Homework 4

CSC522: Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2019

Instructions

- **Due Date:** Apr 18, 2019
- **Total Points**: 100
- Make sure you clearly list each team member's names and Unity IDs at the top of your submission.
- Your submission should be a single zip file containing a PDF of your answers, your code, and a README file with running instructions. Please follow the naming convention for your zip file: G(homework group number)_HW(homework number), e.g. G1_HW4.
- If any question is unclear, please post a question on Piazza. If you make any assumptions in your answer, you must state them explicitly (e.g. "Assuming the data in question is normally distributed...").
- If a question asks you to explain or justify your answer, **give a brief explanation** using your own ideas, not a reference to the textbook or an online source.
- If this homework needs to be updated for any reason, **we will post the update on Piazza** and announce it in class, so please stay alert.
- In addition to your group submission, please also *individually* submit your Peer Evaluation form on Moodle, evaluating yours and your teammates' contributions to this homework.
- While you will submit *one* homework as a team, you are responsible for *all* of the content on the homework. We recommend attempting to solve each question individually.

R Programming Submission Instructions

- Make sure you clearly list each team member's names and Unity IDs at the top of your submission.
- Your code should be named *hw4.R*. Add this file, along with a README to the zip file mentioned in the first page.
- Failure to follow naming conventions or programming related instructions specified below may result in your submission not being graded.
- Carefully read what the function names have been requested by the instructor. **In this homework or the following ones, if your code does not follow the naming format requested by the instructor, you will not receive credit.**
- For each function, both the input and output formats are provided in the *hw4.R*. Function calls are specified in *hw4_checker.R*. Please ensure that you follow the correct input and output formats. Once again, if you do not follow the format requested, you will not receive credit. It is clearly stated which functions need to be implemented by you in the comments in hw4.R
- You are free to write your own functions to handle sub-tasks, but the TA will only call the functions he has requested. If the requested functions do not run/return the correct values/do not finish running in specified time, you will not receive full credit.

- DO NOT set working directory (`setwd` function) or clear memory (rm(list=ls(all=T))) in your *hw4.R* code. TA(s) will do so in their own auto grader.
- The TA will have an autograder which will first run `source(hw4.R)`, then call each of the functions requested in the homework and compare with the correct solution.
- Your code should be clearly documented.
- To **test you code**, step through the `hw4_checker.R` file. If you update you code, make sure to run `source('./hw4.R')` again to update your function definitions. You can also check the "Source on save" option in R Studio to do this automatically on save.
- Please DO NOT include `install.packages()` or **unnecessary print statements** or `View` function calls in your `hw4.R`. Comment them out. Please note, we are specifying the allowed packages, which means that we already have them installed on our test machine. Having any of these in your code would result in a penalty of 5 points.

Problems

1. K-Means Clustering (16 points) [**Ruth Okoilu**] Use the K-means clustering algorithm with *Euclidean Distance* to cluster the 10 data points in Figure 1 into 3 clusters. Suppose that the initial seed centroids are at points: C, H and I. The data are also given in tabular format in Table 1.

   (a) After each iteration of k-means, report the coordinates of the new centroids and which cluster each data point belongs to. **Stop when the algorithm converges and clearly label on the graph where the algorithm converges.** To report your work, either:

      i. Draw the result clusters and the new centroid at the end of each round (including the first round). You can use the image `hw4q1.png`, included with your homework, to mark clusters and centroids. Additionally, indicate the coordinates (x, y) alongside corresponding centroids.

      ii. Give your answer in tabular format with the following attributes: Round (e.g. Round 1, 2, etc), Points (e.g. {A, B, C}), and Cluster_ID. Also report the centroids for each cluster after each round.

