数据挖掘 ---- 作业三

程伟 2120160984 第10组

# 要求：

1. 使用分类模型（至少2个）对数据集进行挖掘；
2. 对挖掘结果进行可视化，并解释其意义；
3. 使用聚类方法（至少2种）对数据集进行分析；
4. 对挖掘结果进行可视化，并解释其意义。

# 实验环境：

本实验程序基于python开发，使用的第三方库如下：

pandas、numpy -- 分析数据，并格式化存储数据

sklearn – 分类聚类模型库

matplotlib – 绘制图表

# 实验结果：

本次实验选择第三个数据集<https://www.kaggle.com/c/titanic/data>，该数据各列含义描述如下表所示：

|  |  |  |
| --- | --- | --- |
| 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 |

## 使用分类模型（至少2个）对数据集进行挖掘

本次实验使用了决策树、svm，以及朴素贝叶斯三个分类模型。其中决策树模型设定树的最大深度为5时效果最好，svm模型使用rbf核最好，而朴素贝叶斯模型使用高斯朴素贝叶斯最好。

## 对挖掘结果进行可视化，并解释其意义

为了最终的可视化，这里将原数据（共计12维）降到2维，并绘制分类模型的分隔面。原数据包含以下属性：'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'。其中'Survived'是最终需要预测的label，由统计结果得知 'PassengerId', 'Name'两个属性 对于每个对象都有不同的值，该属性对于模型和任务没有任何帮助，所以在后面的建模过程中去除了该属性。另外虽然'Ticket'属性值不是一一对应乘客，但是实际情况应该是一一对应(满射)的，之所以不是应该是数据统计上的失误，也加以去除。

最终对模型有意义的属性为以下8个：'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Cabin', 'Embarked'。

#### 决策树模型

图1是决策树模型的建模结果，其中样本由点绘出，分隔面有边绘出。红绿两种颜色的点表示的是两种不同类别的样本，浅红浅绿两种颜色的区域代表的是分类模型预测的类别。红色的点出现在浅红区域或者绿色的点出现在浅绿区域，则意味着模型预测正确，反之则意味着模型预测错误。

