Queen's

Master of Management in Artificial Intelligence

MMAI 869

Machine Learning and AI

Dr. Stephen W. Thomas

Individual Assignment 1

Version 2

November 5, 2018

Sakib Hasan (20139537)

ORDER OF FILES:

Filename	Pages	Comments and/or Instructions
Q1.R	-	
Q1	2-8	
Q2	9	
Q3	10-12	
Q4	13-14	
Q5	15-16	

Additional Comments:

1. ARREST THAT MAN!

You work in Washington, DC. You help politicians do… whatever politicians do. Like most political analysts, you would like to make a strong case about crime in America. In particular, you would like to analyze arrest rates in each state, and determine which states are most similar to each other in that regard. You find the `USArrests` dataset in the `datasets` package. You will now perform clustering on the states to determine groups of similar states.

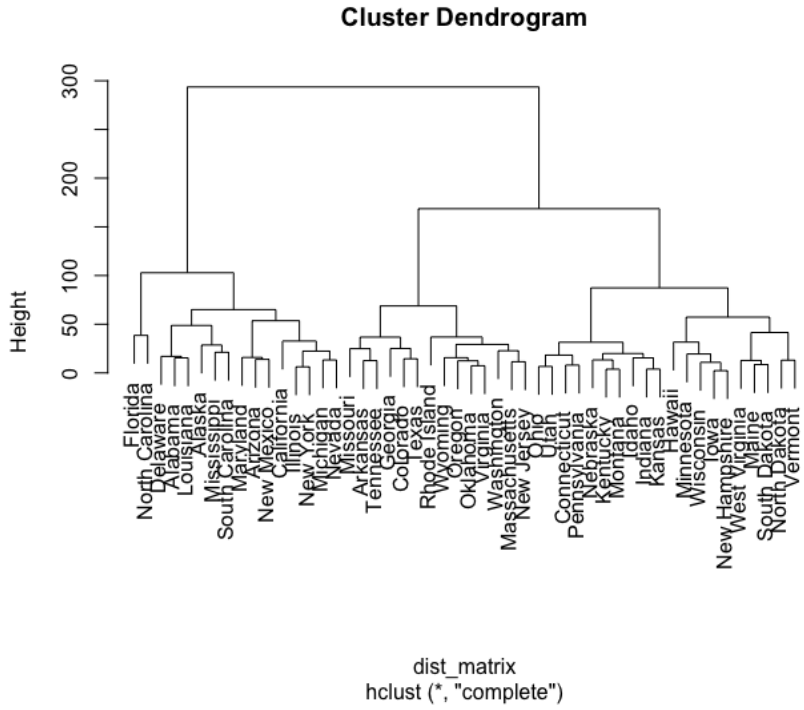
a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Hint: use the `hclust` package in R.
b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters? Provide a description/interpretation of each cluster.
c) Hierarchically cluster the states again, except this time, first scale all of the features to have standard deviation one. Hint: use the `scale()` function.

Answer:

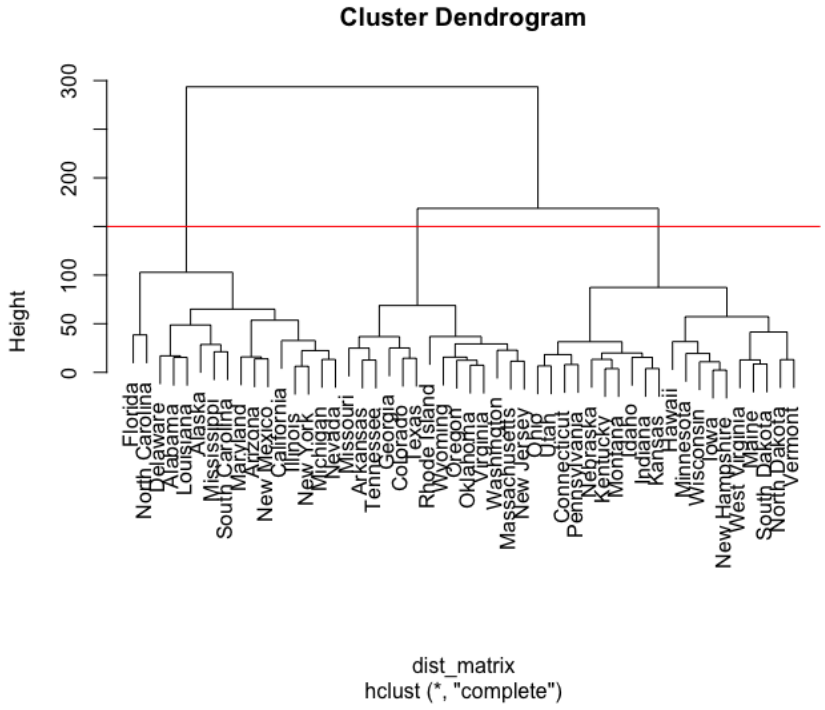
The R script for this question is submitted as *Q1.R*. Please refer to the script to understand how the analysis was done.

US NATIONAL AVERAGE ARRESTS	
Average Murder	7.78
Average Assault	170.76
Average Rape	65.54
Average Urban Population	21.23

a) Below is the dendrogram plot depicting the cluster of states:



b) Cut the dendrogram at height=150 that results in **three** distinct clusters:

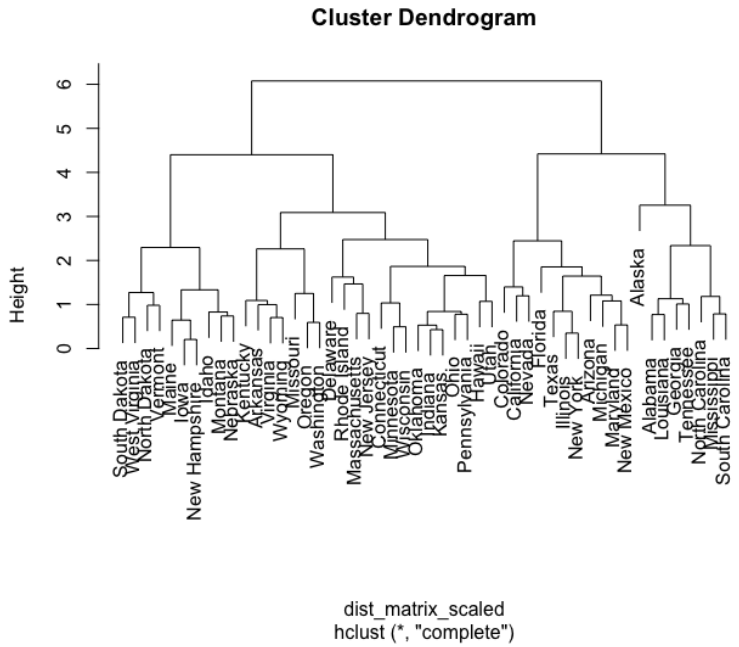


- **Cluster-1**
    - **States:** Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina
    - **Average Murder Rate:** 11.81
    - **Average Assault Rate:** 272.56
    - **Average Rape Rate:** 28.37
    - **Average Urban Population:** 68.31
    - **Interpretation:** All the high-end states fall in cluster-1. The rate of murder, assault and rape is higher than the national average. The urban population of the states in this cluster is also higher than the national average. These are the states where there are lots of tourist visits, major financial hubs and where well-known celebrities live.
- **Cluster-2**
    - **States:** Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming
    - **Average Murder Rate:** 8.21
    - **Average Assault Rate:** 173.28
    - **Average Rape Rate:** 22.84
    - **Average Urban Population:** 70.64
    - **Interpretation:** All the mid-range states fall in cluster-2. The rate of murder, assault and rape is still higher than the national average but not as high as cluster-1. The urban population of the states in this cluster is

