CS 4210 – Assignment #5

Maximum Points: 100 pts.

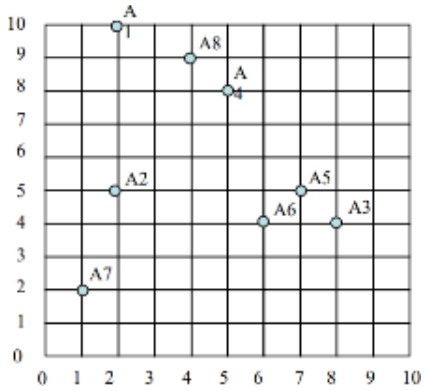
Bronco ID: |0|1|4|4|1|7|0|6|5|

Last Name: Just

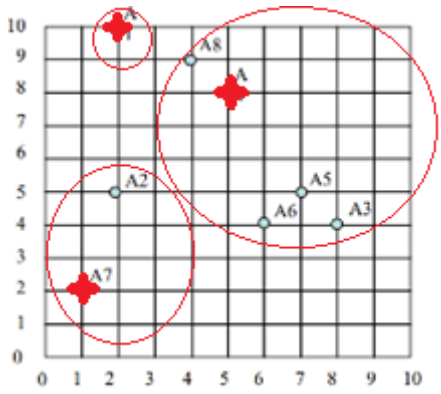
First Name: Kyle

Note 1: Your submission header must have the format as shown in the above-enclosed rounded rectangle.
Note 2: Homework is to be done individually. You may discuss the homework problems with your fellow students, but you are NOT allowed to copy – either in part or in whole – anyone else's answers.
Note 3: Your deliverable should be a .pdf file submitted through Gradescope until the deadline. Do not forget to assign a page to each of your answers when making a submission. In addition, source code (.py files) should be added to an online repository (e.g., github) to be downloaded and executed later.
Note 4: All submitted materials must be legible. Figures/diagrams must have good quality.
Note 5: Please use and check the Canvas discussion for further instructions, questions, answers, and hints. The bold words/sentences provide information for a complete or accurate answer.

1. [25 points] By considering the following 8 2D data points below do:
   a. [20 points] Group the points into 3 clusters by using k-means algorithm with Euclidean distance. Show the intermediate clusters (**by drawing ellipses on this 2D space**) and centroids (**by drawing marks like X on this 2D**) in each iteration until convergence. Also, **fill the solution table** below to show the distances from data points to the current centroids and the new centroids found. Consider the initial centroids as: C1 = A1, C2 = A4, and C3 = A7.



1st iteration								
Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	0	5	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
C2 dist.	$\sqrt{13}$	$\sqrt{18}$	5	0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
C3 dist.	$\sqrt{65}$	$\sqrt{10}$	$\sqrt{53}$	$\sqrt{52}$	$\sqrt{45}$	$\sqrt{29}$	0	$\sqrt{58}$
Cluster Assigned	C1	C3	C2	C2	C2	C2	C3	C2



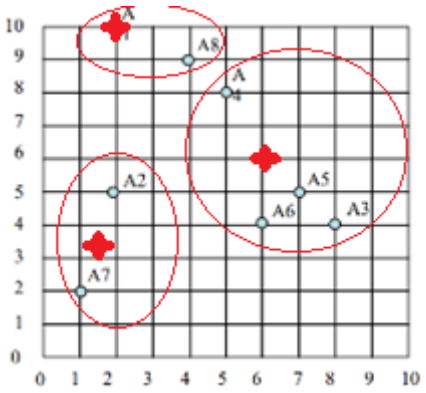
2nd iteration centroid:

C1 = (2, 10)

C2 = (6, 6)

C3 = (1.5, 3.5)

2nd iteration								
Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	0	5	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
C2 dist.	$\sqrt{32}\}$	$\sqrt{17}$	$\sqrt{8}$	$\sqrt{5}$	$\sqrt{2}$	2	$\sqrt{41}$	$\sqrt{13}$
C3 dist.	$\sqrt{42.5}$	$\sqrt{2.5}$	$\sqrt{42.5}$	$\sqrt{32.5}$	$\sqrt{32.5}$	$\sqrt{20.5}$	$\sqrt{2.5}$	$\sqrt{36.5}$
Cluster Assigned	C1	C3	C2	C2	C2	C2	C3	C1



