KMeans实验报告

目录

[数据集来源 1](#_Toc72267746)

[特征和标签的选择 1](#_Toc72267747)

[数据预处理 1](#_Toc72267748)

[填充缺失值 1](#_Toc72267749)

[无量纲化 2](#_Toc72267750)

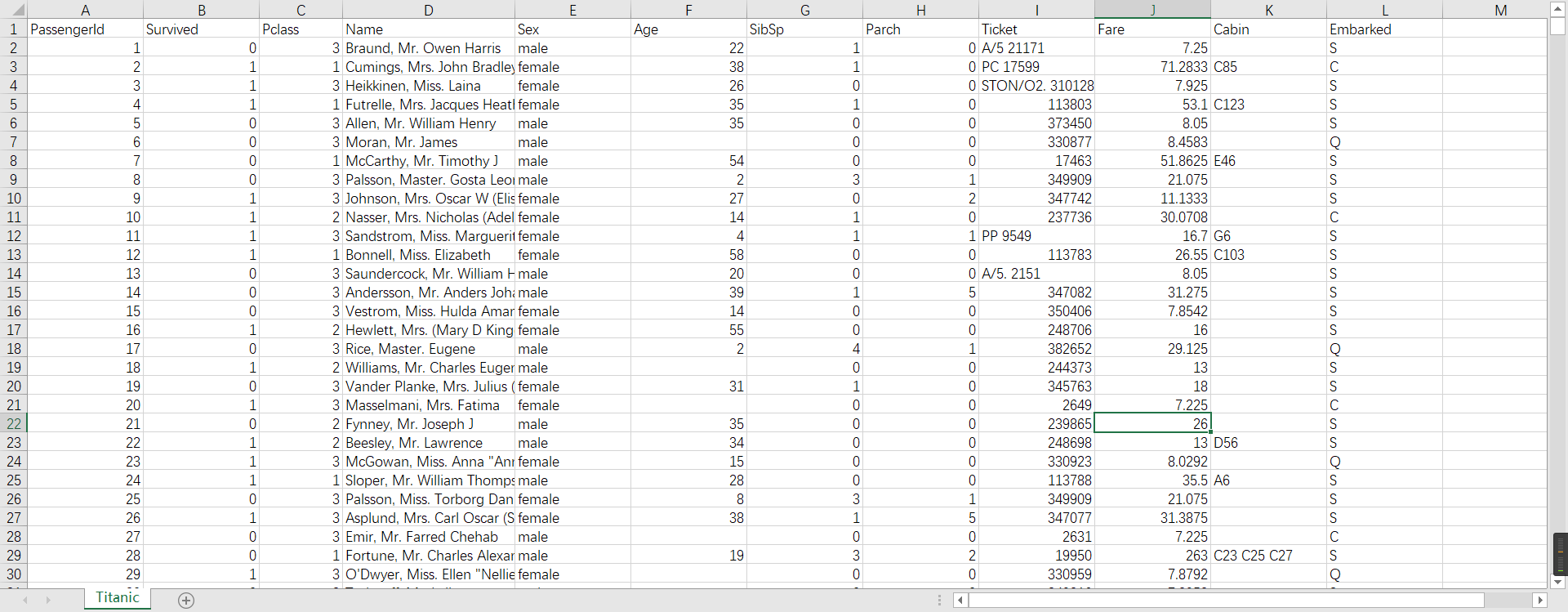
[主体函数 2](#_Toc72267751)

[模型结果： 3](#_Toc72267752)

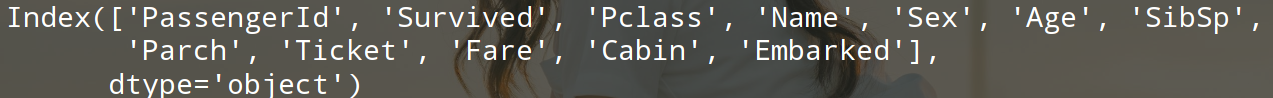
[KMeans模型 3](#_Toc72267753)

[凝聚式层次聚类模型 4](#_Toc72267754)

