**数据挖掘大作业三：分类与聚类分析报告**

**姓名： 康杨**

**学号：2120171024**

# 一．实验环境

本次实验使用环境为Python3.6，使用IDE为Visual Studio Code，使用sklearn进行算法分析，使用pandas进行数据分析，选取Titanic数据集进行数据分析实验。

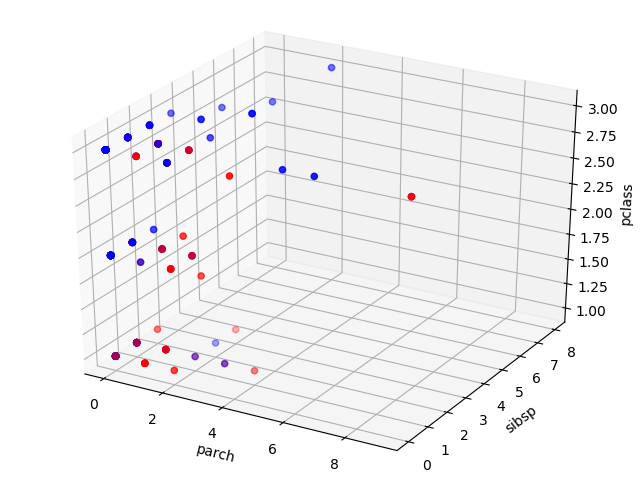
# 二．分类实验

本次分类实验选用CART决策树，随机森林以及梯度提升决策树三个模型分别进行Titanic数据的训练及预测，预测结果的精确度、召回率、F1指数和支持率输出于result目录下的classification\_result.txt，可视化绘图保存于plot目录下。

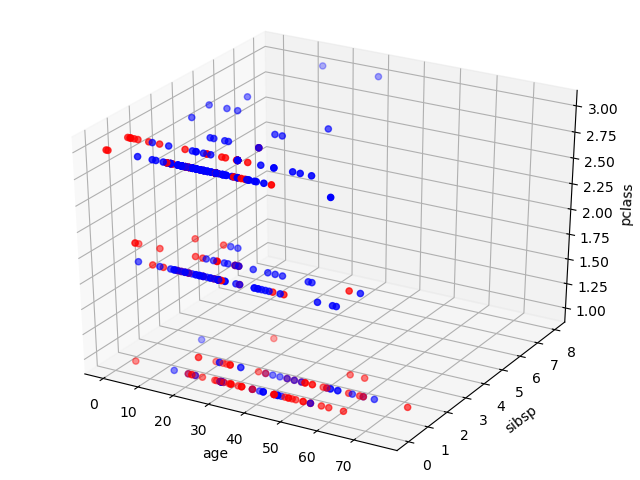
## 2.1 CART决策树

CART分类回归树是一种典型的二叉决策树，可以处理连续型变量和离散型变量。如果待预测分类是离散型数据，则CART生成分类决策树；如果待预测分类是连续型数据，则CART生成回归决策树。

本次实验中，提取每个人的“Sex”,“Embarked”，“PcClass”，“Sibsp”，“Fare”和“Age”六维特征，用来训练CART决策树分类器。结果如下图1和2。



**图1 parch,sibsp,pclass散点图**



**图2 age,sibsp,pclass散点图**

在测试集上的各项指标为：

DecisionTreeClassifier:

