作业三：分类与聚类

2120171042 芦美艳

数据源

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

数据字典：

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

模型：

1. 分类模型：逻辑回归、svm、决策树
2. 结果

参数：LogisticRegression

(C=1.0, class\_weight=None, dual=False, fit\_intercept=True, intercept\_scaling=1, max\_iter=100, multi\_class='ovr', n\_jobs=1, penalty='l1', random\_state=None, solver='liblinear', tol=1e-06, verbose=0, warm\_start=False)

准确率：accuracy: 0.820627802691

详细结果：

precision recall f1-score support

died 0 0.87 0.84 0.85 140

survive 1 0.74 0.80 0.77 83

avg / total 0.82 0.82 0.82 223

参数：SVC

(C=1.0, cache\_size=200, class\_weight=None, coef0=0.0, decision\_function\_shape='ovr', degree=3, gamma='auto', kernel='rbf', max\_iter=-1, probability=False, random\_state=None, shrinking=True, tol=0.001, verbose=False)

准确率：accuracy: 0.690582959641

详细结果

precision recall f1-score support

died 0 0.88 0.69 0.77 171

survive 1 0.40 0.69 0.51 52

avg / total 0.77 0.69 0.71 223

参数：DecisionTreeClassifier

(class\_weight=None, criterion='gini', max\_depth=None, max\_features=None, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, min\_samples\_leaf=1, min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0, presort=False, random\_state=None, splitter='best')

准确率：accuracy: 0.825112107623

详细结果

precision recall f1-score support

died 0 0.89 0.83 0.86 143

survived 1 0.73 0.81 0.77 80

avg / total 0.83 0.83 0.83 223

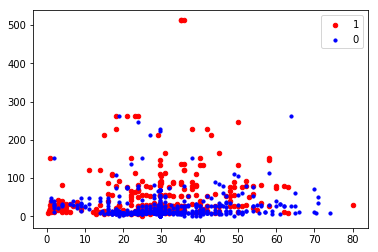
通过比较三个分类算法的结果，可以看出使用决策树得到的结果最好，其准确率最高（0.825），f1值也最好，accuracy和f1成正比关系。

1. 聚类方法：

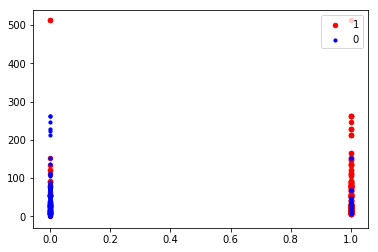
K-means，BSCAN

1. 分析过程

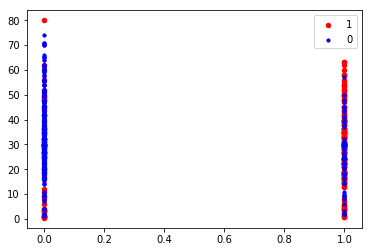
特征分析



按年龄（x轴）和门票价格（y轴）分析survive（红）和died（蓝）的分布



按性别（x轴）和门票价格（y轴）分析survive（红）和died（蓝）的分布

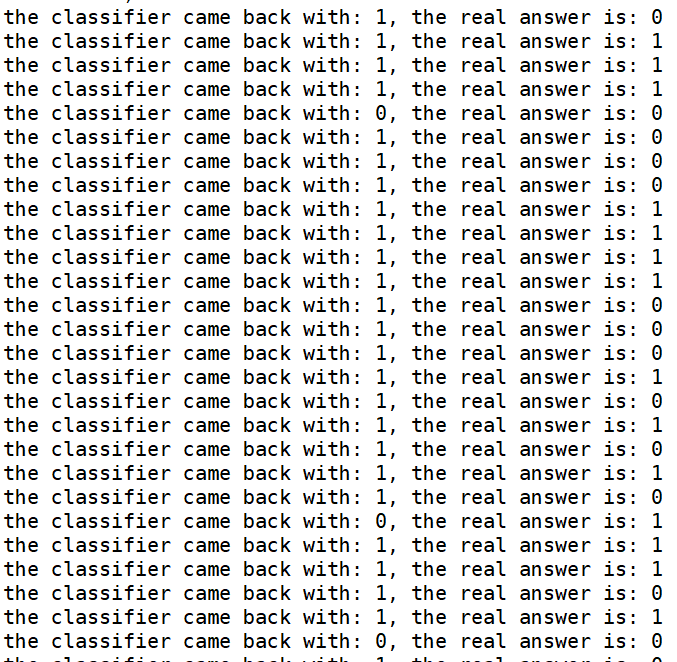


按性别（x轴）和年龄（y轴）分析survive（红）和died（蓝）的分布

可以得出年龄小，门票价格高，女性的存活概率高

使用k-means聚类，发现如果选择较小的k值，其近似误差会小，但估计误差会大，其泛化性能会比较差，即k值越小就意味着整体模型越复杂，容易发生过拟合；相对地，用较大的k值，可以减小学习的估计误差，但是近似误差会增大。该实验选择k=10进行聚类。

部分结果：



错误率为0.23

再用该模型预测测试数据即可

BSCAN分析结果

