2620170052　韩林洁

数据挖掘大作业三：分类与聚类

一、问题描述

1. 数据源

从以下 3 个数据集中任选一个

- [https://www.kaggle.com/c/titanic/data]

2. 要求：

1. 使用分类模型（至少 2 个）对数据集进行挖掘；
2. 对挖掘结果进行可视化，并解释其意义；
3. 使用聚类方法（至少 2 种）对数据集进行分析；
4. 对挖掘结果进行可视化，并解释其意义。

二、问题解答

分类算法：决策树、高斯分布的朴素贝叶斯分类器；

聚类算法：K-Means、DBSCAN 聚类。

1. 预处理操作

分类时：先将数据集中的‘PersonId’与‘Name’属性丢弃，因为这两个属性对于每条数据而言都是唯一的，不适合作为训练分类器的特征。对于二值属性‘Sex’,把取值 male 修改为 1，把取值 female 修改为 0。对于其他标称属性‘Pclass’、‘Cabin’、‘Embarked’,分别根据属性的取值范围用数值进行量化。经过变换，所有字段都是数值型取值。

对于某些属性的空值，在训练集上，采用丢弃行的方式；在测试集，采用均值填充的方式。

聚类时：同样将‘PersonId’与‘Name’属性丢弃，因为这两个属性不适合作为数据的特征。

在聚类中，因测试集缺少‘Survived’属性，所以只在训练集上进行聚类实验。

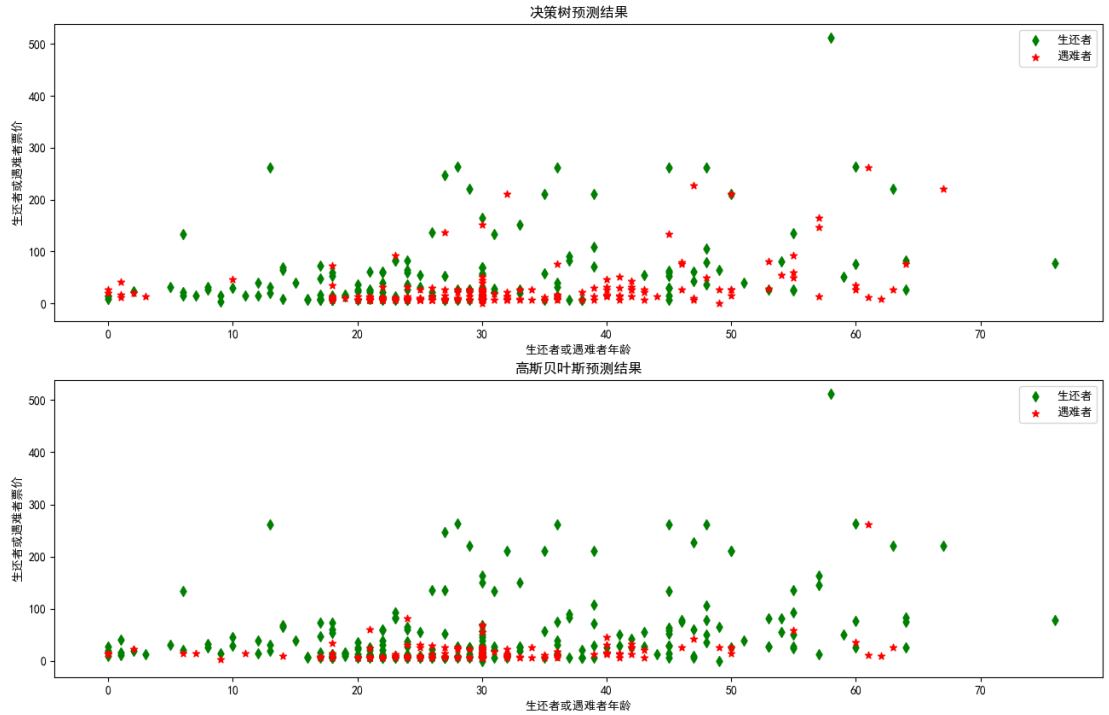
2. 分类结果及可视化

分别训练决策树与朴素贝叶斯分类器，以属性'Survived'作为标签，以属性'Pclass'、'Sex'、'Age'、'SibSp'、'Parch'、'Fare'、'Cabin'、'Embarked'、'Ticket'作为数据的特征，训练二分类模型。设置决策树的最大深度为 10 层，采用高斯函数作为贝叶斯分类器函数。