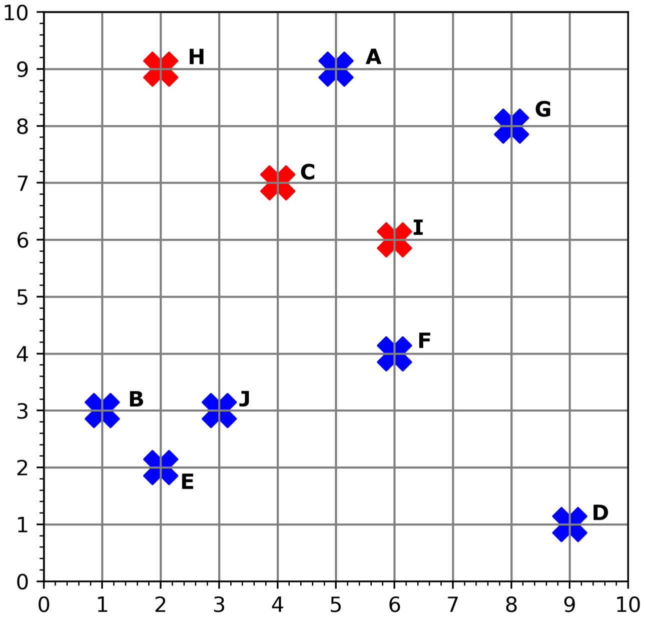


Figure 1: K-means Clustering (a)

Point	**x**	**y**
A	5	9
B	1	3
C	4	7
D	9	1
E	2	2
F	6	4
G	8	8
H	2	9
I	6	6
J	3	3

Table 1: K-means Clustering (b)

(b) How many rounds are needed for the K-means clustering algorithm to converge?

2. Hierarchical Clustering (18 points) [**Ruth Okoilu**] We will use the same dataset as in Question 1 (shown in Figure 1) for Hierarchical Clustering. The *Euclidean Distance* matrix between each pair of the datapoints is given in Table 2 below:

   (a) Perform *single* link hierarchical clustering. Show your work at each iteration by giving the inter-cluster distances. Report your results by drawing a corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged. If possible, use a program to construct your dendrogram (e.g. PowerPoint, LucidChart[1], or VisualParadigm[2]). Scanned hand drawings will also be accepted if they are very clear.

   (b) Perform *complete* link hierarchical clustering on the dataset. As above, show your calculations and report the corresponding dendrogram.

   (c) If we assume there are *two* clusters, will the *single* or *complete* link approach give a better clustering? Justify your answer.

   (d) Consider the single-link hierarchical clustering with **3 clusters**. Compare the quality of this single link clustering with your final K-means clustering in Question 1. To evaluate the quality of each clustering, calculate its corresponding Silhouette Coefficient. Based on this measure, which clustering (k-means, single link), is best? Do you agree with this assessment? Explain why in 1-2 sentences. Note: you may want to write some code to help speed up these calculations, which you can include in lieu of showing your work.

   The Silhouette coefficient[3] [1] is a measure of how well a clustering minimizes intra-cluster distance while maximizing inter-cluster distance. The Silhouette coefficient $s(i)$ for a given data point i in a given clustering can be computing using the formula:

$$s(i) = \frac{(b^i - a^i)}{max(a^i, b^i)}$$

   Where:

   a^i is the average distance from the point i to all *other* data points in the same cluster, and

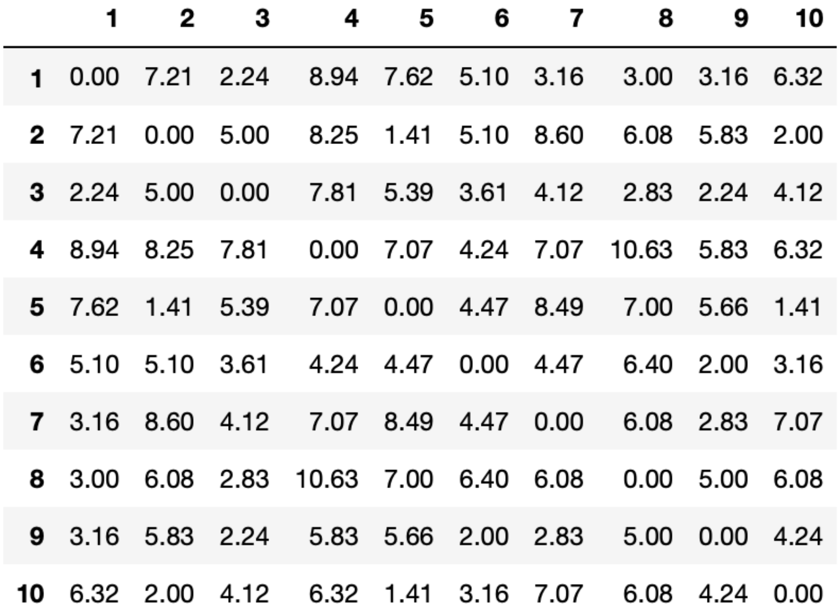
   b^i is the average distance from the point i to all data points in the closest cluster (i.e. the cluster for which b^i would be smallest).

