

Assignment 7: Clustering Analysis

Due: November 12, Thursday, 11:59pm on Canvas

Total Points: 100 points

Submission Requirements: Please submit a *single* R-markdown file and the generated pdf file to Canvas. The pdf file generated (knitted) by R-markdown file should **display all your R code and answers to the questions**. Please set: `echo = TRUE`, `warning = FALSE`, `error = TRUE`, `message = FALSE`. Make sure you clearly mark the problem number. You can use the R-markdown template posed on Canvas in the logistics module. In the R-module on canvas, there is also an R-markdown cheat sheet.

Name your R-markdown file and the generated pdf file as: **K353_lastname_HW1**

Grading Criteria: I will randomly run your R-markdown file on my computer, and they should be executed successfully. If the execution of the R-markdown file is successful, then I will calculate your points for the assignment by comparing the generated R objects to the correct R objects. If the execution of the R-markdown file fails (i.e., an error message pops up during the execution of the R-markdown file), I will let you know and give you 24 hours to resubmit your assignment. After 24 hours, if I still had not received an updated assignment, the assignment receives zero points. So please make sure you submit the pdf generated by the R-markdown file you submitted.

Make sure you:

1. Explain your results. Don't just show your plots or your numerical statistics
2. If the question does not require any R coding, make sure you show the steps. So for example, some questions may ask you to calculate something, make sure you show the equations etc.
3. You can write your answers for the non-coding questions either by hand or outside your R code chunk in Rmarkdown.

Problem 1 [20 points]

Use the built-in dataset **USArrests** to perform hierarchical clustering on the states. This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

1. [5 points] Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
2. [5 points] Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
3. [5 points] Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.
4. [5 points] What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

Problem 2 [25 points]

Use the Customers data set posted on canvas. Name the data set **cust_data**

1. [3 points] Generate the scatter plot of the customers in the **cust_data** dataset. The variable online expense should be on one axis and the income should be on the other. Make sure your labels are professional. Use `ggplot()`
 - a. 1 point each for the correct plot, professional label for x-axis and y-axis
2. [3 points] According to the plot in Question 1A, how many clusters would you use for this dataset?
3. [3 points] Duplicate the **cust_data** to create a new data frame named **cust_norm_data**. The new data frame holds the normalized income and online expense values.
4. [3 points] Using the normalized data from Question 3 and the number of clusters from Question 2, generate the clusters using the k-means algorithm and both income and online_expense columns. Set your seed to zero.
5. [1 point] How many observations are there in each cluster?
6. [1 point] What is the variable values for the cluster centroids?
7. [6 points] Looking at values from Question 6, how would you characterize these clusters in terms of income (low/high/average) and online expense (low/high/average)?
8. [5 points] Generate an elbow chart for clusters of size 2,3,4,5,6,7,8,9,10. Does the elbow chart support your answer to Question 1B? Or after seeing the elbow chart, would you change the number of clusters? Why or why not?

Problem 3 [55 points]

We will use the mortgage.csv data set posted on canvas. Suppose we want to predict if a second-house-buyer would be likely to take out (**TAKEOUT**) a mortgage for a new house.

1. [5 points] Make ONE correlation matrix plot for the correlations among the variables Education, Experience, Income, creditExpen, and Securities. Describe your findings. 2 points for the plot. 3 points for your findings.
2. [6 points] Use ggplot to make the following plots. Make sure you make professional plots with the appropriate labels. 3 points for each plot: 0.5 point for the correct plot, 0.5 point each for the appropriate plot title, x-axis label, and y-axis label. 1 point for a one sentence description of what you found based on the plot.
 - a. Histogram and an added density curve (on the same plot) of the Income variable: 0.5 point for the correct plot, 0.5 point each for the appropriate plot title, x-axis label, and y-axis label. 1 point for a one sentence description of what you found based on the plot.
 - b. Use ggplot to make a pie chart of the variable CreditCard:
 - i. First create a df to summarize the number of 0's and 1's there are in the variable [1 point]
 - ii. Change the value 0 and 1 to characters [0.5 point]
 - iii. Create a pie chart[1.5 points]: <https://www.r-graph-gallery.com/piechart-ggplot2.html>
 - iv. One sentence interpretation/summary of the plot [1 point]
3. [5 points] Use ggplot to create a boxplot of the **income** variable based on three categories of **Education**. 2 points for correct plot. 0.5 point each for the appropriate plot title, x-axis label, and y-axis label. 1.5 point for a description of your finding. The x-axis should be labeled with the three education levels: undergraduate, graduate, advanced/professional
 - Check the customize discrete axis section of the following website for changing the x-axis label.
 - <http://www.sthda.com/english/wiki/ggplot2-axis-ticks-a-guide-to-customize-tick-marks-and-labels>
 -
4. [3 points] Normalize all the numerical data (Age, experience, income, family, creditExpen) and save the normalized data in new columns with the names "Age.n", "Experience.n", "Income.n", "Family.n", "creditExpen.n".
5. [1 point] randomly select 60% of the observations to be your training data set, and the rest to be your validation data set.

KNN model:

6. [6 points] Create 3 knn models, each with the number of clusters being 1, 2, and 3. Save the model as **preds.k.1**, **preds.k.2**, and **preds.k.3**. use the **normalized variables** "Age.n", "Experience.n", "Income.n", "Family.n", "creditExpen.n".
7. [6 point] Create a confusion matrix for each of the three models you just created and report the prediction accuracy, sensitivity and specificity values in the context of this problem. Also indicate whether the model is better at identifying buyers who actually take out a mortgage or better identifying buyers who would not take out mortgage.

Logistic Regression Model:

8. [4 points] Create a logistic regression model on your training data set. Explain why you included these variables. 2 points for explanation.
9. [4 points] Apply your logistic regression model using the validation data set [2 points]. If the predicted probability is larger than 0.5, we classify the buyer as “yes”, meaning that he will take out a mortgage [2 points].
10. [3 points] Create the confusion matrix and calculate the prediction accuracy.

Decision Trees:

11. [2 points] Create a decision tree.
12. [2 points] Plot the decision tree.
13. [3 points] Apply your model to the validation data to predict TAKEOUT
14. [3 points] Create a confusion matrix and calculate the prediction accuracy.

Comparison:

15. [1 point] Compare the three knn models, the one logistical regression, and the one decision tree model (list their prediction accuracy in words), which one performs the best?

Here is a description of the data set

1. ID : unique identifier
2. TAKEOUT : did the home buyer take out a mortgage for his second house (1=Yes, 0=No)
3. Age : buyer's age
4. Experience : number of years of profession experience
5. Income : annual income of the customer (\$000)
6. Zip code: home address zip code
7. Family : family size of the buyer
8. creditExpen: average spending on credit cards per month (\$000)
9. Education: education level (1) undergraduate, (2) graduate, (3) advanced/professional
10. Mortgage : value of a previous house mortgage (\$000)
11. Securities : does the buyer have a securities account with the bank? (1=Yes, 0=No)
12. CDAccount : does the buyer have a certificate of deposit with the bank? (1=Yes, 0=No)
13. Online : does the buyer use Internet banking facilities (1=Yes, 0=No)
14. CreditCard : does the buyer use a credit card issued by Chase? (1=Yes, 0=No)