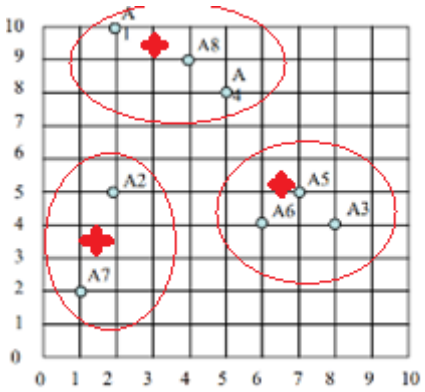
3rd iteration centroid:

C1 = (3, 9.5)

C2 = (6.5, 5.25)

C3 = (1.5, 3.5)

3rd iteration								
Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	$\sqrt{1.25}$	$\sqrt{21.25}$	$\sqrt{55.25}\}$	$\sqrt{6.25}$	$\sqrt{36.25}$	$\sqrt{39.25}$	$\sqrt{60.25}$	$\sqrt{1.25}$
C2 dist.	$\sqrt{42.8125}$	$\sqrt{20.3125}$	$\sqrt{3.8125}$	$\sqrt{9.8125}$	$\sqrt{0.3125}$	$\sqrt{1.8125}$	$\sqrt{40.8125}$	$\sqrt{20.3125}$
C3 dist.	$\sqrt{42.5}$	$\sqrt{2.5}$	$\sqrt{42.5}$	$\sqrt{32.5}$	$\sqrt{32.5}$	$\sqrt{20.5}$	$\sqrt{2.5}$	$\sqrt{36.5}$
Cluster Assigned	C1	C3	C2	C1	C2	C2	C3	C1



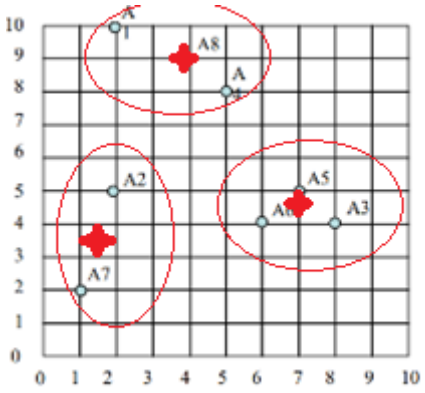
4th iteration centroid:

C1 = (3.665, 9)

C2 = (7, 4.333)

C3 = (1.5, 3.5)

4th iteration								
Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	$\sqrt{3.77}$	$\sqrt{18.77}$	$\sqrt{43.77}$	$\sqrt{2.77}$	$\sqrt{27.11}$	$\sqrt{30.44}$	$\sqrt{56.11}$	$\sqrt{0.11}$
C2 dist.	$\sqrt{57.11}$	$\sqrt{25.44}$	$\sqrt{1.11}$	$\sqrt{17.44}$	$\sqrt{0.44}$	$\sqrt{1.11}$	$\sqrt{41.44}$	$\sqrt{30.77}$
C3 dist.	$\sqrt{42.5}$	$\sqrt{2.5}$	$\sqrt{42.5}$	$\sqrt{32.5}$	$\sqrt{32.5}$	$\sqrt{20.5}$	$\sqrt{2.5}$	$\sqrt{36.5}$
Cluster Assigned	C1	C3	C2	C1	C2	C2	C3	C1



Centroids do not change! Convergence reached!

b. [5 points] Calculate the SSE (Sum of Square Errors) of the final clustering.
SSE = 14.33

2. [15 points] Complete the Python program (clustering.py) that will read the file training_data.csv to cluster the data. Your goal is to run k-means multiple times and check which k value maximizes the Silhouette coefficient. You also need to plot the values of k and their corresponding Silhouette coefficients so that we can visualize and confirm the best k value found. Next, you will calculate and print the Homogeneity score (the formula of this evaluation metric is provided in the template) of this best k clustering task by using the testing_data.csv, which is a file that includes ground truth data (classes).

