数据挖掘大作业三：分类与聚类

数据源

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

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。

实验过程

本实验在三个给定的数据集中选定Titanic数据集，数据集内包含PassengerId，Survived，Pclass，Name，Sex，Age，SibSp，Parch，Ticket，Fare，Cabin，Embarked十二项，它们的含义如下：

|  |  |
| --- | --- |
| 变量名 | 变量描述 |
| PassengerId | 乘客的编号 |
| Survived | 是否存活（0代表没有存活，1代表存活） |
| Pclass | 客舱等级（1,2,3分别代表一二三登记） |
| Name | 姓名 |
| Sex | 性别 |
| Age | 年龄 |
| SibSp | 兄弟姐妹或者配偶同在船上的数目 |
| Parch | 父母和孩子同在船上的数目 |
| Ticket | 船票类型 |
| Fare | 票的费用 |
| Cabin | 客舱号 |
| Embarked | 登船地点(C = Cherbourg; Q = Queenstown; S = Southampton) |

在本实验中，我们着重探索乘客的客舱等级、年龄和性别对于是否存货下来的影响。据此先对数据集进行相应的预处理，从原先的数据集中挑选出所需要的属性，构成新的数据集。代码如下：

library("dplyr")

train\_data <- select(train\_data,Survived, Pclass, Sex, Age)

test\_data <- select(test\_data,Pclass, Sex, Age)

然后对于缺失数据的行做删除处理

train\_data <- na.omit(train\_data)

test\_data <- na.omit(test\_data)

为了更方便的分析，将年龄这个属性划分成四个阶段：儿童（0~10），用1表示； 青年（10~20），用2表示；中年（20~50），用3表示； 老年（>50）,用4表示。具体的实现代码如下;

# 将年龄这一属性划分成四个阶段（儿童，青年，成年，老年）

num = as.numeric(as.character(train\_data$Age))

train\_data[num <10, 'Age'] = 1

train\_data[num >= 10 & num <20, 'Age'] = 2

train\_data[num >= 20 & num <50, 'Age'] = 3

train\_data[num >= 50, 'Age'] = 4

实验要求分别用2种分类模型和2种聚类方法对数据集进行处理。在实验中采用的分类方法分别是：朴素贝叶斯分类器和SVM分类器；采用的聚类方法分别是：kmeans方法和层次聚类的方法。下面分别介绍它们。

**分类模型进行数据挖掘**

**朴素贝叶斯分类器**

首先导入相应的包

library("e1071")

贝叶斯分类器的实现代码为：

#朴素贝叶斯分类器

mod.formula <- as.formula("Survived~Pclass+Sex+Age")

nb.sol <- naiveBayes(mod.formula, train\_data)

nb.predict <- predict(nb.sol, newdata = test\_data)

Survived = as.factor(nb.predict)

Bayes\_Predict <- test\_data;

Bayes\_Predict <- data.frame(Survived, Bayes\_Predict);

#View(Bayes\_Predict)

write.csv(Bayes\_Predict, file = "Bayes\_Predict.csv", row.names = F)

