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"?~~

~~a. A graphic should be clear on its own~~
~~b. A graphic should enable meaningful comparison~~
~~c. A graphic should provide insight beyond the text~~
~~d. A graphic should use color in a visually appealing way~~

~~2. The following is an example of "chartjunk":~~

~~a. Moire effects~~
~~b. Three-dimensional bar charts~~
~~c. Chart legends~~
~~d. Moire effects and three dimensional bar charts~~

3. Which of the following is true for decision tree analysis:
   a. We use decision tree to determine the probability of an event happening based on predictor variable values
   b. There could be only one tree that fits the same data
   c. Changing the order of predictors will lead to different results
   d. Adding more nodes in the tree ensures a smaller model error

4. Increasing the complexity factor parameter in our decision tree analysis can potentially:
   a. Increase the classification accuracy rate of the tree
   b. Decrease the number of nodes in the tree
   c. Decrease the size of the training set
   d. Leads to a more complex tree

5. Which of the following is not the advantage of decision tree analysis:
    a. Require relatively less effort for data preparation
    b. Automatically perform variable screening or feature selection
    c. Handle category variables well
    d. 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 _____.
    a. Clusters vary widely in size
    b. Clusters vary widely in density
    c. Clusters are in rounded shapes
    d. There are high variations in the value of variables

7. Which of the following is not true for K-means clustering analysis:
    a. Prior knowledge of the classes (categories) is unknown
    b. Subjective expectations of the analyst are not useful to figure out if the clusters are good or not
    c. K-means clustering analysis is often used in marketing for market segmentation
    d. Normalization is always needed before running K-means clustering analysis

8. Which of the following is not true about overfitting:
    a. Overfitting is about including too many explanatory variables
    b. Overfitting is about including irrelevant explanatory variables
    c. Overfitting means the decision tree have poor predictive performance for new data
    d. All the other three statements are true

9. Which of the following is not true about normalization?
    a. The average after normalization is 0
    b. The standard deviation after normalization is 1
    c. Help correct outliers
    d. 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 _____.
    a. Increase, Decrease
    b. Increase, Increase
    c. Decrease, Decrease
    d. Decrease, Increase

11. Compared to K-Means, hierarchical clustering analysis is less suitable when:
    a. You don't know an appropriate value for k
    b. You are more interested in how each observation is collected to each other
    c. Then number of observations is large
    d. Shapes of clusters are asymmetry

12. Which of the following statements is true about the confidence of a rule, X→Y:
    a. How often baskets contain both X and Y
    b. How often Y appears in baskets that contain X
    c. Whether X and Y appear together at the same frequency as random chance
    d. When confidence> 1, the occurrence of X → Y together is more likely than what you would expect by chance

13. Which of the following statements is not true about support:
    a. support(X→Z) can be the same as support(Z→X)
    b. support(X→Z) can be the same as support(Z→X,Y)
    c. support(X→Z) must be different from support(Z→X,Y)
    d. 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

14. Based on the output, which of the following itemsets appears most often in shopping baskets?
    a. Green Tea, Plastic Bags and Paper Towels
    b. Hot Dogs and Lucky O's

c. Pop-Tarts and Organic Coffee

d. M&Ms, Plastic Bags and Paper Towels

15. 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?

a. The likelihood of Sue buying Plastic Bags is unchanged

b. It is now more likely than before that Sue will buy Plastic Bags

c. It is now less likely than before that Sue will buy Plastic Bags

d. It is likely that Sue will switch back to M&Ms

16. Which of the following statements is **not true** about implication of association rules:

a. The more rules you produce, the more applicable the result is

b. Some important information not considered

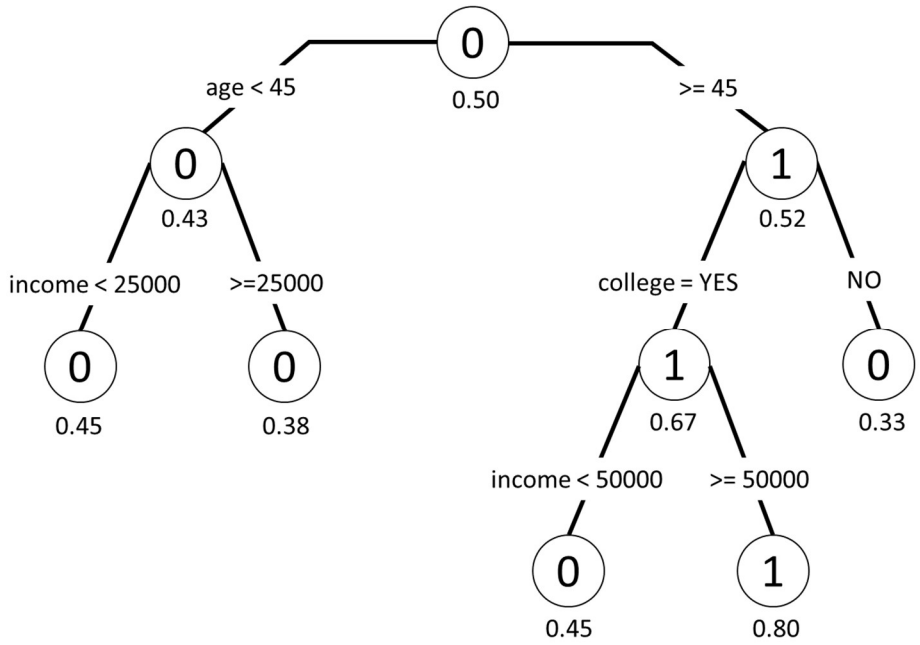
c. Random data can generate apparently interesting association rules