总的来说，该模型具有一定的区分度，但由于这是原数据降维后的建模结果，所以效果比不上原数据建模的结果。

图2是原数据建模生成的判别树(决策树)，X[i]代表的就是样本第i个属性。除去根节点，数的深度为5。

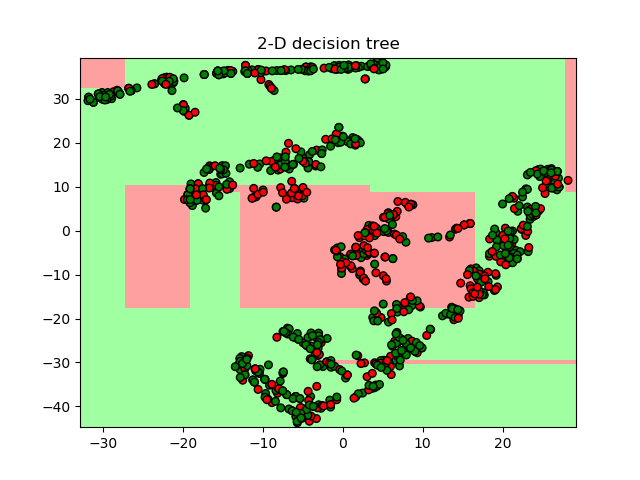


图1

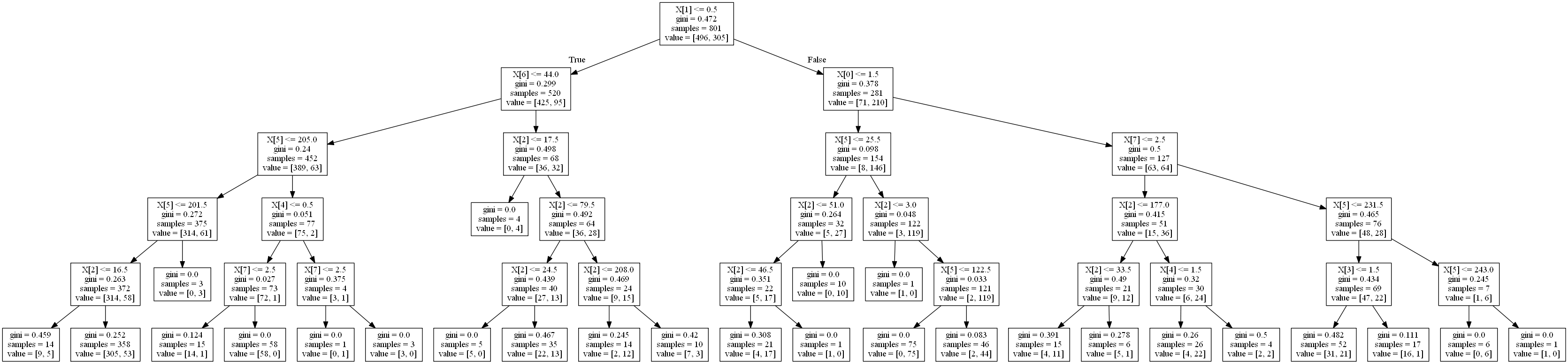


图2

图3是SVM模型的建模结果，其中样本由点绘出，分隔面由边绘出。红绿两种颜色的点表示的是两种不同类别的样本，浅红浅绿两种颜色的区域代表的是分类模型预测的类别。红色的点出现在浅红区域或者绿色的点出现在浅绿区域，则意味着模型预测正确，反之则意味着模型预测错误。

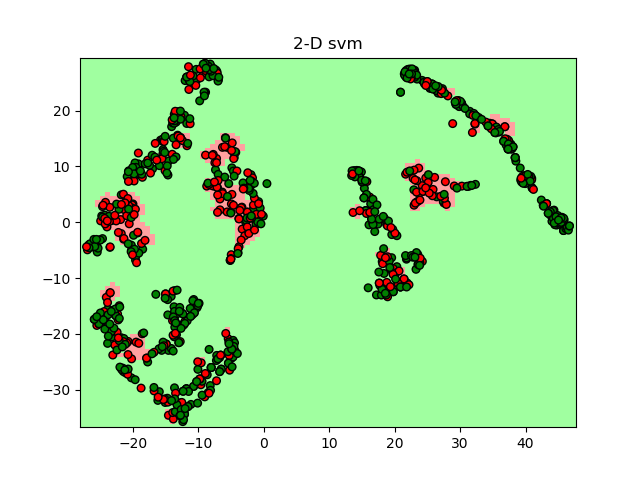


图3

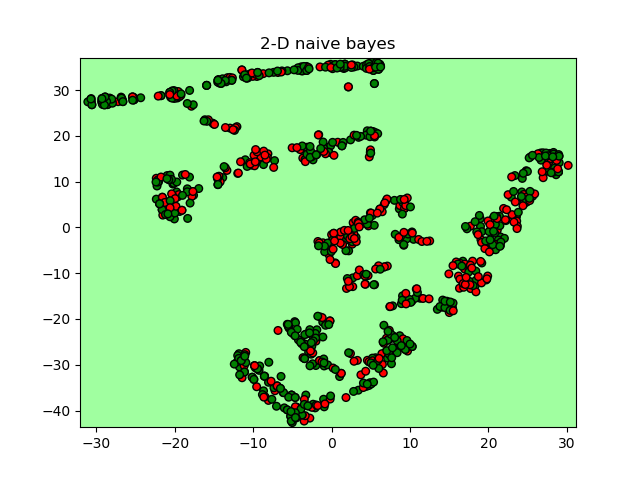


图4

总的来说，该模型具有一定的区分度，但由于这是原数据降维后的建模结果，所以效果比不上原数据建模的结果。

图4是高斯贝叶斯模型的建模结果，其中样本由点绘出，分隔面有边绘出。红绿两种颜色的点表示的是两种不同类别的样本，浅红浅绿两种颜色的区域代表的是分类模型预测的类别。红色的点出现在浅红区域或者绿色的点出现在浅绿区域，则意味着模型预测正确，反之则意味着模型预测错误。

总的来说，该模型没有区分度，该模型将所有数据预测为正样本(对应绿色)，这样一来模型的准确率将由正样本的比例决定。考虑到这是原数据降维后的建模结果，所以不同等效认为原数据建模的结果也没有区分度。

## 使用聚类模型（至少2个）对数据集进行挖掘

本次实验使用了k-means、batch k-means原型聚类，以及DBSCAN密度聚类等三个聚类模型。其中原型聚类设定簇的数量为2，对应样本label的数量。而密度聚类设定的最小距离为0.3，类最小样本数量为10.

## 对挖掘结果进行可视化，并解释其意义

为了最终的可视化，这里将原数据（共计12维）降到2维，并绘制分类模型的分隔面。原数据包含以下属性：'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'。其中'Survived'是最终需要预测的label，由统计结果得知 'PassengerId', 'Name'两个属性 对于每个对象都有不同的值，该属性对于模型和任务没有任何帮助，所以在后面的建模过程中去除了该属性。另外虽然'Ticket'属性值不是一一对应乘客，但是实际情况应该是一一对应(满射)的，之所以不是应该是数据统计上的失误，也加以去除。

最终对模型有意义的属性为以下8个：'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Cabin', 'Embarked'。