3. [20 points] The dataset below presents the user ratings on a 1-3 scale for 6 different rock bands.

	Bon Jovi	Metallica	Scorpions	AC/DC	Kiss	Guns n' Roses
Fred	1	3	-	3	1	3
Lillian	3	-	2	2	3	1
Cathy	2	2	2	3	-	2
John	3	2	2	2	?	?

a. [10 points] Apply **user-based** collaborative filtering on the dataset to decide about recommending the bands Kiss and Guns n' Roses to John. You should make a recommendation when the predicted rating is greater than or equal to 2.0. Use cosine similarity, a neutral value (1.5) for missing values, and the top 2 similar neighbors to build your model.

Kiss

Fred = (1, 3, 1.5, 3, 3)

Lillian = (3, 1.5, 2, 2, 1)

Cathy = (2, 2, 2, 3, 2)

John = (3, 2, 2, 2, 1.5)

---

cos(Fred, John) = 0.848415

cos(Lillian, John) = 0.990865

cos(Cathy, John) = 0.953996

Choose Lillian and Cathy

$$r_{ij} = 2.25 + \frac{\big(0.990865 * (3 - 2.2)\big) + \big(0.953996 * (1.5 - 2.2)\big)}{(0.990865 * 0.953996)} = 2.38$$

Kiss is recommended

Guns n' Roses

Fred = (1, 3, 1.5, 3, 1)

Lillian = (3, 1.5, 2, 2, 3)

Cathy = (2, 2, 2, 3, 1.5)

John = (3, 2, 2, 2, 1.5)

---

cos(Fred, John) = 0.857350

cos(Lillian, John) = 0.955972

cos(Cathy, John) = 0.956989

Choose Lillian and Cathy

$$r_{ij} = 2.25 + \frac{\left(0.955972 * (1 - 2.2)\right) + \left(0.956989 * (2 - 2.2)\right)}{(0.955972 * 0.956989)} = 0.787$$

Guns n' Roses is not recommended

b. [10 points] Now, apply **item-based** collaborative filtering to make the same decision. Use the same parameters defined before to build your model.

Guns n' Roses = (3, 1, 2)
Kiss = (1, 3, 1.5)
AC/DC = (3, 2, 3)
Scorpions = (1.5, 2, 2)
Metallica = (3, 1.5, 2)
Bon Jovi = (1, 3, 2)

Kiss

cos(Kiss, Guns n' Roses) = 0.687243
cos(Kiss, AC/DC) = 0.822346
cos(Kiss, Scorpions) = 0.937043
cos(Kiss, Metallica) = 0.768221
cos(Kiss, Bon Jovi) = 0.992685

Choose Bon Jovi and Scorpions

$$r_{ij} = 1.83 + \frac{\left(0.937043 * (2 - 2)\right) + \left(0.992685 * (3 - 2)\right)}{(0.937043 * 0.992685)} = 2.897$$

Kiss is recommended

Guns n' Roses

cos(Guns n' Roses, Kiss) = 0.687243
cos(Guns n' Roses, AC/DC) = 0.968665
cos(Guns n' Roses, Scorpions) = 0.876523
cos(Guns n' Roses, Metallica) = 0.99236
cos(Guns n' Roses, Bon Jovi) = 0.714286

Choose AC/DC and Metallica

$$r_{ij} = 2 + \frac{\big(0.968665 * (2 - 2.67)\big) + \big(0.99236 * (2 - 2.16)\big)}{(0.968665 * 0.99236)} = 1.160$$

Guns n' Roses is not recommended

4. [25 points] Consider the following transaction dataset.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Suppose that minimum support is set to 30% (*minsup*) and minimum confidence is set to 60%.

a. [5 points] Rank all frequent itemsets according to their support (list their support values).

1) {d} 9/10
2) {b} 7/10
3) {e} 6/10
4) {b, d} 6/10
5) {d, e} 6/10
6) {a} 5/10
7) {c} 5/10
8) {a, d} 4/10
9) {a, e} 4/10
10) {b, e} 4/10
11) {c, d} 4/10

12) {a, d, e} 4/10

13) {b, d, e} 4/10

14) {a, b} 3/10

15) {b, c} 3/10

b. [5 points] For all frequent 3-itemsets, rank all association rules - according to their confidence values - which satisfy the requirements on minimum support and minimum confidence (list their confidence values).

