Yiling Chen

Hw#2 Write-up

Mar. 2, 2021

**Question 1**

The goal for this problem is to cluster the marketing data using a classification tree. This dataset was a questionnaire containing 502 questions distributed to shopping mall customers in the San Francisco Bay area. This dataset contains 8993 observations and 14 variables. However, after deleting all the missing values observations in the dataset, it became 6876 observations and 14 variables.

Graphical user interface, text, application, chat or text message

Description automatically generatedText

Description automatically generated

Next, I generate a reference set with all the unique values from the marketing dataset and insert into a new data frame called ‘combo’. As we can see, combo has all 13752 observations and 14 variables. Then, before I could fit a classification tree with my random dataset, I created a binary response variable called ‘YY’.

Graphical user interface, text, application

Description automatically generated

Finally, I cluster the marketing data using a classification tree. Base on the classification tree, we can see that the householder status, language, household, home type and ethnic are major factors in determining the splits for our classification tree. Base on the terminal nodes we can see few different trends:

1. There is a high possibility that people live in house, condominium and apartment, has a household number less than five people, as well as the language they speak at home is less then 1.5, and the persons in household under 18 is less than 2.5.
2. There is a high possibility that people are pacific islander, white or other, their home type are apartment, mobile home or others, and they tend to have household less than 5.5 people as well as they spoke less than 1.5 languages at home with persons in household under 18 less than 2.5.
3. There is also a high possibility that People that more than 7 to 10 years in SF has a greater chance of having less than 4.5 people in the household under 18, as well as the type of home they lived in are condominium, apartment, mobile home, or others with the high possibility of having more than 2.5 persons in household that are under 18.

Text

Description automatically generatedDiagram

Description automatically generated

In conclusion, compare with the results near table 14.1, which are:

1. When number in household is 1 and no children, the language in home is English.
2. When language in home is English, and the householder status is own, as well as the occupation is ‘professional/managerial’, the income is greater than $40K.
3. When language at home is English, and income is less than $40K, and the marital status is not married as well as with no children, then their educations are not college graduate, or graduate study.

As we can see, the different method of prediction came up with different conclusions, but they are not against each other’s prediction. It also seems that the confidence for both methods is similar as well. Therefore, we could use both methods to come up with more associations.

**Question 2**

The goal for this problem is to use the Boston Housing dataset and find the rules of association and suggest different solutions for individuals that has their specific needs. The Boston Housing dataset contains 506 observations with 14 variables tracks Boston from the 1970 census. Text

Description automatically generated

Text

Description automatically generated

Next, I created histograms for all the variables in the data to visualize our Boston dataset. As shown below, each histogram shows the range of the variables on the x-axis as well as the frequency on the y-axis. For example, based on these histograms, we can see that even though the range of the crime rate is from 0 up to 80, but most of the areas has a crime rate less than 10. Similarly, on the age histogram, we can see that there are a large proportion of owner-occupied units built prior to 1940.

A diagram of a city

Description automatically generated with medium confidence

For the second part of question 2, in order to transform the dataset into binary incidence matrix, I first transform the Boston dataset into categorical data. I categorized each of the variable base on the range, and the median values. I changed all the variables into those categories because it is easy to distinguish as well as not too many categories can confuse us later if we deal with too many binary matrices.

Text, letter

Description automatically generated

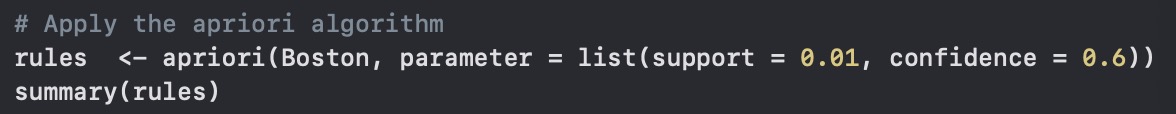
Now, all the Boston dataset has turned into categorical variables. Then, I converted the dataset into binary incidence matrices. Then, I visualize the data used the itemFrequencyPlot and set the support to 0.05 because I want to see a lot of the data but not the non-important ones. From the graph below, we can see that med-crim rate has been droped because it is not as significant as the low and high crimes.

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated



Next, I applied the apriori rule and I see that there are many rules that has been generated. Therefore, I set my parameter for a closer commute to the city as well as a low crime rate neighborhood base on the need of our student. The result came out as shown below. Base on the Apriori rule, the highest supported one is closer to the city and low tax will usually go together with low crim rate.

Table

Description automatically generated

Next, a family wanted to live somewhere where low pupil-teacher ratios, then base on our Apriori rule, when zn is high, the ptratio tend to be low, as well as the age will be low, and low-nox, and low-indus and low crime. It appears to be when the ptratio is low, it tends to be a good neighborhood.

Text

Description automatically generated

Now, we look at a regression model, and the level of significant of each variable to the ptratio. Base on the regression model, the zn, nox, rad, and medv are closely related when it comes to predicting ptratio. Association rule is more easy for interpretation when explaining to non-professionals, whereas regression can be more technical.

Table

Description automatically generated