分类与聚类作业报告

数据集：<https://www.kaggle.com/c/titanic/data>

一．预处理操作

分类采用决策树分类器与高斯分布的朴素贝叶斯分类器；聚类采用KMEANS聚类与DBSCAN聚类算法。

对于分类器，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。

对于聚类，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。

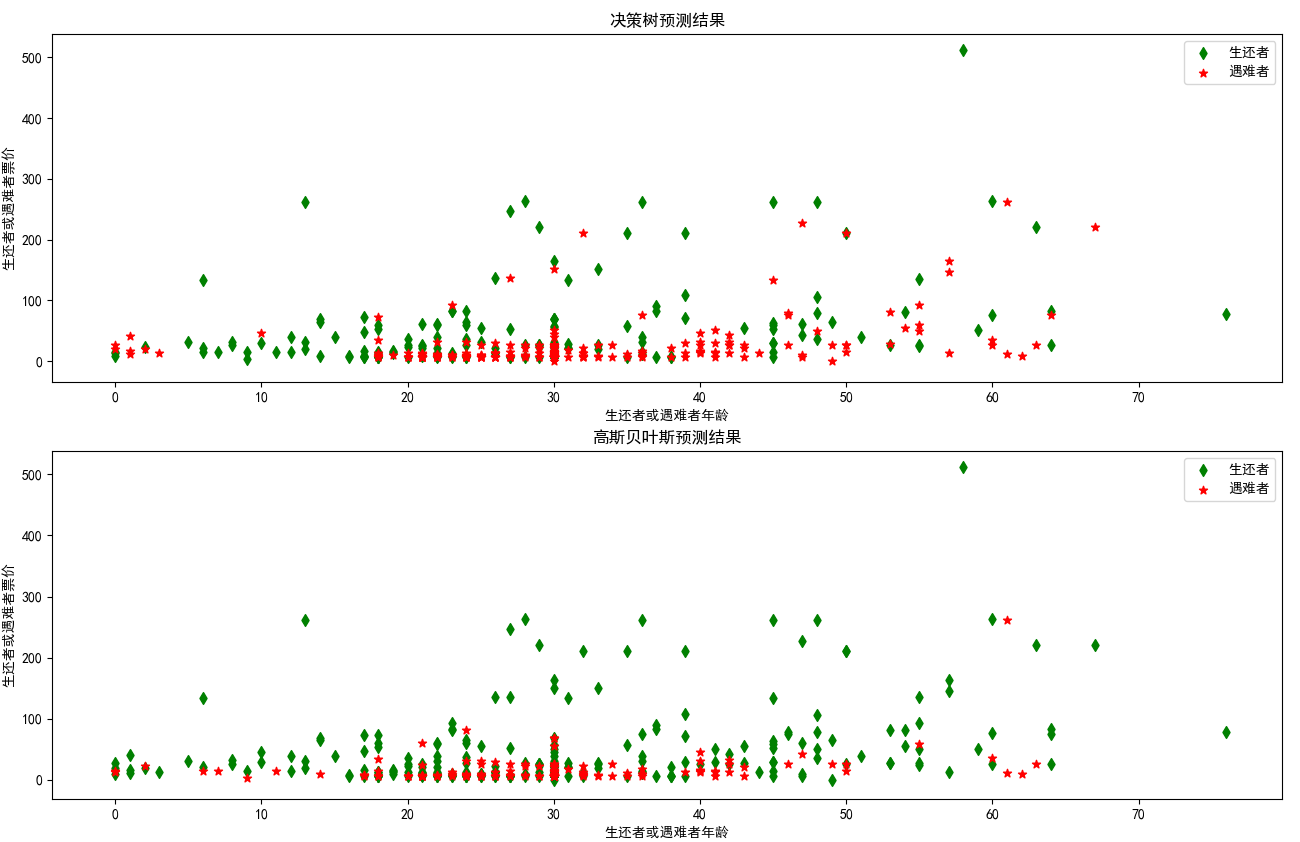
二.分类

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。



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。



三.聚类

聚类采用Kmeans与DBSCAN算法，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。