d. Association may not imply causality

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

The following is a decision tree a major car dealership uses 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:

17. 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)

18. *What does "0" in the circle mean? (2 points) (write your answer in the blank)*

19. 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)*
    a. Everyone under 45 went to college
    b. For people under 45, attending college doesn't determine whether they will buy the car
    c. Most people over 45 didn't go to college while most people under 45 did go to college
    d. Income for people under 45 is lower than income for people over 45

20. 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)

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

Predicted outcome:

Observed outcome:	Sunny	Windy	Rainy	
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.

22. Which number should we look at if we want to compare the separation of this set of clusters with another one? _____Sum of squares error_____
(2 points)

23. Compare the characteristics (RegionDensityPercentile, MedianHouseholdIncome, and AverageHouseholdSize) of Cluster 6 to the population as a whole: (3 points)

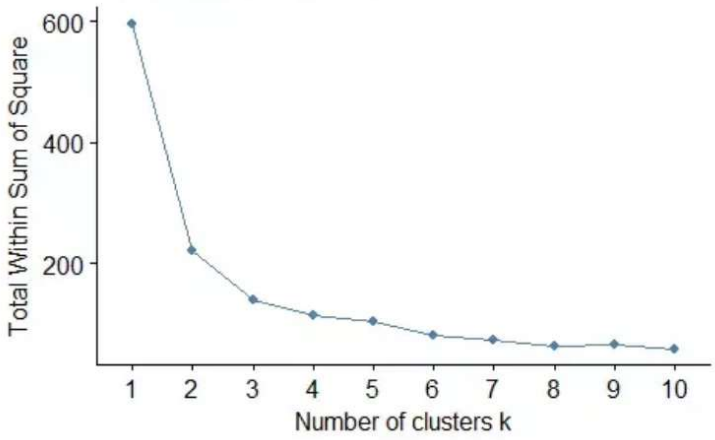
RegionDensityPercentile: _____ _____ than population average

(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)

24. Assume that you have the 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)



Choosing the number of clusters beyond the elbow point can lead to overfitting and unnecessarily complex cluster structures. It may result in smaller, more fragmented clusters that do not 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.

Therefore, it is generally recommended to select the number of clusters at or near the elbow point as the optimal value. This allows for a reasonable trade-off between capturing meaningful patterns in the data and maintaining a manageable number of clusters.

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

25. **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**

$$c(X \to Y) = \frac{s(X \to Y)}{s(X)}$$

$$= \frac{s(\text{Milk, Diapers, Beer})}{s(\text{Milk, Diapers})} = \frac{0.4}{0.6} = 0.67$$

$$Lift(A \to B) = \frac{s(A \to B)}{S(A) * S(B)} = \frac{\frac{s(A \to B)}{S(A)}}{S(B)} = \frac{c(A \to B)}{S(B)}$$

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	
		No	**Yes**
Bought SuperSalsa	**No**	3,500	1,000
	Yes	1,000	4,500

Total: 10,000

26. Compute the lift value for the rule: { CrazyQueso } => { SuperSalsa } _____1.488_____

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

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

The lift value of 1.488 for the rule {CrazyQueso} => {SuperSalsa} indicates that there is a positive association between buying CrazyQueso and buying SuperSalsa. It suggests that customers who buy CrazyQueso are 1.488 times more likely to buy SuperSalsa compared to the overall probability of buying SuperSalsa.

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')

28. 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?

Analysis: Decision Tree

Data Mining Task: Decision-making for lifting the lockdown in NY

Input:

Historical data on COVID-19 cases, hospitalization rates, and vaccination rates in NY.

Socio-demographic data such as population density, age distribution, and healthcare resources.

Economic data such as unemployment rates, business closures, and revenue loss.

Output:

Decision on when to lift the lockdown in NY based on various factors and criteria.

Data and Analysis: To conduct the analysis, you would gather relevant data from sources such as the New York State Department of Health, Census Bureau, and economic indicators. The data would include daily or weekly records of COVID-19 cases, hospitalizations, and vaccinations. Additionally, you would collect socio-demographic data at the county or regional level, including population density, age distribution, and healthcare resources. Economic data such as unemployment rates, business closures, and revenue loss would also be necessary.

