CIS9660: Data Mining for Business Analytics

Exam 2

*May 20, 2021*

NAME (PRINT CLEARLY!!): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

**Instructions**

* This is a closed-book, closed-notes exam.
* There are **four** 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.)

(This page is for grading purpose.)

|  |  |  |
| --- | --- | --- |
| **Part** | **Points Possible** | **Points Assigned** |
| Part 1 | 48 |  |
| Part 2 | 10 |  |
| Part 3 | 8 |  |
| Part 4 | 7 |  |
| Part 5 | 27 |  |
| **Total** | **100** |  |

**Part 1: Multiple Choice (3 points each; 48 points total)**

Please write down the answers for Part 1 (Questions 1-22) in the following table. Choose only **one** answer for each question.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 |
| 16 |  |  |  |  |
|  |  |  |  |  |

1. ~~Which of the following is~~ **~~not~~** ~~an aspect of the principle that a graphic should “tell a story”?~~
   1. ~~A graphic should be clear on its own~~
   2. ~~A graphic should enable meaningful comparison~~
   3. ~~A graphic should provide insight beyond the text~~
   4. ~~A graphic should use color in a visually appealing way~~
2. ~~The following is an example of “chartjunk”:~~
   1. ~~Moire effects~~
   2. ~~Three-dimensional bar charts~~
   3. ~~Chart legends~~
   4. ~~Moire effects and three dimensional bar charts~~
3. Which of the following is true for decision tree analysis:
   1. We use decision tree to determine the probability of an event happening based on predictor variable values
   2. There could be only one tree that fits the same data
   3. Changing the order of predictors will lead to different results
   4. Adding more nodes in the tree ensures a smaller model error
4. Increasing the complexity factor parameter in our decision tree analysis can potentially:
   1. Increase the classification accuracy rate of the tree
   2. Decrease the number of nodes in the tree
   3. Decrease the size of the training set
   4. Leads to a more complex tree
5. Which of the following is not the advantage of decision tree analysis:
   1. Require relatively less effort for data preparation
   2. Automatically perform variable screening or feature selection
   3. Handle category variables well
   4. Best models to provide information on the non-linear relationship between the predictors and the response.
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 not true for K-means clustering analysis:
   1. Prior knowledge of the classes (categories) is unknown
   2. Subjective expectations of the analyst are not useful to figure out if the clusters are good or not
   3. K-means clustering analysis is often used in marketing for market segmentation
   4. Normalization is always needed before running K-means clustering analysis
8. Which of the following is not true about overfitting:
   1. Overfitting is about including too many explanatory variables
   2. Overfitting is about including irrelevant explanatory variables
   3. Overfitting means the decision tree have poor predictive performance for new data
   4. All the other three statements are true
9. Which of the following is not true about normalization?
   1. The average after normalization is 0
   2. The standard deviation after normalization is 1
   3. Help correct outliers
   4. Normalizing variables is likely to hide the true groupings present in the data
10. As a general rule, if we increase the number of clusters, cohesion between clusters will\_\_\_\_ and separation between clusters will \_\_\_\_\_.
    1. Increase, Decrease
    2. Increase, Increase
    3. Decrease, Decrease
    4. Decrease, Increase
11. 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. Then number of observations is large
    4. Shapes of clusters are asymmetry
12. Which of the following statements is true about the confidence of a rule, X→Y:
13. How often baskets contain both X and Y
14. How often Y appears in baskets that contain X
15. Whether X and Y appear together at the same frequency as random chance
16. When confidence> 1, the occurrence of X → Y together is more likely than what you would expect by chance
17. Which of the following statements is not true about support:
18. support(X🡪Z) can be the same as support(Z🡪X)
19. support(X🡪Z) can be the same as support(Z🡪X,Y)
20. support(X🡪Z) must be different from support(Z🡪X,Y)
21. support(X🡪Z) cannot be negative

