CIS9660: Data Mining for Business Analytics

Exam 2

*Spring, 2023*

NAME (PRINT CLEARLY!!): Siddha Deshpande

Baruch ID (PRINT CLEARLY!!): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Instructions**

* This is a closed-book, closed-notes exam.
* There are **five** parts on the exam and a total of 100 points possible.
* The last three pages are blank. You can write on them if you need additional space.
* **Calculator Policy**: You can use a calculator that does not have the ability to communicate

with other electronic devices. (You are not allowed to use your smartphone’s calculator.)

Please write down the answers for Part 1 (Questions 1-20) in the following table. Choose only **one** answer for each question.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 **A** | 2  **C** | 3 **C** | 4 **A** | 5 **C** |
| 6 C | 7 A | 8 D | 9 B | 10 C |
| 11 C | 12 C | 13 D | 14 B | 15 B |
| 16 B | 17 E | 18 D | 19 C | 20 B |
|  |  |  |  |  |

Please write down the answers for Part 2, 3 and 4 (Questions 21-32) in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 21(2’) 0.20 | 22(2’) No Default | 23(2’) | 24(2’) | 25(2’) 0.1875 |
| 26(2’) | 27(2’) | 28(3’) | 29(3’) | 30(6’) |
| 31(2’) | 32(2’) |  |  |  |

Please write down the answers for Part 5 (Questions 33-35) below. No more than half page for each question.

Q33:

**Model: Association**

Data: I would use background data like interests, schooling, university, demographics, and social connections.

Analysis: Then, based on these factors, I would try to find out thumb rules or associations and then Generate association rules that capture user relationships based on their profiles, interests, and connections. I would also consider various parameters and thresholds to refine the rules, such as minimum support and confidence levels.

Then based on the generated association rules, develop an algorithm to suggest friends to users, and utilize the rules to recommend potential friends who have similar characteristics, interests, or mutual connections.

Q34:

**Model: Decision tree**

Data: To conduct the analysis, I would consider historical data related to previous iPhone models, including their specifications, features, release prices, and corresponding market responses. Additionally, I would also add relevant data on competitor models and their pricing strategies would be valuable and factors like technological advancements, market demand, customer preferences, and pricing trends.

Analysis:

First, I will construct the decision tree based on the selected features and their relationships with the pricing and analyze the decision tree to understand the most influential factors and their impact on pricing. Then I would try to apply the trained decision tree model to predict the most competitive price for the forthcoming iPhone model. It is important to consider the input features of the new model and follow the decision path in the decision tree to determine the predicted price. Then I can evaluate the performance of the decision tree model by comparing the predicted prices with the actual prices of previous iPhone models and refine the model, if necessary, by adjusting parameters, considering additional features, or using ensemble methods to improve accuracy.

Q35

**Model: Clustering**

Data: For this analysis, I would focus on customer data from the online dating platform, including customer profiles, preferences, and other relevant information.

Analysis: I would analyze the resulting clusters to understand individual characteristics and preferences and identify the key attributes to differentiate the clusters and find the best matches. Then I would develop an algorithm to utilize that generated cluster to suggest potential match for each user taking into account who belongs to the same cluster or have high similarity in their profiles and preferences.