然后对贝叶斯分类器得到的分类结果进行可视化：

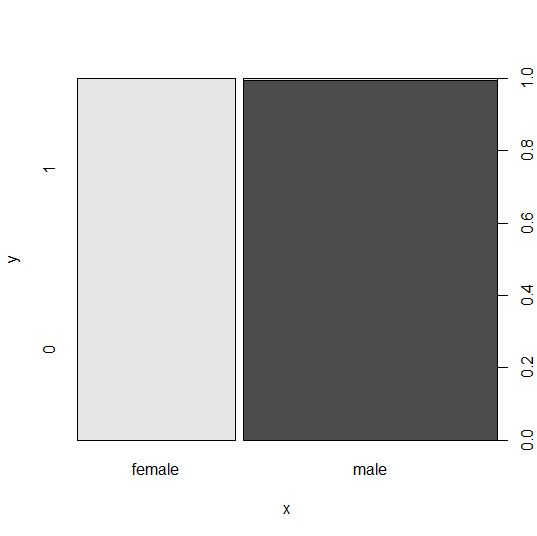
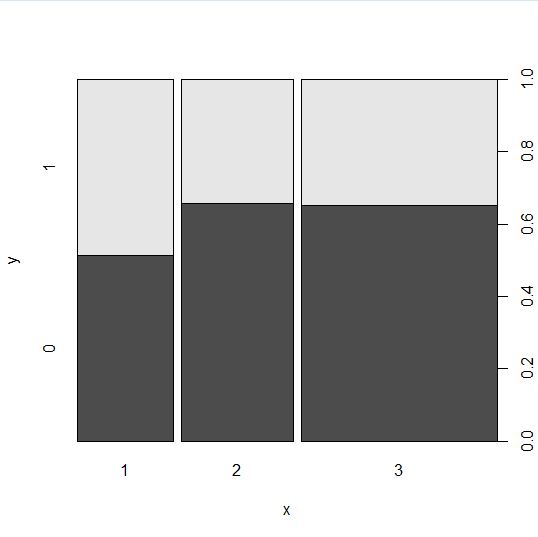
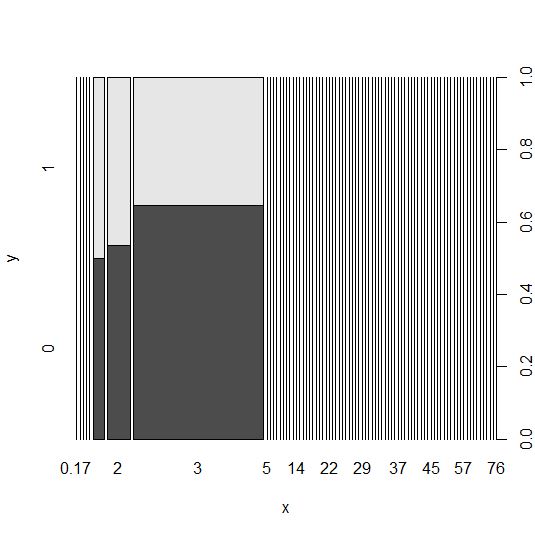
# 可视化朴素贝叶斯分类器的效果

plot(Bayes\_Predict$Pclass, Bayes\_Predict$Survived)

plot(Bayes\_Predict$Sex, Bayes\_Predict$Survived)

plot(Bayes\_Predict$Age, Bayes\_Predict$Survived)

可视化的结果如下：



实验结果表明，生还的乘客中，年纪小的居多；头等舱的乘客的生存率高于二等舱和三等舱；生存下来的乘客中女性居多。

**SVM分类器**

首先导入相应的包

library("e1071")

贝叶斯分类器的实现代码为：

#朴素贝叶斯分类器

mod.formula <- as.formula("Survived~Pclass+Sex+Age")

svm.sol <- svm(mod.formula, train\_data)

svm.predict <- predict(svm.sol, train\_data)

Survived = as.factor(svm.predict)

Svm\_Predict <- select(train\_data,Pclass, Sex, Age);

Svm\_Predict <- data.frame(Survived, Svm\_Predict);

write.csv(Svm\_Predict, file = "Svm\_Predict.csv", row.names = F)write.csv(Bayes\_Predict, file = "Bayes\_Predict.csv", row.names = F)