**Interpreting Association Rule Mining**

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

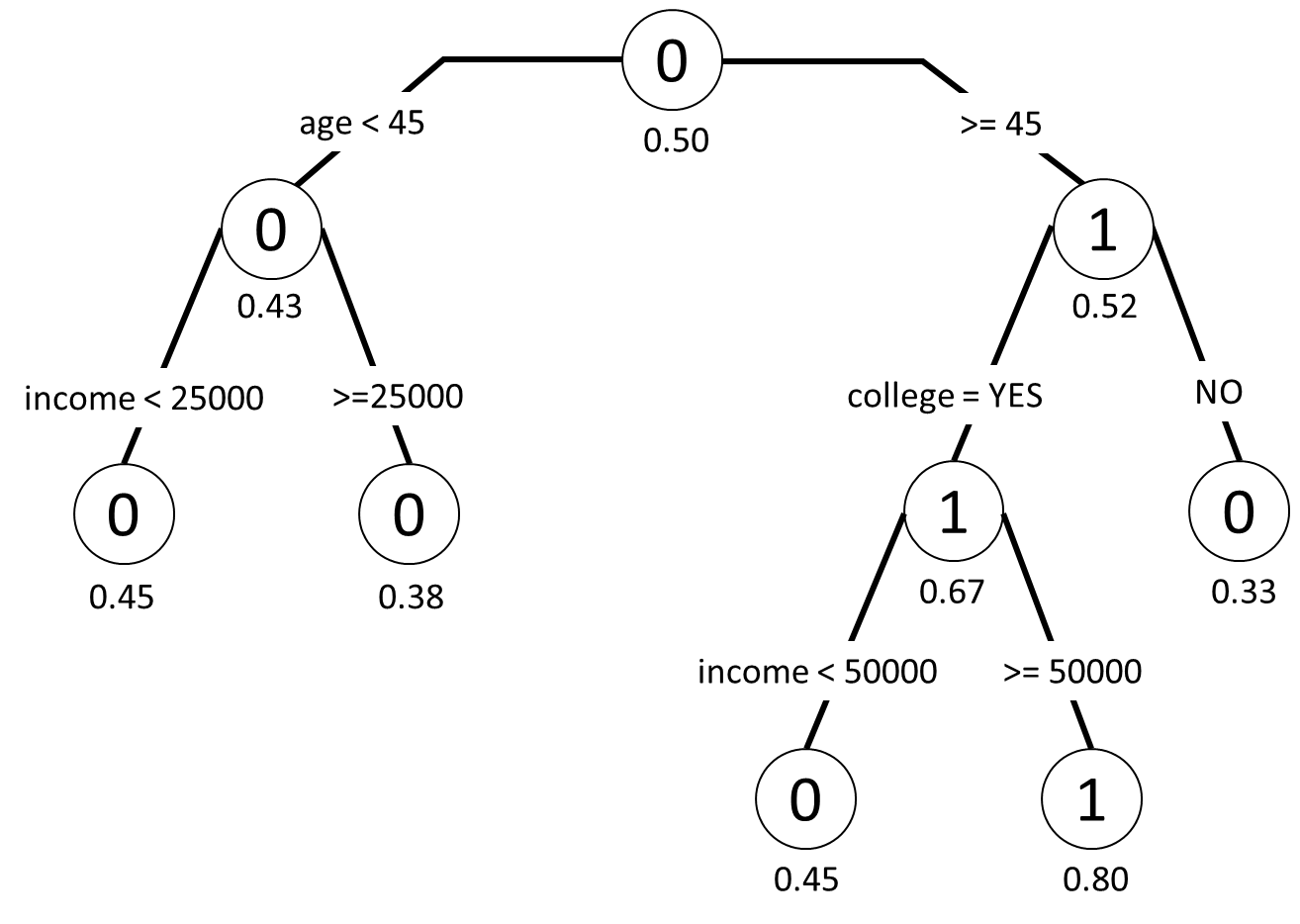
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Rule** | **Support** | **Confidence** | **Lift** |
| 1 | {Green Tea, Plastic Bags} => {Paper Towels} | 0.0167 | 0.41 | 2.80 |
| 2 | {Paper Towels, Green Tea } => {Plastic Bags} | 0.0167 | 0.46 | 2.68 |
| 3 | {Hot Dogs} => {Lucky O’s} | 0.0218 | 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
11. Which of the following statements is **not true** about implication of association rules:
12. The more rules you produce, the more applicable the result is
13. Some important information not considered
14. Random data can generate apparently interesting association rules
15. Association may not imply causality

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

The following is a decision tree used by a major car dealership to determine whether a customer will buy a new car this year.