precision recall f1-score support

0 0.93 0.94 0.93 263

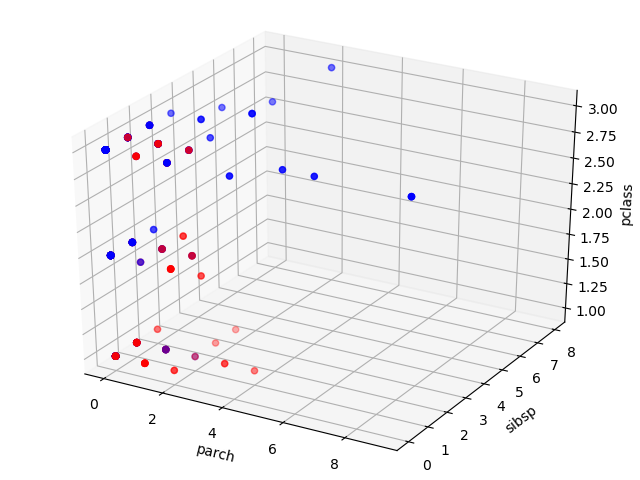
1 0.89 0.88 0.89 155

avg / total 0.92 0.92 0.92 418

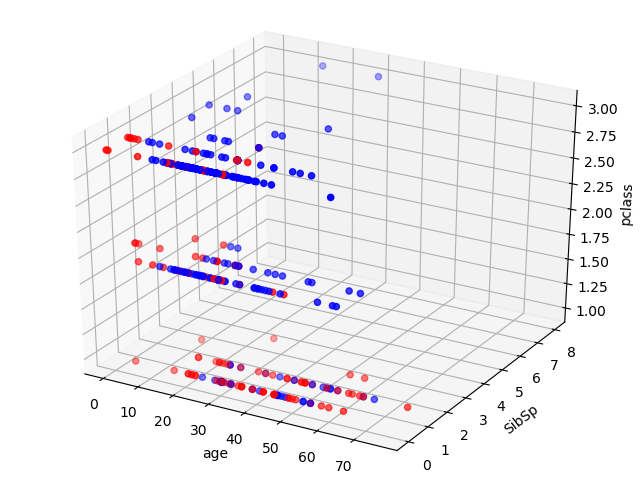
## 2.2 随机森林

在机器学习中，随机森林（Random Forest）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。Leo Breiman和Adele Cutler发展出推论出随机森林的算法。

本次实验中，提取每个人的“Sex”，“Embarked”，“PcClass”，“Sibsp”，“Fare”和“Age”六维特征，用来训练随机森林分类器。结果如下图3和4。



**图3 parch,sibsp,pclass散点图**



**图4 age,sibsp,pclass散点图**

在测试集上的各项指标为：

RandomForestClassifier:

precision recall f1-score support

0 0.83 0.83 0.83 268

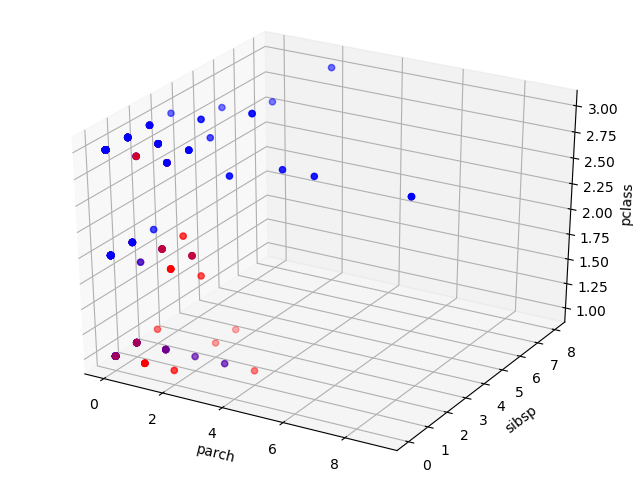
1 0.70 0.71 0.70 150

avg / total 0.79 0.78 0.78 418

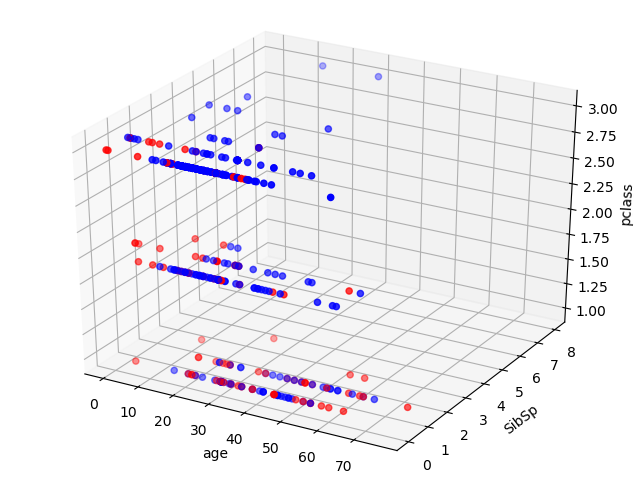
## 2.3 梯度提升决策树

梯度提升决策树GBDT 又叫 MART（Multiple Additive Regression Tree)，是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来做最终答案。它在被提出之初就和SVM一起被认为是泛化能力（generalization)较强的算法。