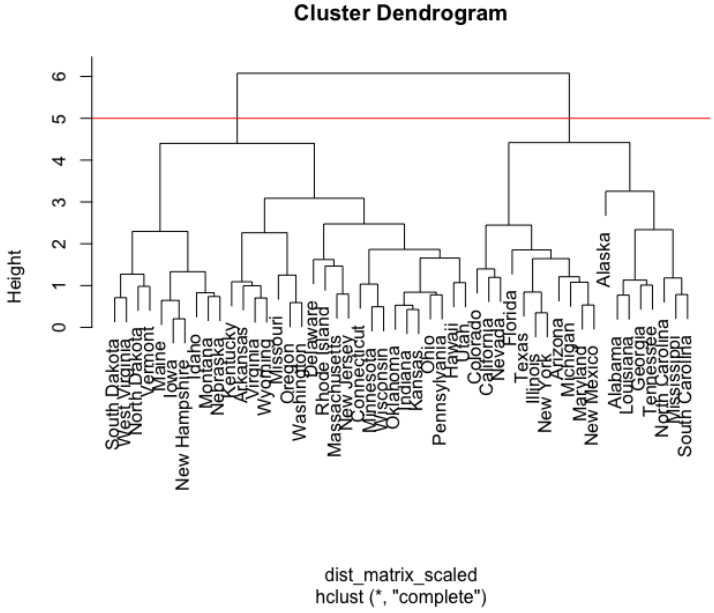
also higher than the national average. These are the states where there are some big financial institutions and national monuments.

- **Cluster-3**
    - **States:** Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, Vermont, West Virginia, Wisconsin
    - **Average Murder Rate:** 4.27
    - **Average Assault Rate:** 87.55
    - **Average Rape Rate:** 14.39
    - **Average Urban Population:** 59.75
    - **Interpretation:** All the low-end rural states fall in cluster-3. The rate of murder, assault and rape is lower than the national average. The urban population of the states in this cluster is also lower than the national average. These are the states where mostly aboriginal people reside and not much tourist activities are seen.

c) Hierarchically cluster the states again but instead by scaling all features to have a standard deviation of one. Below is the dendrogram plot depicting the cluster of states:



Cut the dendrogram at a height, (h=5), where we will have **two** distinct clusters:



- **Cluster-1**
    - **States:** Alabama, Alaska, Arizona, California, Colorado, Florida, Georgia, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina, Tennessee, Texas
    - **Average Murder Rate:** 12.33
    - **Average Assault Rate:** 259.31
    - **Average Rape Rate:** 29.21
    - **Average Urban Population:** 68.31
    - **Interpretation:** All the high-end states fall in cluster-1. The rate of murder, assault and rape is higher than the national average. The urban population of the states in this cluster is also higher than the national average. These are the states where there are lots of tourist visits, major financial hubs and well-known celebrities live.
- **Cluster-2**
    - **States:** Arkansas, Connecticut, Delaware, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Massachusetts, Minnesota, Missouri, Montana, Nebraska, New Hampshire, New Jersey, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Dakota, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming
    - **Average Murder Rate:** 5.00
    - **Average Assault Rate:** 116.48
    - **Average Rape Rate:** 16.33
    - **Average Urban Population:** 63.83
    - **Interpretation:** All the low-end rural states fall in cluster-2. The rate of murder, assault and rape is lower than the national average. The urban

population of the states in this cluster is also lower than the national average. These are the states where mostly aboriginal people reside and not much tourist activities are seen.

2. WHO'S THE BEST?

You work at a large financial services company. The marketing department presents your analytics team with a customer dataset, which has thousands and thousands of features, and millions of instances. Your goal is to classify whether customers will likely respond to a given offer, but first you must choose a classifier algorithm to use. Of the five classifier algorithms we've discussed in class (i.e., Decision Trees, Naïve Bayes, KNN, SVM, and Neural Networks), which algorithm would you choose? Why? State any additional assumptions you are making.