## 数据集来源

原始的泰坦尼克号的数据集：

特征值：一共包含12个特征。



## 特征和标签的选择

标签："Survived"作为标签列

特征："Name", "Cabin", "Embarked", "Ticket"，'PassengerId'五列的意义相对较小，将其剔除。选择 'Pclass', 'Sex', 'Age', 'SibSp','Parch', 'Fare'六列作为特征列。

由于使用的KMeans聚类算法，属于无监督学习，KMeans模型只需要特征列进行无监督学习，标签列可以用来观察模型的准确率。

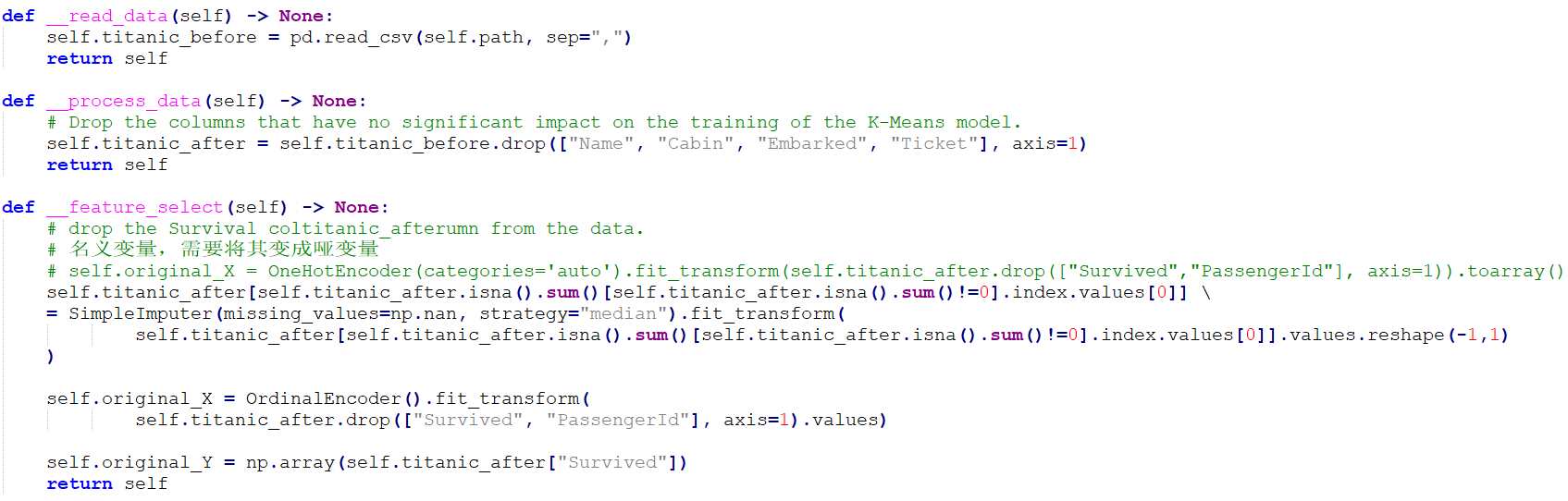
## 数据预处理

### 填充缺失值

由于特征列内含有缺失值，首先使用SimpleImputer方法填充缺失值，由于特征字段包含数值型和字符型，而且根据经验，strategy选用平均数进行填充缺失值相对于平均值和中位数更加合适。

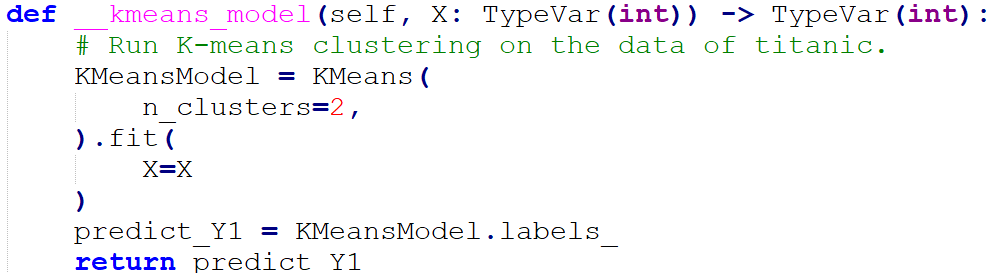
### 无量纲化

由于数据的规格和数据的分布不同，KMeans模型对噪声比较敏感，需要进行无量纲化，使用基于中心化处理和缩放处理的归一化MinMaxScaler模型对特征列进行无量纲化，使得数据分布在[0,1]区间内，并将数据转换成矩阵。