两个分类器的预测效果如下。

总人数：418 决策树判断生还者：182 贝叶斯判断生还者：268 两者的相同预测数量：290

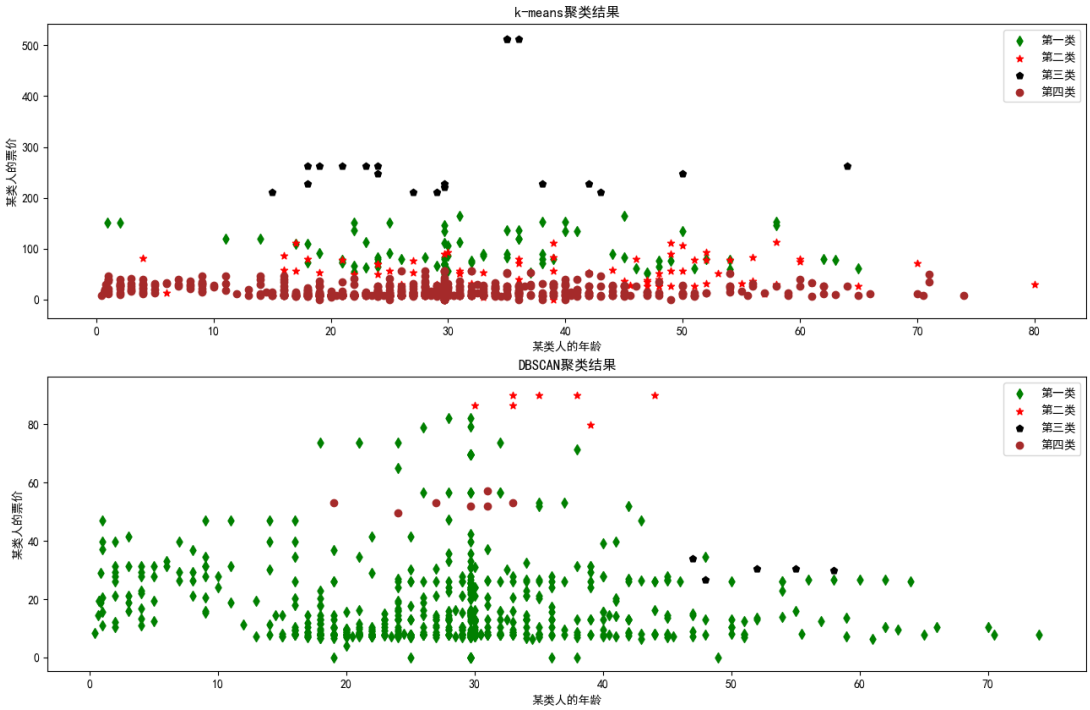
两个分类器预测的生还者、遇难者的年龄与船票价格分布如下。



3. 聚类结果及可视化

因聚类不需要标签，所以把属性'Survived'、'Pclass'、'Sex'、'Age'、'SibSp'、'Parch'、'Fare'、'Cabin'、'Embarked'、'Ticket'均作为数据聚类特征。Kmeans 设置为四个聚类中心，DBSCAN 设置密度半径为 10，密度数量为 5。

两个算法都把训练集的样本聚为四类，每类人的年龄与船票价格分布如下。


k-means聚类结果
第一类
第二类
第三类
第四类
某类人的票价
某类人的年龄
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80

DBSCAN聚类结果
第一类
第二类
第三类
第四类
某类人的票价
某类人的年龄
0
20
40
60
80
0
10
20
30
40
50
60
70