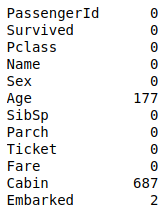
**分类与聚类**

数据源：Titanic [https://www.kaggle.com/c/titanic/data]

1. **数据预处理**

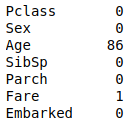
观察数据可以看出，这是泰坦尼克号上乘客的信息，用于预测该是否生还，训练集共有891个样本，有姓名，性别，年龄，船舱类别，船票价格等特征。首先删除部分可挖掘信息较少的属性”Name”,”PassengerId”,“Ticket”，这几个属性对于每个乘客来说都是独特的，从中挖掘信息比较难。

然后统计发现，训练集中“Age”，“Cabin”和“Embarked”属性包含缺失值，见下图：



对于缺失值较多的Cabin属性，选择直接删除。对于“Embarked”属性，观察数据发现这是终点站属性，有“C”, “Q”, “S”三种取值，直接将缺失值填充为“S”值。对于“Age”属性，这是一个连续型数据，选择用均值来填充缺失值。

对于测试集的缺失值，“Age”属性同样使用均值进行填充。“Fare”属性使用中位数进行填充。

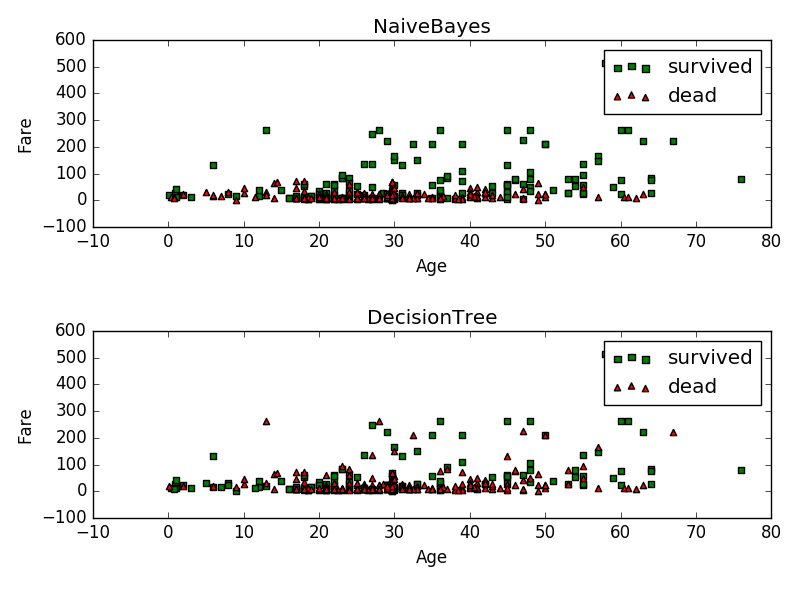


对于字符串类型的的标称属性“Sex”，包含“male”和“female”值，需要将其映射为0和1，便于后续挖掘。同理，“Embarked”也需要将“C”, “Q”, “S”映射到数值0，1，2。

在训练时，将“Survived”属性提取出来，作为标签值。

1. **分类模型及结果**

分类模型选择朴素贝叶斯和决策树，其中决策树的最大深度设置为。选取”Fare”和“Age”两个连续型的数值属性作为X轴和Y轴，便于可视化，分类结果如下图：



1. **聚类模型及结果**

聚类选择K-means和MeanShift，其中，K-means的类别个数，也就是K值设置为5，而MeanShift自动产生的类别个数为6。选取”Fare”和“Age”两个连续型的数值属性作为X轴和Y轴，便于可视化，聚类结果如下图：