Answer:

For this business problem I would use the KNN (k-nearest neighbor) algorithm. As mentioned in the problem, the customer dataset has thousands and thousands of features and millions of instances, and most likely contains valuable information about each customer. KNN works best with large sets of data and this business problem is perfect suit for it. Using KNN we could look into each customer's previous behavior on a similar historical offer by using Euclidian distance matrixes to find the closest neighbors and classify them into group of customers that will respond to the new offer and group of customers that will not respond to the new offer. Some disadvantages of using KNN includes medium accuracy, the choice of K values drastically effects the computation and sometimes not as easy to interpret. Despite the disadvantages, KNN has exceptional training speed and could scale thousands and thousands of features. Hence, I would choose KNN classifier algorithm for this business problem.

3. THE INTERN

You are an in-demand, world-traveling, work-all-night consultant who specializes in designing supervised machine learning solutions for clients in a wide-range of industries. You have seen it all and you know what to do. To help you get more done in less time, you have hired an intern from Ivey, who, unfortunately, needs some handholding. Your intern does not understand when to use which classification measure. Your intern keeps getting it wrong. To help your intern learn from your experience, you have decided to look at some previous projects and describe which measure you used, and more importantly, why.

For each project below, describe which measure(s) are best, and why. Also, give an example of a measure which would be horrible to use, and why. List any assumptions you are making, about the dataset, problem, or business priorities that were involved in the project.

a) The fraud department at a bank wanted to predict which transactions were fraudulent. The training dataset had 100K credit card transactions, of which 97K are legit and 3K are fraud.
b) A hospital wanted to predict whether a MRI scan contained cancer.
c) A marketing team at a telco wanted to predict which customers were going to leave next month.
d) A city government wanted to build a system to monitor Twitter to see if any local residents were tweeting about emergencies that needed quick response from the police department.
e) A data science team wanted to find the best possible classifier for a given problem.
f) *[Describe one more project, whereby the best measure is one that you have not yet listed in parts a-e above.]*

Answer:

a) For this business problem **Precision** would be the best classification measure to use. Precision measures, out of the ones predicted to be "yes", how many were actually correct. This is the best classification measure for this business problem because at the fraud department of a bank, they would want to know out of the all the transactions made if they are predicted to be fraud, did they actually turn out to be fraud. For this business problem, accuracy would be a horrible measure to use. This is because the data could be imbalanced, and we could achieve high accuracy by always predicting "no fraud" even for a "fraud" labeled transaction.

b) For this business problem **Specificity** would be the best classification measure to use. Specificity measures how well the classifier can predict to the "no" case. A higher specificity means the classifier will predict "no" for most "no" cases (True Negative Rate) and when the

classifier predicts "yes", it is most likely the certain answer is also "yes". This is the best classification measure for this business problem because at a hospital where you want to predict whether an MRI scan contained cancer, you need to be sure about the fact if the classifier predicted no, the patient actually did not have cancer. Similarly, if the classifier said the patient had cancer, it is most likely the patient did have cancer. For this business problem, accuracy would also be a horrible measure to use. This is because the hospital could be okay with having false positive results (when the classifier predicts the patient to have cancer, but the patient does not have cancer), however, they would not want any false negative results (when the classifier predicts the patient to not have cancer, but the patient has cancer).

c) For this business problem **Sensitivity** would be the best classification measure to use. Sensitivity measures how well the classifier can predict to the "yes" case. It measures, out of the actual "yes", how many were predicted "yes" correctly. A higher sensitivity means the classifier will predict "yes" for most "yes cases (True Positive Rate) and when the classifier predicts "no", it is most likely the certain is also "no". This is the best classification measure for this business problem because if the classifier predicts the customer will be leaving next month, then most likely the customer will leave. For this business problem, accuracy would be a horrible measure to use. This is because the company would not want to have any false positive results as they might give the wrong customer a discounted deal to stay longer even though they were not thinking of leaving. However, they could be okay with false negative results where the customer actually left but the classifier predicted no. In that case, the company actually did not lose anything by providing discount to a small set of customers who left.