然后对贝叶斯分类器得到的分类结果进行可视化：

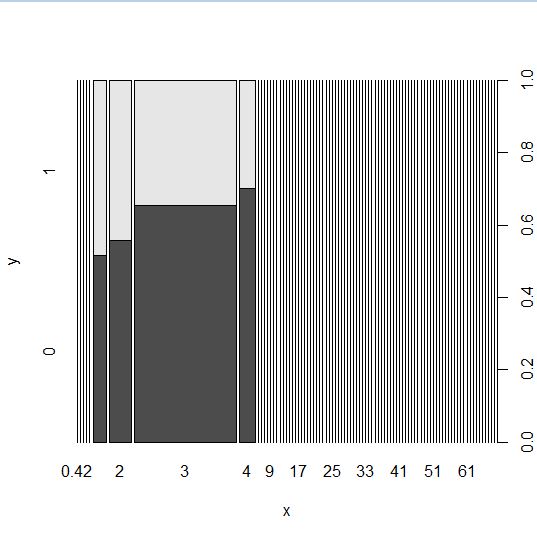
# 可视化朴素贝叶斯分类器的效果

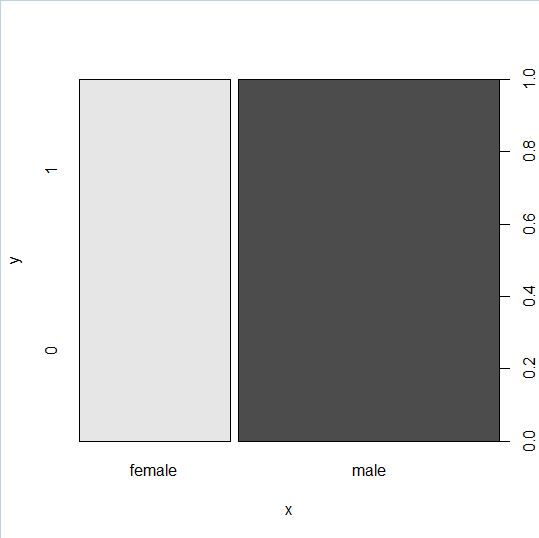
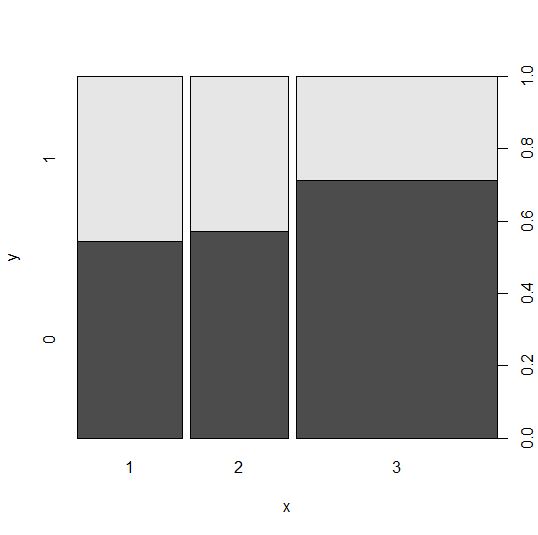
plot(Bayes\_Predict$Pclass, Bayes\_Predict$Survived)

plot(Bayes\_Predict$Sex, Bayes\_Predict$Survived)

plot(Bayes\_Predict$Age, Bayes\_Predict$Survived)

可视化的结果如下：





实验结果与贝叶斯分类器的结果类似，生还的乘客中，年纪小的居多；头等舱的乘客的生存率高于二等舱和三等舱；生存下来的乘客中女性居多。

**聚类模型进行数据挖掘**

**Kmeans聚类**

首先导入相应的包

library("kernlab")

library("magrittr")

分别分析了生存与年龄的关系以及生存与客舱等级的关系，代码如下：

# 生存与年龄

cluster\_data <- select(train\_data, Survived, Age)

res <- kmeans(cluster\_data,4)

plot(cluster\_data, col = res$cluster)

#jpeg(file = paste('survived\_age','.jpg'))