本次实验中，提取每个人的“Sex”，“Embarked”，“PcClass”，“Sibsp”，“Fare”和“Age”六维特征，用来训练随机森林分类器。结果如下图5和6。



**图5 parch,sibsp,pclass散点图**



**图6 age,sibsp,pclass散点图**

在测试集上的各项指标为：

GradientBoostingClassifier:

precision recall f1-score support

0 0.96 0.86 0.91 298

1 0.72 0.92 0.81 120

avg / total 0.89 0.88 0.88 418

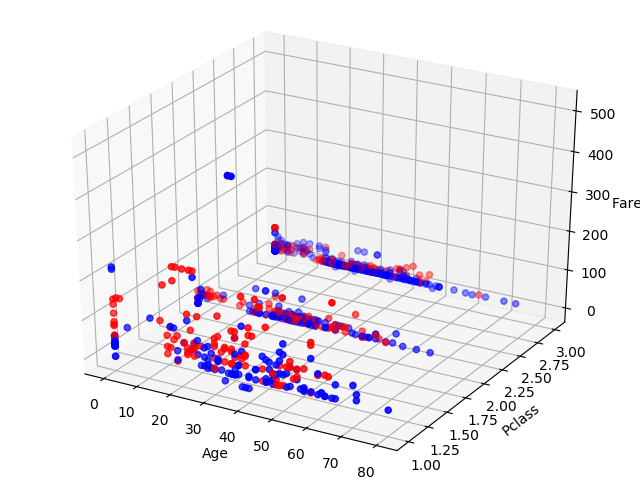
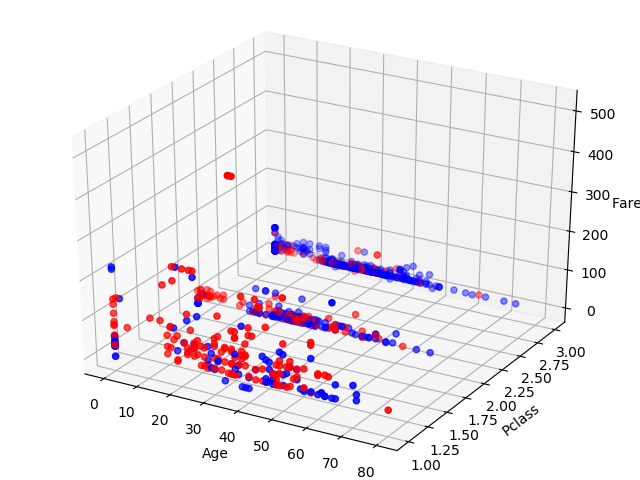
# 三．聚类实验

本次聚类实验选用K-means，MeanShift以及MiniBatchKMeans三个模型分别进行Titanic数据的训练及预测，预测结果的精确度、召回率、F1指数和支持率输出于result目录下的cluter\_result.txt，可视化绘图保存于plot目录下。

## 3.1 K-means

K-means算法是最为经典的基于划分的聚类方法，是十大经典数据挖掘算法之一。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。

本次实验中，提取每个人的“Sex”,“Embarked”，“PcClass”，“Sibsp”，“Fare”和“Age”等九维特征（其中标称属性均转化为数值），用来训练K-means聚类器。和正确的分类对比如下图7。



**图7 age，pclass, fare散点图对比**

与原始数据对比的各项指标为：

KMeans:

precision recall f1-score support

0 0.84 0.81 0.83 571

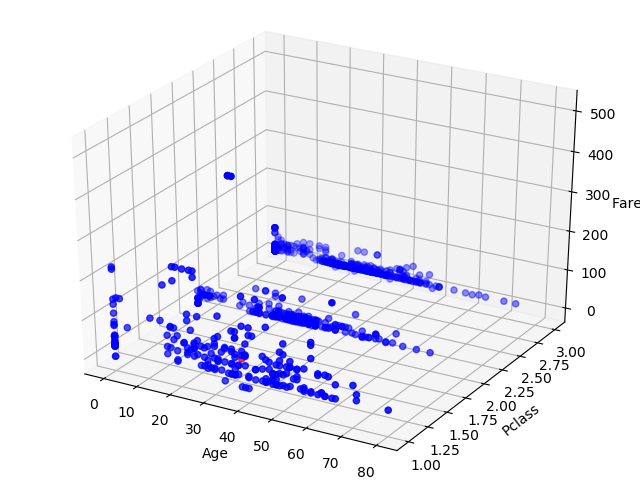
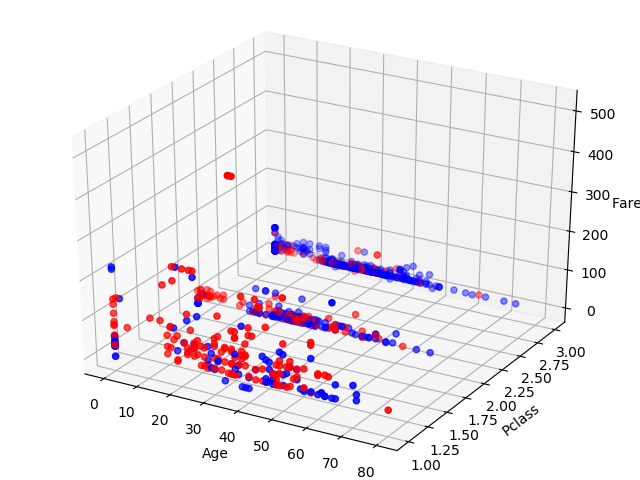
1 0.68 0.73 0.70 320

avg / total 0.78 0.78 0.78 891

## 3.2 MeanShift

Mean Shift算法：指一个迭代的步骤，即先算出当前点的偏移均值，移动该点到其偏移均值，然后以此为新的起始点，继续移动，直到满足一定的条件结束。

本次实验中，提取每个人的“Sex”，“Embarked”，“PcClass”，“Sibsp”，“Fare”和“Age”等九维特征（其中标称属性均转化为数值），用来训练Mean Shift聚类器。和正确的分类对比如下图8。



**图8 age，pclass, fare散点图对比**

与原始数据对比的各项指标为：

MeanShift:

precision recall f1-score support

0 0.94 0.62 0.75 824

1 0.00 1.00 0.01 1

2 0.00 0.00 0.00 1

3 0.00 0.00 0.00 1

4 0.00 0.00 0.00 1

5 0.00 0.00 0.00 1

6 0.00 0.00 0.00 1

7 0.00 0.00 0.00 1

8 0.00 0.00 0.00 1

9 0.00 0.00 0.00 1

10 0.00 0.00 0.00 1

11 0.00 0.00 0.00 1

12 0.00 0.00 0.00 1

13 0.00 0.00 0.00 1

14 0.00 0.00 0.00 1

15 0.00 0.00 0.00 1

16 0.00 0.00 0.00 1

17 0.00 0.00 0.00 1

18 0.00 0.00 0.00 1

19 0.00 0.00 0.00 1

20 0.00 0.00 0.00 1

21 0.00 0.00 0.00 1

22 0.00 0.00 0.00 1

23 0.00 0.00 0.00 1

24 0.00 0.00 0.00 1

25 0.00 0.00 0.00 1

26 0.00 0.00 0.00 1

27 0.00 0.00 0.00 1

28 0.00 0.00 0.00 1

29 0.00 0.00 0.00 1

30 0.00 0.00 0.00 1

31 0.00 0.00 0.00 1

32 0.00 0.00 0.00 1

33 0.00 0.00 0.00 1

34 0.00 0.00 0.00 1

35 0.00 0.00 0.00 1

36 0.00 0.00 0.00 1

37 0.00 0.00 0.00 1

38 0.00 0.00 0.00 1

39 0.00 0.00 0.00 1

40 0.00 0.00 0.00 1

41 0.00 0.00 0.00 1

42 0.00 0.00 0.00 1

43 0.00 0.00 0.00 1

44 0.00 0.00 0.00 1

45 0.00 0.00 0.00 1

46 0.00 0.00 0.00 1

47 0.00 0.00 0.00 1

48 0.00 0.00 0.00 1

49 0.00 0.00 0.00 1

50 0.00 0.00 0.00 1

51 0.00 0.00 0.00 1

52 0.00 0.00 0.00 1

53 0.00 0.00 0.00 1

54 0.00 0.00 0.00 1

55 0.00 0.00 0.00 1

56 0.00 0.00 0.00 1

57 0.00 0.00 0.00 1

58 0.00 0.00 0.00 1

59 0.00 0.00 0.00 1

60 0.00 0.00 0.00 1

61 0.00 0.00 0.00 1

62 0.00 0.00 0.00 1

63 0.00 0.00 0.00 1

64 0.00 0.00 0.00 1

65 0.00 0.00 0.00 1

66 0.00 0.00 0.00 1

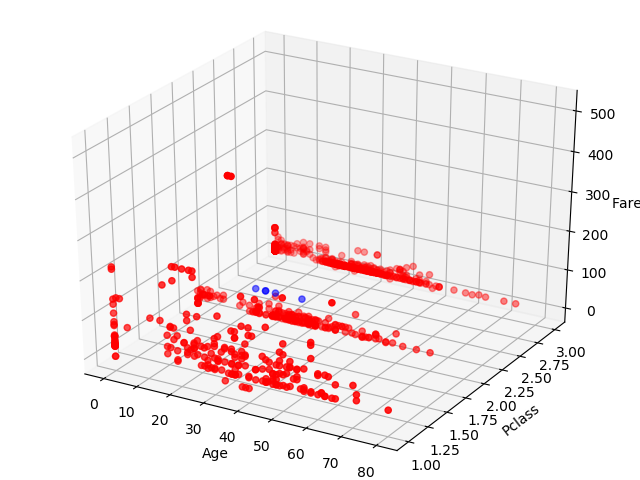
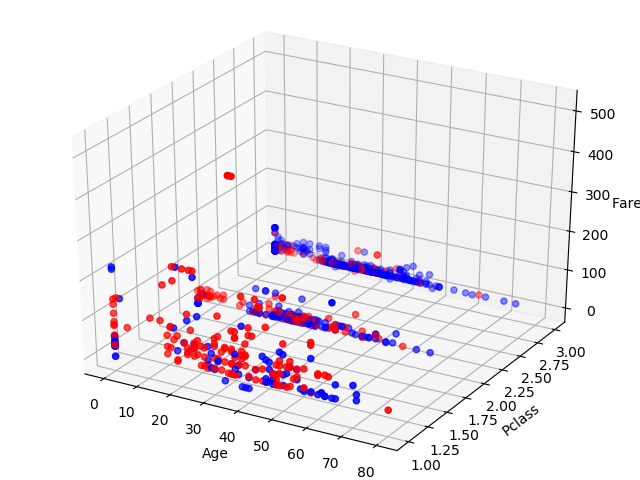
67 0.00 0.00 0.00 1

avg / total 0.87 0.58 0.69 891

## 3.3 MiniBatchKMeans

Mini Batch K-Means算法是K-Means算法的变种，采用小批量的数据子集减小计算时间，同时仍试图优化目标函数，这里所谓的小批量是指每次训练算法时所随机抽取的数据子集，采用这些随机产生的子集进行训练算法，大大减小了计算时间，与其他算法相比，减少了k-均值的收敛时间，小批量k-均值产生的结果，一般只略差于标准算法。

本次实验中，提取每个人的“Sex”，“Embarked”，“PcClass”，“Sibsp”，“Fare”和“Age”等九维特征（其中标称属性均转化为数值），用来训练Mean Shift聚类器。和正确的分类对比如下图9。



**图9 age，pclass, fare散点图对比**

与原始数据对比的各项指标为：

MiniBatchKMeans:

precision recall f1-score support

0 0.01 1.00 0.02 5

1 1.00 0.39 0.56 886

avg / total 0.99 0.39 0.55 891

# 四．实验总结

本次实验通过编写程序实现了Titanic数据集的分类和聚类。通过sklearn中的特征提取功能将标称属性转化为数值，实现了分类器的训练和数据的非监督聚类。通过此次实验，学习并实践了经典的分类模型CART决策树、随机森林和梯度提升决策树，经典的聚类算法K-means、MeanShift和MiniBatchKMeans，详细地了解了算法的各个步骤。深入理解了分类和聚类算法在数据挖掘中的原理和作用。