**Part 1: Multiple Choice (2 points each; 40 points total)**

1. Which of the following is not true for decision tree analysis:
   1. A statistical method used to identify to which of a set of categories (or “subpopulations”) a new observation belongs
   2. If the attribute at a node is numerical, the number of children is usually the number of possible values of the attribute.
   3. On the basis of a training dataset containing observations of which categories were known
   4. A “divide-and-conquer” approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree.
2. Which of the following are not advantages of Decision Trees:
   1. Require relatively less effort from users for data preparation
   2. Automatically perform variable screening or feature selection
   3. Over Fitting is of less concern for decision tree models
   4. Non-linear relationships between the different parameters do not affect the performance of the tree
3. In Decision Trees, which of the following is not true about how to choose the best attribute for splitting?
   1. Each time, the algorithm examines each predictor variable and all possible split values for each variable to find the best split.
   2. We will need to use rationale to select features for splitting.
   3. Gini impurity shows which predictor variable produces the greatest “separation” in the outcome variable
   4. While our goal is to find partition with uniform category — pure leaf, attributes rarely split a group perfectly
4. Which of the following is not true about overfitting:
   1. Overfitting means the decision tree have poor performance for the training data
   2. Overfitting is the tendency to tailor models at the expense of generalization.
   3. As a model gets more complex, it is allowed to pick up harmful spurious correlations which do not represent characteristics of the population in general
   4. To avoid overfitting, we stop grow the tree before it gets too complex.
5. Which of the following is not true about clustering:
   1. Clustering or cluster analysis is the art of finding groups based on similarity in data
   2. Observations are not labeled, viz. don’t know which group one observation belongs to before the analysis
   3. The goal is to find a rule that can predict labels of new observations
   4. Unsupervised learning
6. K-Means is more suitable than other methods when \_\_\_\_\_.
   1. Clusters vary widely in size
   2. Clusters vary widely in density
   3. Clusters are in rounded shapes
   4. There are high variations in the value of variables
7. Which of the following is the right process for K-means clustering analysis:
   1. Specify the number of clusters, k
   2. k points are chosen at random as cluster centers.
   3. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric.
   4. Iteration continues until the same points are assigned to each cluster in consecutive rounds
   5. Repeated with the new cluster centers.
   6. Next the centroid, or mean, of the instances in each cluster is calculated—this is the “means” part. These centroids are taken to be new center values for their respective clusters.
   7. 123654
   8. 123456
   9. 123645
   10. 123465
8. Which of the following is least likely to be true about how we choose the number of clusters (K):
   1. Choose based on external reasons
   2. Try out different possibilities and see which is best
   3. Visually select k to make sure the clusters are describing distinct groups
   4. Using results of other analysis such as decision trees
9. Which of the following is not true about normalization?
   1. The average after normalization is 0
   2. The value of each attribute after normalization is always less than 1
   3. Normalizing variables is likely to hide the true groupings present in the data
   4. Normalizing the variables is important to keep a variable with high variability from dominating the cluster analysis
10. As a general rule, if we increase the number of clusters, cohesion between clusters will\_\_\_\_ and separation between clusters will \_\_\_\_\_.
    1. Decrease, Decrease
    2. Decrease, Increase
    3. Increase, Decrease
    4. Increase, Increase
11. Which of the following is not true about limitations of K-Means clustering?
    1. The result is locally optimal
    2. It is almost infeasible to find globally optimal clusters
    3. The final clusters are not sensitive to the initial cluster centroids
    4. K-Means gives unreliable results in many cases
12. Compared to K-Means, hierarchical clustering analysis is less suitable when:
    1. You don’t know an appropriate value for k
    2. You are more interested in how each observation is collected to each other
    3. The number of observations is large
    4. Shapes of clusters are asymmetry
13. Which of the following statements is not true about association rule mining?
    1. Attempt to find interesting associations between entities based on transactions involving them
    2. Not a good algorithm for numeric data
    3. Unsupervised machine learning
    4. The challenge is that it is hard to find all possible combinations of items
    5. c and d
14. Which of the following statements is not true about the support of a rule:
    1. Support is an indication of how frequently the itemset appears in the dataset
    2. The support is sometimes expressed as a percentage of the total number of A and B together in the database
    3. The support measures the coverage of the association rule
    4. A constraint of maximum support is often applied in order to control complexity
15. Which of the following statements is true about the confidence of a rule, X→Y:
16. How often baskets contain both X and Y
17. How often Y appears in baskets that contain X
18. Whether X and Y appear together at the same frequency as random chance
19. When confidence> 1, the occurrence of X → Y together is more likely than what you would expect by chance
20. Which of the following statements is not true:
21. Confidence(X🡪Z) =Support(X🡪Z)/Support(X)
22. Lift(X🡪Z) =Confidence(X🡪Z)/Support(Z)
23. Lift(X🡪Z) = Support(X🡪Z)/(Support(X) Support(Z))
24. b and c
25. None
26. Which of the following is limitation of association rule mining?
27. Some important information such as price and quantity not considered.
28. Analytics insights aren't always actionable
29. Random data can generate apparently interesting association rules
30. The more rules you produce, the greater the danger is
31. All above
32. Which of the following is the least likely to be answered using decision tree?
33. Detect watermelon ripeness
34. Detect competitors
35. Detect future tend for housing sales
36. Suggest friends on Facebook

**Interpreting Association Rule Mining**

The following is an excerpt from the association rule analysis of a local grocery store on Temple’s Main Campus:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Rule** | **Support** | **Confidence** | **Lift** |
| 1 | {Green Tea, Plastic Bags} => {Paper Towels} | 0.0167 | 0.41 | 2.80 |
| 2 | {Paper Towels, Green Tea } => {Plastic Bags} | 0.0167 | 0.46 | 2.68 |
| 3 | {Hot Dogs} => {Lucky O’s} | 0.0218 | 0.24 | 3.60 |
| 4 | {M&Ms, Paper Towels} => {Plastic Bags} | 0.0132 | 0.41 | 2.40 |
| 5 | {M&Ms, Plastic Bags} => {Paper Towels} | 0.0132 | 0.33 | 2.25 |
| 6 | {M&Ms, Green Tea} => {Plastic Bags} | 0.0137 | 0.43 | 2.53 |
| 7 | {Pop-Tarts} => {Organic Coffee} | 0.0372 | 0.49 | 0.79 |