#生存与仓次

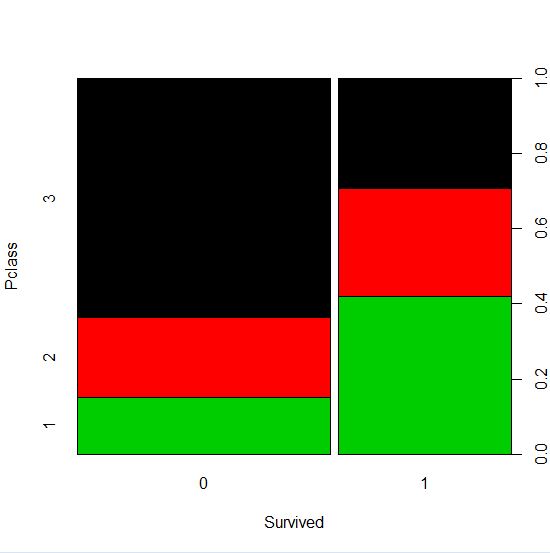
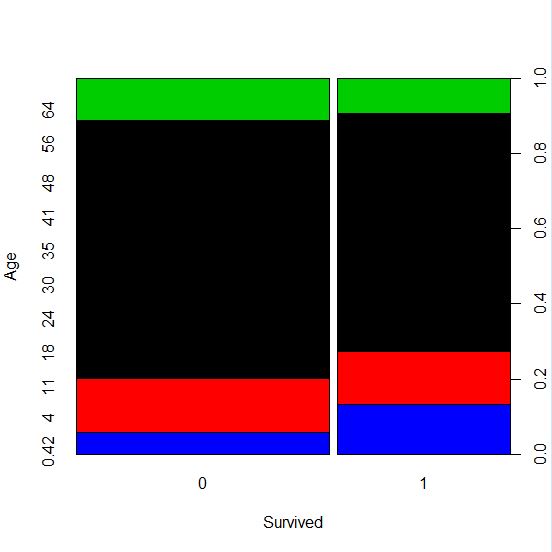
cluster\_data <- select(train\_data, Survived, Pclass)

res <- kmeans(cluster\_data,3)

plot(cluster\_data, col = res$cluster)

#jpeg(file = paste('survived\_pclass','.jpg'))

实验结果如下：



实验结果表明，年龄对于生存率影响不是很明显，不过年纪小的生还率稍微偏高；一等舱的生还概率明显高于二等舱，二等舱的生还概率高于三等舱。

**层次聚类**

实验代码为：

# 层次聚类

cluster\_data <- select(train\_data, Survived, Age, Pclass)

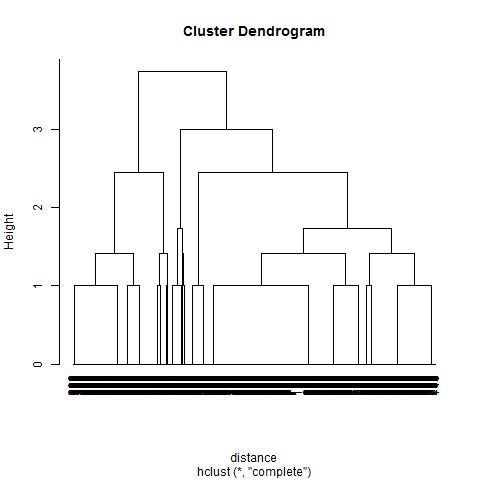
distance <- dist(cluster\_data)

cluster\_data.hc <- hclust(distance)

plot(cluster\_data.hc, hang = -1)

re <- rect.hclust(cluster\_data.hc, k = 2)

实验结果如下：



从图中可以看出，根据数据的分布，最终将数据聚成两类，分别是生存和未生成两种情况。

实验结论

本次实验，通过查阅资料，独立完成了实验要求。在本次实验中，灵活运用了前两次实验中的学到的内容，如数据的摘要分析，缺失数据处理以及数据可视化等。通过实验分析得知，泰坦尼克号上生存的乘客中，女性偏多，一等舱的乘客偏多。