Yiling Chen

Hw#4 Write-up

Apr. 23, 2021

**Question 2**

The goal for this problem is to investigate on the Titanic dataset on whether the famous titanic followed the women and children first policy base on the survival rate. I will answer the following questions: Is there evidence that women and children were the evacuated first? What characteristics are more likely in surviving passengers? What characteristics are more likely in passengers that perished? What is the probability that Rose would survive, and Jack would not survive?

First, I find the total survival rate for the Titanic tragic. 61% of the total passengers perished, and 39% of survival rate as a whole.

A picture containing text

Description automatically generated

Then, starting by exploring the dataset without considering passenger’s Pclass, and age group, I see that there are more male on board than female, yet the survival number of females is significantly exceeded males.

A picture containing text

Description automatically generated

Next, I use gg plot to look at the survival rate and perish rate base on people’s class and sex. From here we can see that the survival rate for someone locate in class 1 is around 60% but for someone locate in class 3 is a little less than 25%. We also can see that the survival rate for female is a little less than 75% but for male is less than 25%.

Chart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

In addition, I categorize the passengers on board by age groups such as people under 18 as child, people from 18-60 as adults, and people over 60 as elder. Not surprisingly, the survival rate of children is higher than adult and elder. However, the number of adult passengers is the largest population.

Chart, histogram

Description automatically generated Text

Description automatically generated with medium confidenceText

Description automatically generated with medium confidence

Next, we want to know what is the percentage of survival for Rose, a first class female adult is much more higher than Jack, a third class adult male.

Text, letter

Description automatically generated

Chart, bar chart

Description automatically generated

By combineing all the factors above, I use a mosaic plot to better illustrate the chances of survival base on different characteristics.

In conclusion, there is a clear evidence that women and children are evacuated first. Some characteristics of passengers that are likely to survive are: female, adult, locate in pclass 1. As well as female, adult and children in pcalss 2. Some characteristics of passengers that are likely to perish are: male adults in pclass 3 and female adults in pclass 3. The probability that Rose much higher than Jack’s survival rate.

**Question 3**

Specify the structure of a Bayesian Network that contains four nodes {W,X,Y,Z} and has satisfies the following set of independencies.

Diagram

Description automatically generated with medium confidence

**Question 4**

The goal for this problem is to create a user based collaborative filtering recommender system. The given dataset is MovieLense data in the recommenderlab package which contains about 100,000 ratings from 943 users on 1664 movies.

My first step is to look at the matrix rating and get a generalize picture for the dataset. From the diagram, we see the top 50 raters’s rating on the first 50 movies in a normalize form. Then in the code below, we can see the detail in numbers of each of the raters. Where the empty cells are the ones raters didn’t rate.

Chart, scatter chart

Description automatically generatedTable

Description automatically generated

Next, I use the recommender package functions to create a user based collaborative filtering recommender system.



Then, I can use the predict function to predict five recommendation of movies for some random users. All the movies were predicted using similar users who have seen the movie and gave that movie a great rating, hence our system believes these users are likely to like those movies.

Graphical user interface, text, application

Description automatically generated

Finally, I use the predict function again to predict the same three users rating for six movies as shown below. Some movies were not rated, instead showing NA because the user already rated the movie.

A picture containing graphical user interface

Description automatically generated

**Question 5**

The purpose for this problem is to test the performance of my recommender system previously built using cross validation method where k =5. The average errors are shown below:

Text

Description automatically generated

Next, I plot the ROC and the precision/recall curves:

Graphical user interface, text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

In conclusion, as the percentage of rated movies is recommended increase, the precision increase as well. The Mean Square Error of the recommender system I built using UBCF is 1.42.