The decision tree model would suit this analysis as it can handle numeric and categorical variables. The model could be trained on the historical data, with the target variable being whether to lift the lockdown. The features or attributes used in the decision tree could include COVID-19 indicators (cases, hospitalizations, vaccinations), socio-demographic factors, and economic factors. The model would provide insights into the key factors contributing to the decision, such as the thresholds for COVID-19 cases or vaccination rates indicating a safer reopening. It would help the state government make an informed decision by considering the various factors and their interplay.

29. 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?

Analysis: Association Rule Mining

Data Mining Task: Identifying potential customers for the BestSofa product

Input:

Customer purchase history data from Amazon, including information on products purchased, customer demographics, and browsing behavior.

Product data for the BestSofa, including its features, price, and customer reviews.

Customer feedback and ratings on similar sofas or furniture products.

Output:

Association rules that indicate which customers are likely to be interested in purchasing the BestSofa.

A list of potential customers who can be targeted for promotion on the recommendation page.

Data and Analysis: To find potential customers for the BestSofa, association rule mining can be utilized. The analysis would involve the following steps:

Data Preparation:

Gather customer purchase history data from Amazon, including details on products purchased, customer demographics, and browsing behavior.

Collect product data for the BestSofa, including its features, price, and customer reviews.

Incorporate customer feedback and ratings on similar sofas or furniture products.

Data Preprocessing:

Clean and preprocess the data, removing any irrelevant or duplicate entries.

Transform the data into a suitable format for association rule mining, typically using transactional data format.

Association Rule Mining:

Apply association rule mining algorithms, such as the Apriori algorithm, to discover patterns and associations between customers' purchasing behavior and the BestSofa.

Set appropriate thresholds for support and confidence to filter out meaningful and actionable rules.

Rule Evaluation and Interpretation:

Evaluate the generated association rules based on their support and confidence values.

Identify the rules that indicate a strong association between the purchase of certain products or characteristics and the likelihood of purchasing the BestSofa.

Interpret the rules to understand the characteristics, preferences, or purchase patterns of potential customers for the BestSofa.

Targeted Promotion:

Based on the identified association rules and potential customer segments, promote the BestSofa on the recommendation page for customers who exhibit similar characteristics or purchasing behavior.

Monitor the effectiveness of the promotion and iterate on the analysis to optimize targeting and increase conversion rates.

By leveraging association rule mining, Amazon can identify potential customers who are more likely to be interested in purchasing the BestSofa based on their past behavior and preferences. This targeted promotion strategy increases the chances of reaching relevant customers and driving sales for the specific product.

30. 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?

Analysis: Clustering Analysis

Data Mining Task: Selecting a book for a TV show that is likely to be loved by customers

Input:

Data on a collection of books, including their genres, authors, publication dates, ratings, reviews, and popularity.

Customer preferences and viewing history data, such as genres they enjoy, favorite TV shows, and ratings/reviews they have provided.

Demographic data of the target audience, such as age, gender, and location.

Output:

Clusters of books that share similar characteristics and are likely to be loved by similar groups of customers.

Identification of the book cluster that aligns well with the target audience's preferences.

Data and Analysis: To choose a book that is most likely to be loved by customers and suitable for adaptation into a TV show, a clustering analysis can be performed. The analysis would involve the following steps:

Data Collection:

Gather data on a collection of books, including their genres, authors, publication dates, ratings, reviews, and popularity.

Collect customer preferences and viewing history data, such as genres they enjoy, favorite TV shows, and ratings/reviews they have provided.

Obtain demographic data of the target audience, such as age, gender, and location.

Data Preprocessing:

Clean and preprocess the book and customer data, ensuring data quality and consistency.

Select relevant features from the book and customer data, such as genres, ratings, and demographics.

Feature Engineering:

Extract or derive additional features that may be relevant for capturing customer preferences, such as sentiment analysis of book reviews or popularity metrics.

Clustering Analysis:

Apply clustering algorithms, such as k-means or hierarchical clustering, to group books based on their similarities in terms of genres, ratings, popularity, and other relevant features.

Consider incorporating customer preferences and demographics as additional inputs to the clustering analysis to tailor the results to the target audience.

Cluster Evaluation and Interpretation:

Evaluate the quality and coherence of the generated clusters, considering metrics such as silhouette score or cluster compactness.

Interpret the clusters to understand the characteristics and preferences of each group.

Selection of Book:

Analyze the clusters and identify the cluster(s) that align well with the preferences of the target audience and have the highest potential for customer love and engagement.

Choose the book(s) within the selected cluster(s) that are most suitable for adaptation into a TV show based on factors like plot, character development, and visual appeal.

By performing a clustering analysis on the book and customer data, the TV show director can identify the book that is most likely to be loved by customers. This analysis considers customer preferences, demographics, and book characteristics to align the TV show with the target audience's preferences and maximize its potential for success.

[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.