，增加pca以及随机选择一些特征而不是使用所有特征，结果如下：



可以看出使用pca并不是总是能够提高性能，特征组合的选择还需要进一步的分析。

多类分类：

Logistic进行分类一般适合处理二分类问题，但是可以使用一些其他的方法，比如构建不同的两两分类器，然后进行投票。或者直接使用树形结构不断的进行二分类，直到细分到我们希望的程度。具体过程不再赘述。简单进行多分类，其混淆矩阵如下：



## 附录：补充说明

1. 正态性检验方法
2. Kolmogorov-Smirnov正态性检验

Kolmogorov-Smirnov检验法是检验单一样本是否来自某一特定分布。比如检验一组数据是否为正态分布。它的检验方法是以样本数据的累积频数分布与特定理论分布比较，若两者间的差距很小，则推论该样本取自某特定分布族。

1. Lilliefor正态分布检验

该检验是对Kolmogorov-Smirnov检验的修正，当总体均值和方差未知时，Lilliefor提出用样本均值和标准差代替总体的期望和标准差，然后使用Kolmogorov-Smirnov正态性检验法，它定义了一个D统计量：



参数未知，由计算得到统计量，查表得Lilliefor检验的临界值，确定拒绝域，得出结论。

1. 方差齐性的检验方法

Levene方差齐性检验也称为Levene检验（Leverie’s Test），由H.Levene在1960年提出。Levene检验主要用于检验两个或两个以上样本间的方差是否齐性。要求样本为随

检验原理及方法如下：

①检验假设：，即各处理组方差相等；：，各处理组方差不全相等。取或。

②计算检验统计量W值



式中：W为Levene检验统计量，k为样本组数，Ni为第个样本的含量，N为各样本含量之和，将原始数据经数据转换后的新的变量值。为第个样本的均数，为全部数据的总的均数。

③判断原则Levene检验统计量W服从自由度为，的分布。当时，则，在水准上拒绝，接受，可认为各样本方差不全相等。当时，则，不拒绝，可以认为各样本方差齐性。