1. Based on the output, which of the following itemsets appears most often in shopping baskets?
2. Green Tea, Plastic Bags and Paper Towels
3. Hot Dogs and Lucky O’s
4. Pop-Tarts and Organic Coffee
5. M&Ms, Plastic Bags and Paper Towels
6. Sue has M&Ms and Green Tea in her cart. You then see her take out the M&Ms and replace it with Paper Towels. Which of the following is true?
7. The likelihood of Sue buying Plastic Bags is unchanged
8. It is now more likely than before that Sue will buy Plastic Bags
9. It is now less likely than before that Sue will buy Plastic Bags
10. It is likely that Sue will switch back to M&Ms

**Part 2: Interpreting Decision Tree Output (10 points total)**

The following is a decision tree used by the president of the bank who wants to predict how likely a future customer is to pay back their loan so she can make better loan approval decisions.

A picture containing diagram

Description automatically generated

**Answer the following questions regarding this tree:**

1. How likely is a customer to pay back their loan if they are married but have no children? (2 points)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *(write the answer in the blank)*

1. *What does “0” in the circle mean? (2 points) (write your answer in the blank)*

*The 0 in the circle means no default, as in the condition above is not true or no default or 0.*

1. Children>=3 and <3 is list as the first splitter. What does this mean? Does this make sense to you? Explain your reasons. (2 points)

The first splitter is the initial condition based on which the whole tree is divided. The analysis aims to split the whole data based on whether a customer is more or less likely to repay the loan based on the number of children he/she has. More children mean more expenses and fewer savings, which is a huge factor in loan payback.

1. According to the tree, will marriage increase or decrease people’s propensity to pay back their own? Does this make sense to you? Explain your reason *(2 points)*

According to the tree, marriage decreases people’s propensity to pay back. Yes, since anything that increases the expenses can threaten a customer's ability to pay back.

1. Compute the correct classification rate based on the following confusion matrix:

Table, calendar

Description automatically generated

What is the correct classification rate for this decision tree: \_\_\_\_\_  
*(write the number, 2 points)*

**Part 3: Interpreting Clustering Output (10 points total)**

Consider the output from a cluster analysis of Census Data when **six clusters** are specified:

> # Display the cluster sizes

> cat("\nCluster s ..." ... [TRUNCATED]

Cluster size:

> MyKMeans$size

[1] 4963 7156 4242 7965 4492 2074

> # Display the cluster means (means for each input variable)

> cat("\nCluster Means (centroids):")

Cluster Means (centroids):

> MyKMeans$centers

RegionDensityPercentile MedianHouseholdIncome AverageHouseholdSize

1 1.13503301 -0.2236964 -0.77154666

2 0.85819972 1.4012665 0.32650016

3 -1.14288405 -0.5566597 -0.55985291

4 -0.98068538 -0.2391447 0.66813667

5 0.01195197 -0.1603545 -0.04070976

6 1.02091360 -0.3220717 1.31884198

> # Display withinss (i.e. the within-cluster SSE for each cluster)

> cat("\nWithin cluster SSE for each cluster (Cohesion):")

> MyKMeans$withinss

[1] 4577.141 4598.839 4187.275 3366.116 3860.349 2531.689

> # Display betweenss

> cat("\nTotal between-cluster SSE (Seperation):")

Total between-cluster SSE:

> MyKMeans$betweenss

[1] 48276.65

> cat("\nAverage between-cluster SSE:")

Average between-cluster SSE:

> MyKMeans$betweenss/NUM\_CLUSTER

[1] 8046.108

Some terminology:

* Region density percentile: Population density where households in that segment reside
* Median household income: Median household income for households in that segment
* Average household size: Average number of people in the households in that segment