   The Silhouette coefficient for a *clustering* is the average $s(i)$ over all data points i.

[1] `https://www.lucidchart.com/`

[2] `https://online.visual-paradigm.com/features/dendrogram-software/`

[3] `https://en.wikipedia.org/wiki/Silhouette_(clustering)`



	1	2	3	4	5	6	7	8	9	10
1	0.00	7.21	2.24	8.94	7.62	5.10	3.16	3.00	3.16	6.32
2	7.21	0.00	5.00	8.25	1.41	5.10	8.60	6.08	5.83	2.00
3	2.24	5.00	0.00	7.81	5.39	3.61	4.12	2.83	2.24	4.12
4	8.94	8.25	7.81	0.00	7.07	4.24	7.07	10.63	5.83	6.32
5	7.62	1.41	5.39	7.07	0.00	4.47	8.49	7.00	5.66	1.41
6	5.10	5.10	3.61	4.24	4.47	0.00	4.47	6.40	2.00	3.16
7	3.16	8.60	4.12	7.07	8.49	4.47	0.00	6.08	2.83	7.07
8	3.00	6.08	2.83	10.63	7.00	6.40	6.08	0.00	5.00	6.08
9	3.16	5.83	2.24	5.83	5.66	2.00	2.83	5.00	0.00	4.24
10	6.32	2.00	4.12	6.32	1.41	3.16	7.07	6.08	4.24	0.00

Figure 2: Hierarchical Clustering Dataset

3. Association Rule Mining (12 points) [**Song Ju**]. Consider the following market basket transactions shown in the Table 2 below.

Transaction ID	**Bread**	**Milk**	**Butter**	**Eggs**	**Beer**	**Cola**
1	1	1	0	0	0	1
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	0	1	0	0
5	1	1	1	0	0	1
6	0	0	1	0	1	1
7	0	1	0	1	0	0
8	1	0	1	0	1	0
9	1	1	0	0	0	1
10	0	0	0	1	0	1

Table 2: For each transaction (row), a 1 indicates that a given item was present in that transaction, and a 0 indicates that it was not.

(a) What is the maximum number of unique itemsets that can be extracted from this data set (only including itemsets that have ≥ 1 support)? Briefly explain your answer in 1-2 sentences.

(b) What is the maximum number of association rules that can be extracted from this data set (including rules that have zero support)? Briefly explain your answer in 2-3 sentences.

(c) Compute the support of the itemset: $\{Eggs, Cola\}$?

(d) Compute the support and confidence of association rule: $\{Bread\} \rightarrow \{Butter\}$?

(e) Given min support = 0.3 and min confidence = 0.6, identify all valid association rules of the form $\{A, B\} \rightarrow \{C\}$.

(f) In a different dataset, the support of the rule $\{a\} \rightarrow \{b\}$ is 0.46, and the support of the rule $\{a, c\} \rightarrow \{b, d\}$ is 0.23. What can we say for sure about the support of the rule $\{a\} \rightarrow \{b, d\}$. Explain in 1-2 sentences.

4. Apriori algorithm (16 points) [**Song Ju**]. Consider the data set shown in Table 3 and answer the following questions using apriori algorithm.

TID	Items
t_1	A,B,C,D
t_2	A,B,D,E
t_3	A,B
t_4	A,C,D
t_5	A,C,E
t_6	B,C
t_7	C,D
t_8	C,D,E

Table 3: Apriori algorithm

(a) Show (compute) each step of frequent itemset generation process using the apriori algorithm, with a minimum support count of 3.

(b) Show the lattice structure for the data given in table above, and mark the pruned branches if any. (Scanned hand-drawing is acceptable as long as it is clear.)

5. **Clustering and Neural Networks** In this question, you will be performing a variety of machine learning operations - clustering and classification.