d) For this business problem **Accuracy** would be the best classification measure to use. Accuracy measures the percentage of cases that are classified correctly. As the city government you would want to make sure no emergencies go unnoticed and want to achieve

the highest accuracy measure and reduce the error measure. This is the case where the True Positive and True Negative cases are important to be classified correctly. The classifier needs to also minimize false negative cases when there is an actual emergency, but the classifier predicted no. For this business problem, precision would be a horrible measure to use because the classifier would only be measuring the ones where it predicted to be "yes" and have not taken into account any of the "no" predictions.

e) For this business problem **Area Under Curve** of the ROC plot would be the best classification measure to use. This measurement allows multiple classifiers to compared together and given the data science team does not have chosen classifier algorithm, the AUC measure would be the correct suit. For this business problem, no exact measure is horrible to use without additional knowledge about the problem the data science team is trying to solve.

f) An example project where the best measure is **F1-score** is to classify emails as spams and redirect them to the junk folder. F1-score is the measurement of the harmonic means of precision and recall scores. Sometimes precision and recall are far apart, however, the F1-score keeps that balance between the two values to uneven the class distribution. For this business problem, F1-score is the best measure because we want to predict correctly when an email is spam and redirect to the junk folder and don't want to predict an email that is not spam as a spam email (False Positive) and redirect to the junk folder. A classification measure that would be horrible to use for this project is specificity. This is because we don't want our classifier to predict an email is a spam and put in the junk folder when it is not a spam.

4. UNCLE STEVE'S GROCERY STORE

Uncle Steve runs a small, local grocery store. Looking for some customer insights, he has hired you to do some data science. He has given you a few years' worth of customer transactions, i.e., sets of items that customers have purchased. You have applied an association rules learning algorithm to the data, and the algorithm has generated a large set of association rules.

For each of the following scenarios, provide an example of one of the discovered association rules that satisfies the following conditions. (Just make up the rule, using your human experience and intuition!) Also, describe whether and why each rule would be considered subjectively interesting or uninteresting.

Answer:

Below is a snapshot of example transactions for customers. 1 represent the item was brought in that transaction and 0 represent it was not brought in that transaction.

TID	Cereal	Milk	Bread	Eggs	Chips	Coke
1	1	1	1	0	0	1
2	1	1	0	1	1	0
3	1	1	1	1	0	1
4	0	0	1	0	1	1
5	1	0	1	1	0	1

a) A rule that has high support and high confidence.

An association rule for this condition could be: $\{\boldsymbol{Cereal}\} \rightarrow \{\boldsymbol{Milk}\}$

This rule returns a support of 60% and confidence of 75% from our transaction database.

Intuitively, this rule makes sense because as a customer if they are buying cereal, they would most likely buy milk as well. Hence, I as a manager will take the chance to increase the price of either one of cereal or milk and decrease the price of the other one to promote a sale and maximize profit.

b) A rule that has reasonably high support but low confidence.

An association rule for this condition could be: $\{\boldsymbol{Coke}\} \rightarrow \{\boldsymbol{Milk}\}$

This rule returns a support of 40% and confidence of 50% from our transaction database. Intuitively, this rule is interesting because they are both liquid drinks which contradicts one another health wise. One is a healthy drink and the other is a not a healthy drink. I as a manager might resist from

combining both into a promotional deal, however, in the future if there are more transactions where milk and coke appear together, I may put them into a promotional deal.

c) A rule that has low support and low confidence.

An association rule for this condition could be: $\{\boldsymbol{Milk}, \boldsymbol{Bread}, \boldsymbol{Eggs}\} \rightarrow \{\boldsymbol{Chips}\}$

This rule returns a support and confidence of 0% from our transaction database. Intuitively, this is not an interesting rule because I as a manager know that it is very unlikely that if my customer buys milk, bread and eggs, then they will buy chips as well. Hence, I will resist from increasing the price of milk, bread or eggs to decrease the price of chips.

d) A rule that has low support and high confidence.