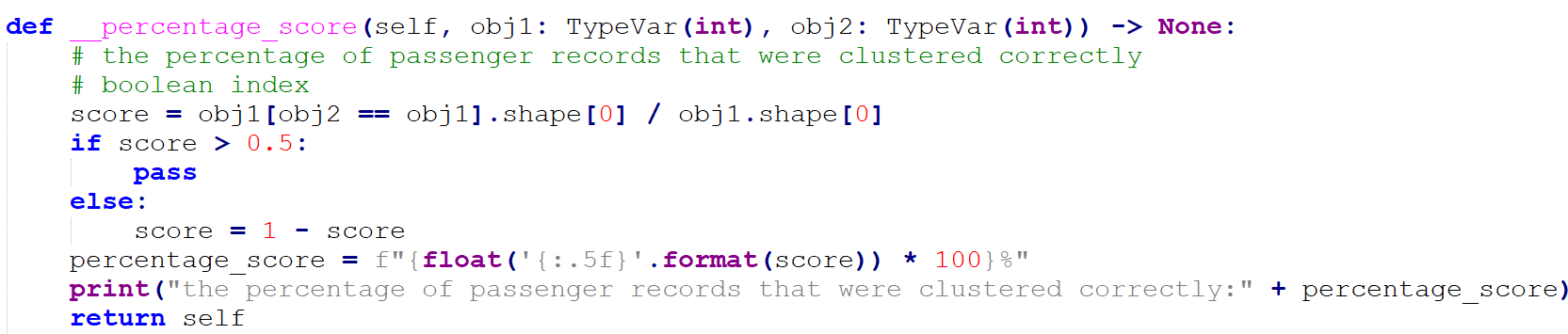


## 主体函数

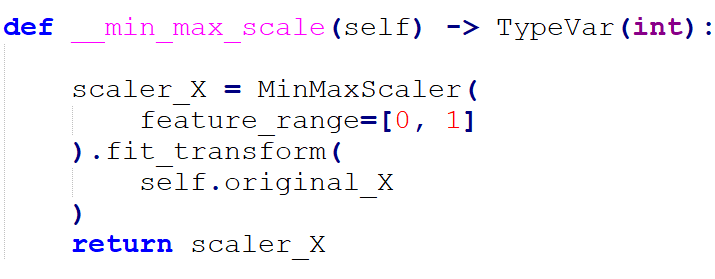
使用KMeans模型训练数据，设置超参数n\_clusters为2，即质心的数量为2。



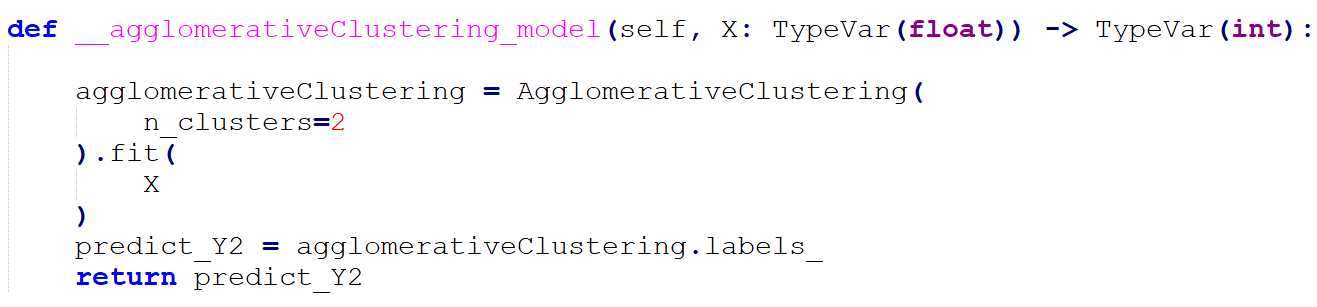
自定义方法对模型进行打分，以百分比返回模型的精确率。在此规定survived是1，not survived 是0，由于模型返回的标签值的不确定性，设计判断分数假如小于0.5，则可以判断模型返回标签的值与我们预定值不符合，只需用1减去得到数值即可得到模型准确率正确结果。