1) $\{a, d\} \rightarrow \{e\} = \frac{4}{4}$
2) $\{a, e\} \rightarrow \{d\} = \frac{4}{4}$
3) $\{b, e\} \rightarrow \{d\} = \frac{4}{4}$
4) $\{d, e\} \rightarrow \{a\} = \frac{4}{6}$
5) $\{b, d\} \rightarrow \{e\} = \frac{4}{6}$
6) $\{d, e\} \rightarrow \{b\} = \frac{4}{6}$

c. [5 points] Show how the 3-itemsets candidates can be generated by the F_{k-1} X F_{k-1} method and if these candidates will be pruned or not.

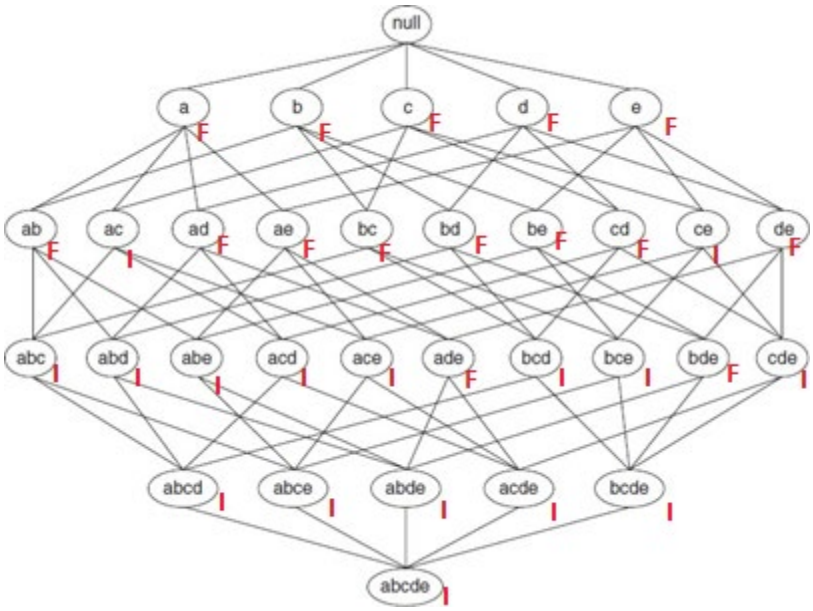
K-1 frequent sets (in lexical order):

{a, b}
{a, d}
{a, e}
{b, c}
{b, d}
{b, e}
{c, d}

{d, e}
Merge({a, b}, {a, d}) = {a, b, d}
Merge({a, b}, {a, e}) = {a, b, e}
Merge({a, d}, {a, e}) = {a, d, e}
Merge({b, c}, {b, d}) = {b, c, d}
Merge({b, c}, {b, e}) = {b, c, e}
Merge({b, d}. {b, e}) = {b, d, e}

Prune {b, c, e} because {c, e} is infrequent
We are left with:
{a, b, d}
{a, b, e}
{a, d, e}
{b, c, d}
{b, d, e}

d. [10 points] Consider the lattice structure given below. Label each node with the following letter(s): *F* if it is frequent and *I* if it is infrequent.



5. [15 points] Complete the Python program (association_rule_mining.py) that will read the file retail_dataset.csv to find strong rules related to supermarket products. You will need to install a python library this time. Just use your terminal to type: pip install mlxtend. Your goal is to output the rules that satisfy *minsup* = 0.2 and *minconf* = 0.6, as well as the priors and probability gains of the rule consequents when conditioned to the antecedents. The formulas for this math are given in the template.

GitHub link: https://github.com/meap02/Assignment-5_CS4210-Machine-Learning-and-its-Applications

Important Note: Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may resubmit multiple times until the deadline (the last submission will be considered).

NO LATE ASSIGNMENTS WILL BE ACCEPTED. ALWAYS SUBMIT WHATEVER YOU HAVE COMPLETED FOR PARTIAL CREDIT BEFORE THE DEADLINE!