An association rule for this condition could be: $\{\boldsymbol{Milk}\} \rightarrow \{\boldsymbol{Cereal}, \boldsymbol{Eggs}\}$

This rule returns a support of 40% and confidence of 67% from our transaction database. This is an interesting rule because milk, cereal and eggs are not seen that often in the transactions together leading to a low support. However, the frequency that cereal and eggs appear in a transaction if a milk has been brought is high resulting in a higher confidence. Since the confidence is high, I as a manager would try to increase the price of cereal and eggs and decrease the price of milk to maximize profit.

5. VIVA LA VINO

Some Smith faculty have started a wine club. At each meeting, members of the club perform blind taste tests of different wine varietals. Members indicate how much they enjoy each varietal, using an integer scale of 1 (worst) to 7 (best). After the most recent meeting, here are the ratings.

	Zin	Pinot Noir	Chard	Merlot	Cab	Pinot Gris
Yuri	7	6	7	4	5	4
Steve		7	6	4	3	4
Gary	3	3	3	1	1	5
Qurat	2	2	1	3	7	4
Brigid	5	6	7	2	3	3

Unfortunately, the club ran out of Zin before Steve had a chance to try it. Luckily, the club has you, a data-driven, clever, and charming Queen's student. Use your skills to predict what Steve would rate Zin. Use user-based collaborative filtering with cosine distance. To predict the rating, find the two nearest neighbors, and take a weighted average of their scores of the item in question.

Hints:

- Recall that the cosine distance is calculated as $\boldsymbol{dist_{cos}(A,B) = 1 - \frac{\sum_{i=0}^{n} A_i B_i}{\sqrt{\sum_{i=0}^{n} A_i^2}\sqrt{\sum_{i=0}^{n} B_i^2}}}$
- When comparing two users, only include items that both users have rated. For example, to compare Yuri and Steve, the calculation would ignore Zin (since Steve hasn't rated it) and would be $1 - \frac{7*6+6*7+4*4+5*3+4*4}{\sqrt{7^2+6^2+4^2+5^2+4^2}\sqrt{6^2+7^2+4^2+3^2+4^2}} = 0.021$.
- To find the weighted average of two ratings R1 and R2, which have a distances D1 and D2 respectively, use the formula: Weighted_Average = ((1-D1)*R1 + (1-D2)*R2) / (2-D1 - D2).
- Round the predicted rating to the nearest integer.

Answer:

First, we find the cosine distance of Steven versus each member:

- $dist_{cos}(Steven, Yuri) = 1 - \frac{(6*7+7*6+4*4+5*3+4*4)}{\sqrt{7^2+6^2+4^2+3^2+4^2}*\sqrt{6^2+7^2+4^2+5^2+4^2}} = 0.021$
- $dist_{cos}(Steven, Gary) = 1 - \frac{(7*3+6*3+4*1+3*1+4*5)}{\sqrt{7^2+6^2+4^2+3^2+4^2}*\sqrt{3^2+3^2+1^2+1^2+5^2}} = 0.123$
- $dist_{cos}(Steven, Qurat) = 1 - \frac{(7*2+6*1+4*3+3*7+4*4)}{\sqrt{7^2+6^2+4^2+3^2+4^2}*\sqrt{2^2+1^2+3^2+7^2+4^2}} = 0.308$
- $dist_{cos}(Steven, Brigid) = 1 - \frac{(7*6+6*7+4*2+3*3+4*3)}{\sqrt{7^2+6^2+4^2+3^2+4^2}*\sqrt{6^2+7^2+2^2+3^2+3^2}} = 0.026$

The two minimum distances for Steven versus each member is 0.021 and 0.026 which corresponds to Yuri and Brigid respectively. We use the two distances and their corresponding ratings for Zin to recommend what Steven might rate Zin.

$$rating(Steven, Zin) = \frac{(0.979)(7) + (0.974)(5)}{(2 - 0.021 - 0.026)} = \frac{11.723}{1.953} \approx 6$$

∴ We can conclude Steven would most likely rate Zin a 6.