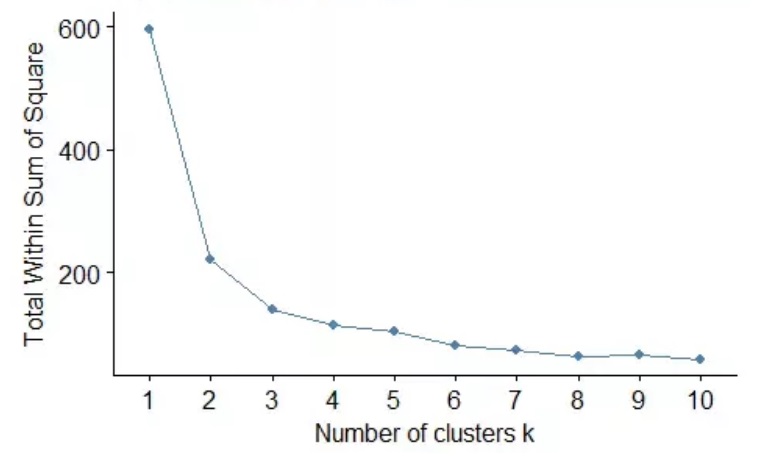
**Answer the questions on the next page regarding the cluster analysis.**

Which number should we look at if we want to compare the separation of this set of clusters with another one? \_\_\_\_\_\_\_\_ Total between cluster SSE\_\_\_\_\_\_\_\_\_\_\_\_\_  
*(2 points)*

1. Which number should we look at if we want to compare the cohesion of this set of clusters with another one? \_\_\_\_\_\_Within cluster SSE\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
   *(2 points)*
2. Describe the characteristics of group 1 compared to the mean of the population: (3 points)

Group 1 has a higher population density (1.13503301), lower median household income (-0.2236964), and lower average household size (-0.77154666).

1. Assume that you have a following graph from the data. What number of clusters would you suggest to improve this clustering analysis? (Write one number and explain why. Credits are given based on logic and reasoning) (3 points)

****

I would choose 4 or 5 because selecting the number of clusters at or near the elbow point as the optimal value is generally recommended. This allows us to capture meaningful patterns in the data and to maintain a reasonable number of clusters.

If not, we may face overfitting and unnecessarily complex cluster structures, resulting in smaller, more fragmented clusters that need to capture meaningful patterns in the data. This can make interpretation and analysis more challenging and may not provide substantial benefits in terms of understanding the underlying data distribution.

**Part 4: Computing Support, Confidence, and Lift (10 points total)**

Consider the following set of baskets for a new burrito restaurant on Temple’s Main Campus. Each basket represents a customer order:

|  |  |
| --- | --- |
| Basket | Items |
| 1 | White Rice, Black Beans, Chicken, Mild Salsa, Soda |
| 2 | White Rice, Steak, Corn Salsa, Soda |
| 3 | Brown Rice, Tofu, Hot Salsa, Lettuce, Soda |
| 4 | White Rice, Black Beans, Chicken, Corn Salsa |

1. **Compute the support, confidence, and lift for the following rules**

Write the values in the table below. Keep 3 digits after the decimal point**.** (6 points)

|  |  |  |  |
| --- | --- | --- | --- |
| **Rule** | **Support** | **Confidence** | **Lift** |
| {White Rice} => {Soda} | **2/4=0.50** | **2/3=0.66** | **0.5/ 0.5625=0.889** |

The burrito restaurant also wants to determine how its new CrazyQueso is affecting sales of its SuperSalsa. They have the following customer data for 10,000 customers:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Bought CrazyQueso** | | |  |
| **Bought SuperSalsa** |  | **No** | **Yes** |  |
| **No** | 3,500 | 1,000 |  |
| **Yes** | 1,000 | 4,500 | Total: 10,000 |

1. Compute the lift value for the rule: { CrazyQueso } => { SuperSalsa } \_\_\_0.45/0.55\*0.55\_\_\_1.488\_\_\_\_\_\_\_

*(Write the number. Keep 3 digits after the decimal point. 2 points)*

People who bought both items= 4500/10,000=0.45

People who bought supersalsa= 5500/10000=0.55

People who bought CrazyQueso= 5500/10000=0.55

1. Explain the lift value for the rule: { CrazyQueso } => { SuperSalsa }. (Use less than 2 sentences, 2 points)

The rule’s lift value indicates a positive relationship between CrazyQueso and super salsa. It says that CrazyQueso customers are 1.488 times more likely to buy SuperSalsa.

**Part 5: Choose the right models (10 points each; 30 points total)**

Transform each of the business problems into a data mining task by:

* Choose the right model from ***decision tree, clustering, or association rule mining*** for each analysis (3’)
* Roughly explain what data you will use and how you plan to conduct the analysis (7’)

1. Create an algorithm to suggest friends on Facebook.
2. Determine the most competitive price for a forthcoming model of iPhone
3. Find the best match for each customer on an online dating platform.

[BLANK PAGE]

*You can write here if you need additional space or as scratch papers.*

[BLANK PAGE]

*You can write here if you need additional space or as scratch papers.*

[BLANK PAGE]

*You can write here if you need additional space or as scratch papers.*