图5是原数据降维后的二维散点图，其中红绿两种颜色的点表示的是两种不同类别的样本。

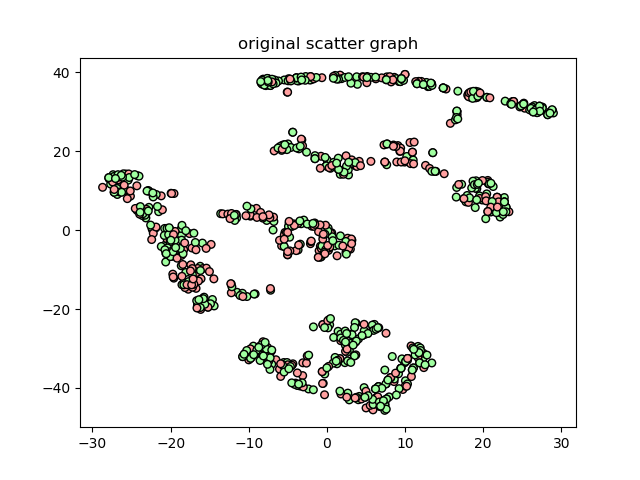


图5

#### k-means

图6是k-means聚类后的结果，其中红绿两种颜色的点表示的是两种不同类别的样本。与图5比较可知该模型聚类的效果距离期望差距较大。

图7是batch k-means聚类后的结果，其中红绿两种颜色的点表示的是两种不同类别的样本。该结果与kmeans一样，除了聚类中心有些差别。与图5比较可知该模型聚类的效果距离期望差距较大。

图8 是 dbscan聚类后的结果，不同的颜色代表不同的类别。因为该模型最终产生的簇数量不定，所以很难将其与原数据对比。但从样本点之间的稠密度来看，该模型确实对样本有较好的区分。

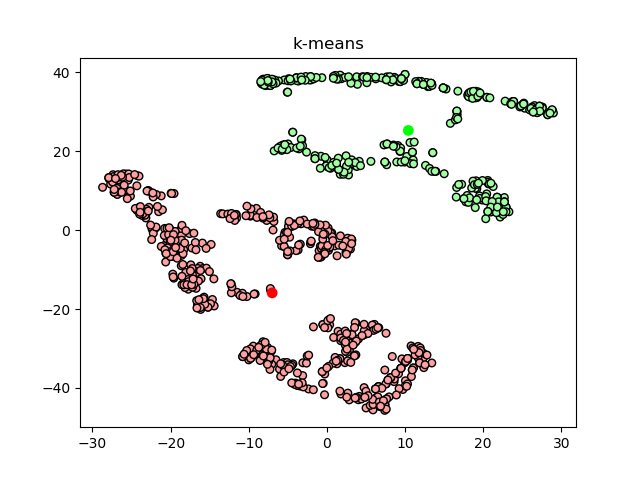


图6

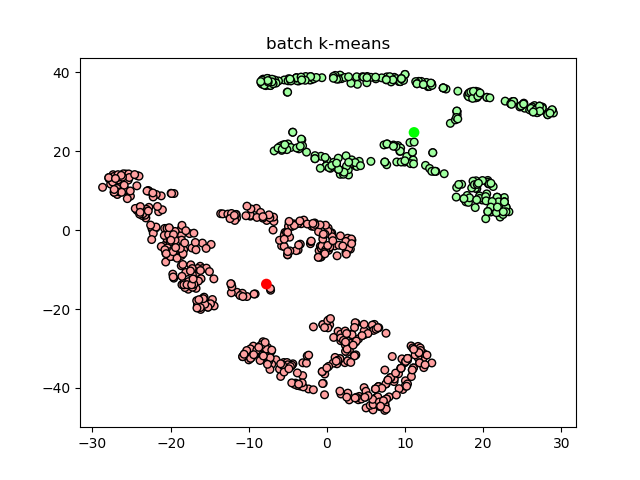


图7

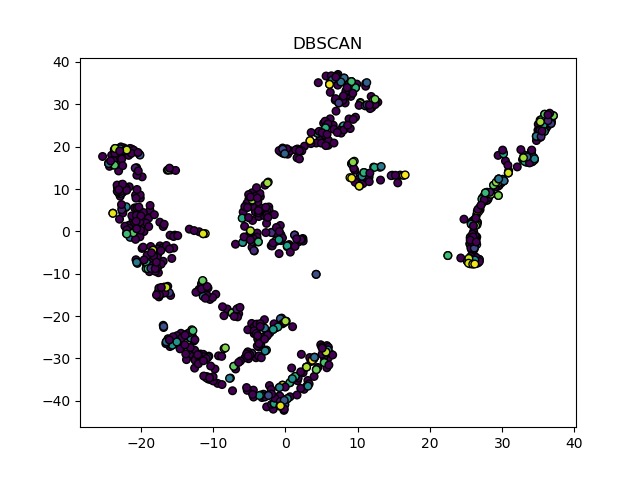


图8