   (a) **PART-1: Clustering** (28 points)

      i. **Dataset description:** You are provided a dataset with 2 variables (x, y). Your data is stored in the file `data/clustering-sample.csv`.

      ii. **Note:** The TA will use a different version of `data/clustering-sample.csv`. The format (variables x , y, will be similar, but TA's file may contain different number of data points, and may look visually different than the file supplied to you. Please ensure you take this into account, and do not hard code any dimensions/outputs.

      iii. In this exercise, you will apply three different types of clustering methods to the dataset supplied to you, and then compare their results:

         A. **Clustering:** You will write code in the function `alda_cluster()` to perform KMeans, Single Link and Complete Link clustering. Detailed instructions for implementation and allowed packages have been provided in `hw4.R` and in `hw4_checker.R`.

         B. **SSE Calculation:** You will write code in the function `alda_calculate_sse()` to calculate the SSE given the dataset and cluster assignments. Please note, you are not allowed to use any libraries related to SSE calculation for this. You must **implement** the SSE calculation.

         C. **Analysis-1: KMeans Elbow Plot:** You will write code in the function `alda_kmeans_elbow_plot()` to generate an elbow plot for a specific set of values of k on the dataset supplied to you. Note that you can use `alda_calculate_sse()` to calculate the SSE for each value of k. Generate this elbow plot, save it as instructed in `hw4.R` and place this plot in your PDF. Using this plot, report what you think is the best value of k and your justification for this choice in the PDF. Detailed instructions for implementation and allowed packages have been provided in `hw4.R` and in `hw4_checker.R`.

         D. **Analysis-2: Comparison of three clustering methods in terms of SSE:** Use the method `alda_calculate_sse()` for calculating SSE. I've given you code in `hw4_checker.R` which prints the SSE values for $k = 2$ for all the three clustering methods. Purely based on SSE, which clustering method do you think is the best? Report this answer in your PDF.

         E. **Analysis-3: Visual comparison of three clustering methods:** I've given you code for visualizing the clusters in `hw4_checker.R`, which plots the data with their cluster assignments for $k = 2$ for all the three clustering methods. Based on these plots, which clustering method do you think is the best? Report this answer in your PDF.

         F. **Analysis-4: Visual comparison vs SSE**: Is your answer for visual comparison (C) the same as for SSE (D) the same? What conclusions can you draw from this? How do visualizations compare with numeric measures of cluster quality? Are numeric measures always reliable? Explain your answers in 3-4 sentences.

G. **Sanity Checks:** The SSE values for KMeans, Single-Link and Complete-Link on the dataset given to you for 2 clusters approximately are 204.63, 319.7, 221.04.

(b) **PART-2: Classification** (10 Points)

i. **Dataset description:** You are provided a dataset with 5 variables. Variables $x1 - x4$ refer to the independent variables, while variable *class* is your class variable. Training data is stored in the file `data/classification-train.csv`, and test data is stored in the file `data/classification-test.csv`.

ii. **Note:** The TA will use a different version of `data/classification-test.csv`. The format (independent variables $x1 - x4$, dependent variable *class*) will be similar, but TA's file may contain different number of data points than the file supplied to you. Please ensure you take this into account, and do not hard code any dimensions.

iii. In this exercise, you will apply an Artificial Neural Network (ANN) classification method to the dataset supplied to you:

A. **Artificial Neural Networks:** You will use the `nnet` library for this purpose. In this exercise, you will implement perform grid search to identify the best set of hyperparameters such as the number of units in the hidden layers and decay. Use the function `alda_nn()` to perform hyperparameter tuning using grid search using `caret` and `nnet` packages. Detailed instructions for implementation, along with relevant functions and allowed packages have been provided in `hw4.R` and `hw4_checker.R`. Report the best hyperparameters you identified in your PDF.

B. **Sanity Checks:** If you followed all the instructions exactly and use no other additional parameters, you should achieve an overall accuracy of approximately 0.947 (on a scale of 1) on the test dataset you were given.

NOTE: Your entire solution `hw4.R` should not take more than 2 minutes to run. Any solution taking longer will be awarded a zero.

References

[1] D. heng and Q.-P. Wang, "Selection algorithm for k-means initial clustering center," in *Journal of Computer Applications 32, no. 8*, 2013. [Online]. Available: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a