* “age” is the age of the customer in years
* “income” is the yearly income of the customer in dollars
* “college” is whether the customer graduated from college



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

1. What is the probability that a 50 year-old without a college degree who makes $55,000 per year will buy a car this year? (1 points)

\_\_\_\_\_\_\_\_\_0.33\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *(write the answer in the blank)*

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

1. College is a split variable in for people 45 years old and older, but not for people under 45 years old. What does this imply? *(2 points)* 
   1. Everyone under 45 went to college
   2. For people under 45, attending college doesn’t determine whether they will buy the car
   3. Most people over 45 didn’t go to college while most people under 45 did go to college
   4. Income for people under 45 is lower than income for people over 45
2. For people over 45, College is listed ***on top of*** income. What does this mean? Does this make sense to you? Explain your reasons. (3 points)
3. Compute the correct classification rate based on the following confusion matrix:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Predicted outcome:** | | |  |
|  |  | Sunny | Windy | Rainy |  |
| **Observed outcome:** | Sunny | 301 | 207 | 199 |  |
| Windy | 120 | 482 | 20 |  |
| Rainy | 112 | 158 | 401 | Total: 2000 |

What is the correct classification rate for this decision tree: \_\_\_\_\_0.592\_\_\_\_\_\_\_  
*(write the number, 2 points)*

**Part 3: Interpreting Clustering Output (8 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):")

Within cluster SSE for each cluster (Cohesion):

> MyKMeans$withinss

[1] 4577.141 4598.839 4187.275 3366.116 3860.349 2531.689

> # Display betweenss (i.e. the SSE between clusters)

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

Total between-cluster SSE (Seperation):

> MyKMeans$betweenss

[1] 48276.65

> # Compute average separation: more clusters = less separation

> 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.**

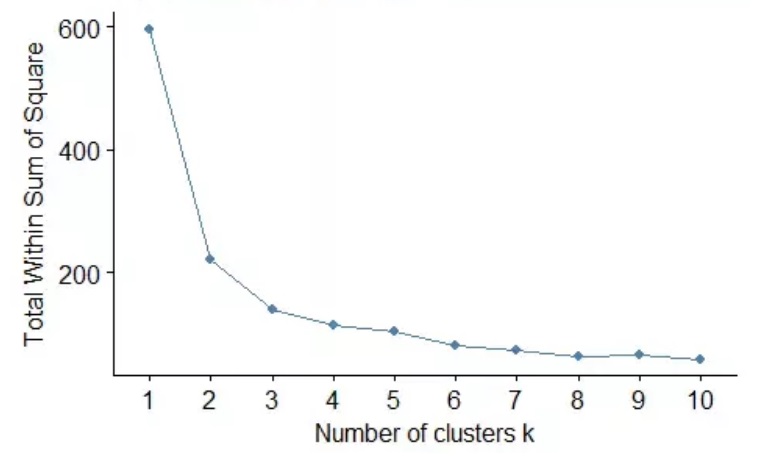
1. Which number should we look at if we want to compare the separation of this set of clusters with another one? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
   *(2 points)*
2. Compare the characteristics (RegionDensityPercentile, MedianHouseholdIncome, and AverageHouseholdSize) of Cluster 6 to the population as a whole: (3 points)

RegionDensityPercentile: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ than population average

*(write "higher” or “lower” in the blank)*   
  
MedianHouseholdIncom: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ than population average  
*(write "higher” or “lower” in the blank)*   
  
AverageHouseholdSize: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ than population average

*(write "higher” or “lower” in the blank)*

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. As long as your reasons make sense to me, you will get the points) (3 points)

****

**Part 4: Computing Support, Confidence, and Lift (7 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**.** (3 points)

|  |  |  |  |
| --- | --- | --- | --- |
| **Rule** | **Support** | **Confidence** | **Lift** |
| {White Rice} => {Soda} | **0.5** | **0.66** | **0.89** |

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 } \_\_\_\_\_\_\_1.488\_\_\_\_\_\_

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

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

**Part 5: Choose the right models (9 points each; 27 points total)**

* Transform each of the business problem into a data mining task by explaining what is the input and what is the output of each analysis. (3’)
* Choose the right model from ***decision tree, clustering, or association rule mining*** for each analysis (2’)
* Roughly explain what data you will use and how you plan to conduct the analysis (4’)

1. Assume you are the state governor of NY. You need to decide when to lift the lockdown. What type of analysis might be helpful for you to make the right decision?
2. Assume you are a marketing analyst of Amazon. Now you want to promote a sofa from a seller “BestSofa” by putting the product in the recommendation page of potential customers. What type of analysis you can do to find the potential customers for this sofa?
3. Assume you are a TV show director. Now you want to choose one among several books and turn it into a TV show. What type of analysis you can do to find the one that is most likely to be loved by customers?

[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.*