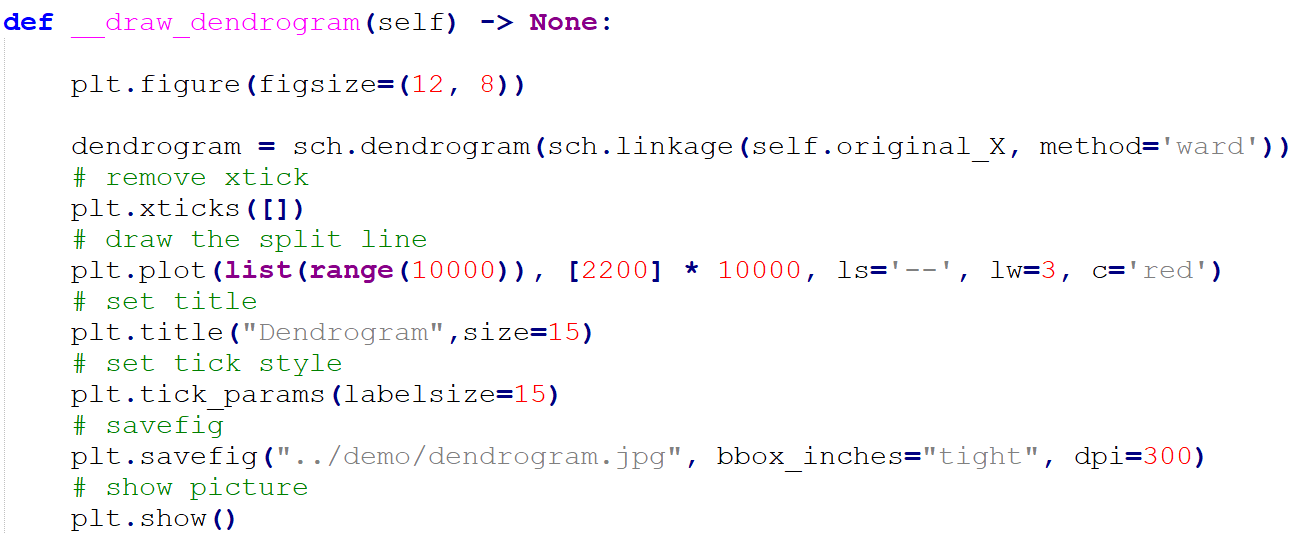
对特征值进行归一化处理，返回归一化处理过后的数据。



实例化凝聚式层次聚类模型，设置超参数n\_clusters为2。



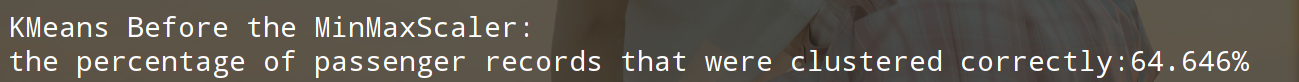
使用scipy模型下的dendrogram函数绘制树形图，这里选择沃德距离进行计算，采用残差平方和SSE，既考虑了簇间的距离，也考虑每个簇内的样本个数。



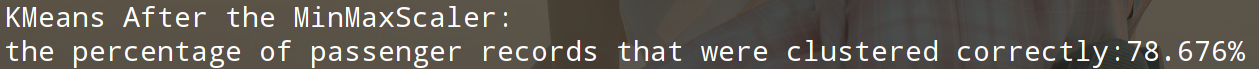
## 模型结果：

### KMeans模型

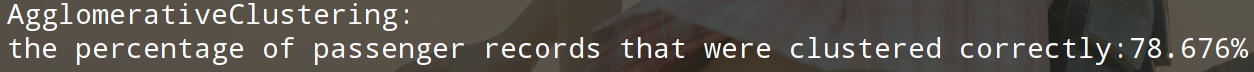
特征未进行归一化处理时，模型的准确率为64.646%



特征进行归一化处理后，模型的准确率为78.676%



### 凝聚式层次聚类模型



树形图：

红色的虚线为模型的切割线。