分类与聚类

2120171107

朱佳琪

计算机科学与技术

# 1. 数据源

本次作业中，对下面数据集进行分类和聚类方法分析。

* 数据集: Titanic数据集

# 2. 实验环境

* 语言及环境依赖

语言： Python 3.6

依赖的包：pandas、matplotlib、sklearn

# 3. 实验过程

## 3.1对数据集进行分析和预处理

## 对数据集各属性的分布和缺失情况，选取'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked'七个属性对乘客的生还情况进行预测。

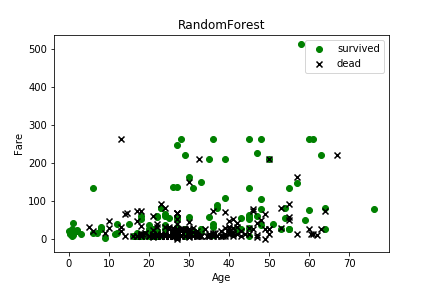
## 其中对数据集的数据分析以训练集为例，如下表所示，容易看出Age列的数据存在缺失值，可以通过中位数来对缺失值进行填补，并且将分类数据用0和1来进行代替，对Sex这一列,将男性用0表示,女性用1表示，对Embarked的缺失值使用S填充。

|  |
| --- |
| PassengerId Survived Pclass Age SibSp |
| count 891.000000 891.000000 891.000000 714.000000 891.000000 |
| mean 446.000000 0.383838 2.308642 29.699118 0.523008 |
| std 257.353842 0.486592 0.836071 14.526497 1.102743 |
| min 1.000000 0.000000 1.000000 0.420000 0.000000 |
| 25% 223.500000 0.000000 2.000000 20.125000 0.000000 |
| 50% 446.000000 0.000000 3.000000 28.000000 0.000000 |
| 75% 668.500000 1.000000 3.000000 38.000000 1.000000 |
| max 891.000000 1.000000 3.000000 80.000000 8.000000 |
| Parch Fare |
| count 891.000000 891.000000 |
| mean 0.381594 32.204208 |
| std 0.806057 49.693429 |
| min 0.000000 0.000000 |
| 25% 0.000000 7.910400 |
| 50% 0.000000 14.454200 |
| 75% 0.000000 31.000000 |
| max 6.000000 512.329200 |

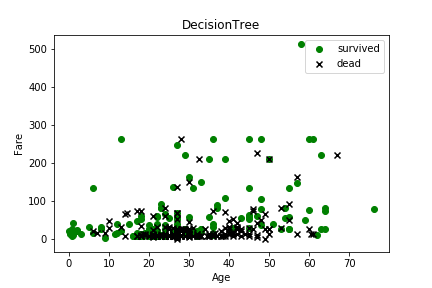
## 对测试集数据进行相同数据预处理操作，其中Fare属性在测试集中存在缺失情况，同样使用中位数进行填充。

## 3.2分类方法

* 在机器学习中，随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。使用RandomForestClassifier方法，将处理过的属性作为特征输入，对Survived属性进行预测。以连续性且分布较为均匀的Age和Fare属性为例，对分类情况进行可视化，如下图所示：



* 在机器学习中，决策树是一个预测模型，代表的是对象属性与对象值之间的一种映射关系。使用DecisionTreeClassifier方法对训练集数据进行训练，对测试集数据进行预测，分类效果如下图所示：



* 两种分类算法的classification\_report如下所示：

RandomForestClassifier:

precision recall f1-score support

0 0.82 0.85 0.83 257

1 0.74 0.70 0.72 161

avg / total 0.79 0.79 0.79 418

DecisionTreeClassifier:

precision recall f1-score support

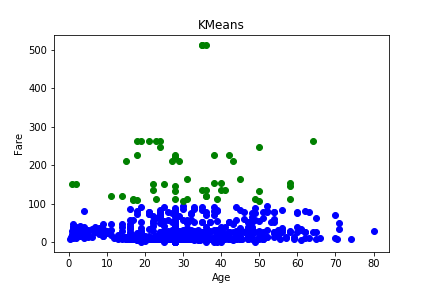
0 0.79 0.86 0.82 244

1 0.77 0.67 0.72 174

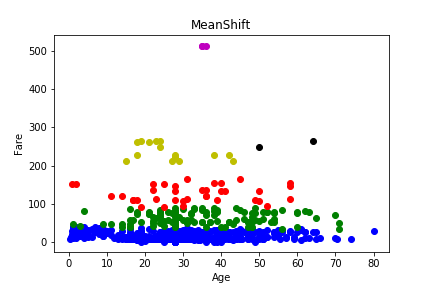
avg / total 0.78 0.78 0.78 418

## 3.3聚类方法

* K-means算法的基本思想是：以空间中k个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。使用K-means算法对训练集进行聚类分析后，效果如下：



* MeanShift算法，是一个迭代的步骤，即先算出当前点的偏移均值，将该点移动到此偏移均值，然后以此为新的起始点，继续移动，直到满足最终的条件。使用MeanShift算法对训练集数据进行聚类分析，形成了6个聚类，效果如下：



* 对对两种聚类算法进行评估，结果如下：

KMeans:

precision recall f1-score support

0 0.97 0.64 0.77 838

1 0.11 0.74 0.20 53

avg / total 0.92 0.64 0.74 891

MeanShift:

precision recall f1-score support

0 0.87 0.68 0.76 704

1 0.22 0.57 0.31 130

2 0.00 0.00 0.00 37

3 0.00 0.00 0.00 15

4 0.00 0.00 0.00 3

5 0.00 0.00 0.00 2

avg / total 0.72 0.62 0.65 891