**数据挖掘第三次作业报告**

计算机学院 2120171017 黄焯恒

1. **主要目的**

数据分类、聚类及其可视化分析

**二、实现过程**

**1. 数据集准备和处理**

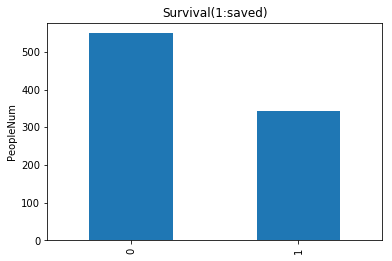
**（1）数据集选择**

Titanic数据集，包含文件train.csv，test.csv， gender\_submission.csv

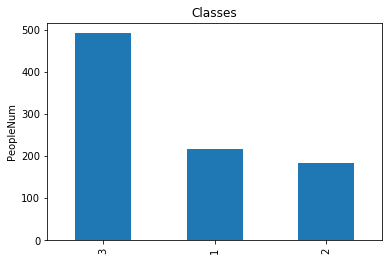
**（2）对数据集train.csv预处理**

数据集中的属性包括：

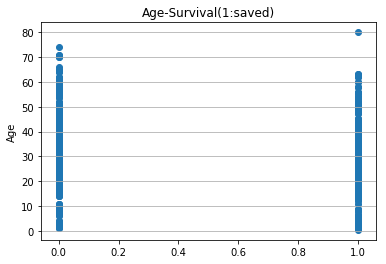
PassengerId、Survived、Pclass、Name、Sex、Age、SibSp、Parch、Ticket、Fare、Cabin、Embarked。首先，对其中的Survived、Pclass、Sex、Age、Embarked数据进行统计处理，并可视化，存活人数如下图所示。



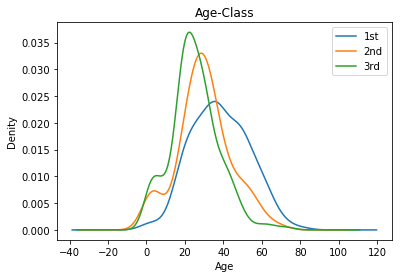
0表示非存活人数，1表示存活人数。然后对船舱等级分类，对不同船舱的人数进行分析统计，如下图所示。



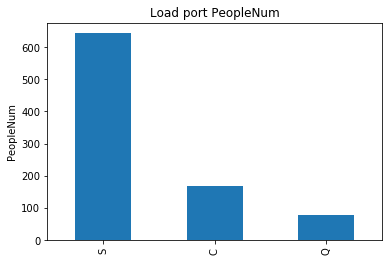
之后，对存活人数和非存活人数按年龄分布进行统计，如下图所示。

****

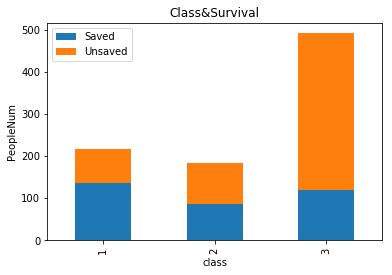
然后，统计每个船舱里面按年龄分的密度曲线，如下图所示。

****

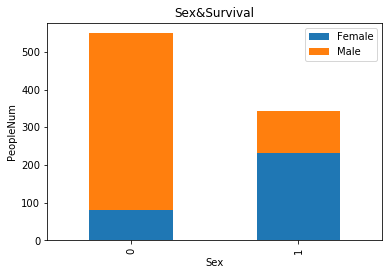
最后是分析从不同港口登上船的人数分布，如下图所示。

****

接着，结合不同的属性进行分析和预测，如人们所住船舱等级与存活人数进行分析，可以简单地看到，船舱等级越高的存活比例也越高。

****

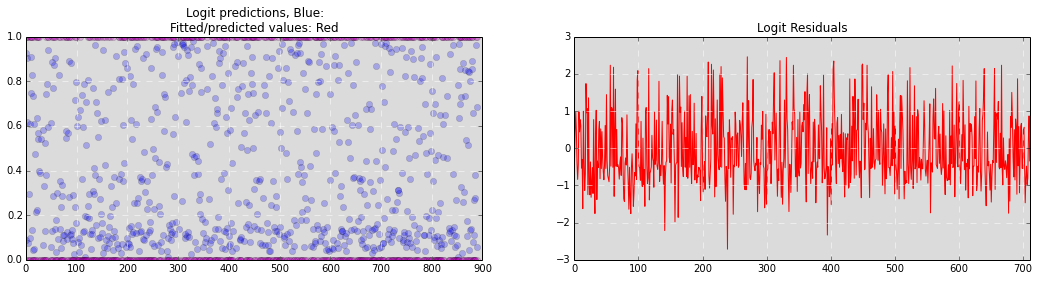
同时，在电影和纪录片中曾有人描述过，泰坦尼克号遇难时，让女人和孩子先走，从性别和存活人数的统计分析得出，女性获救的比例的确是高于男性。

****

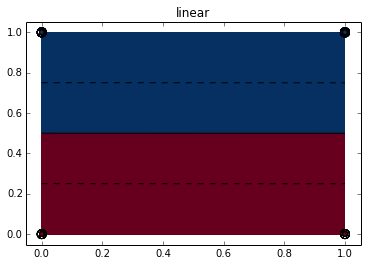
**2.对数据进行分类、聚类，并进行可视化分析**

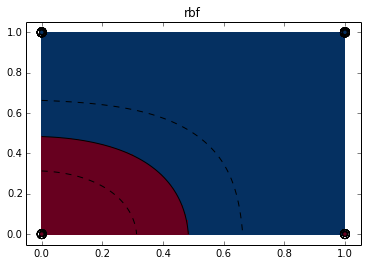
本次实验利用Logistic Regression模型、SVM模型和随机森林模型进行分类和聚类。

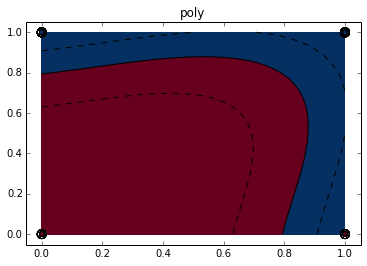
使用Logistic Regression模型分类并可视化的结果如下图所示。



而使用SVM方法处理结果如下图所示。







使用随机森林模型处理结果如下

*Mean accuray of Random Forest Predictions on the data was: 0.945224719101*

**三、实现总结**

通过本次的实验，了解到很多不同的分类、聚类算法和其性能和应用的场景，对数据分类、聚类及其可视化分析有了更深刻的理解。