Politecnico di Milano
School of Industrial and Information Engineering

Data Mining and Text Mining

June 28, 2019

NAME

CODICE PERSONA/ID

GENERAL INSTRUCTIONS

- **Answers must be clearly written inside the answer box designated for each problem.**
- **All the answers must be adequately motivated.**
- **Pencils are not allowed. The exam consists of 7 sheets of paper. It must be returned with all the 7 sheets. No any other sheet can be added. No sheet can be removed.**
- **This is a closed-book/closed-notes exam.**
- **Only non-programmable calculators are allowed.**
- **Notes/books/mobile phones are not allowed.**
- **If you are caught using forbidden material, the exam will immediately end and an RP grade will be recorded; then, your Data Mining exam will consist of an oral examination from then on.**

COURSE PROJECT SCORE

FINAL TIME

GRADES

1	2	3
4	5	6

SCORING

- **A problem left unsolved will amount to zero points.**
- **A completely wrong solution will amount to -3 points**

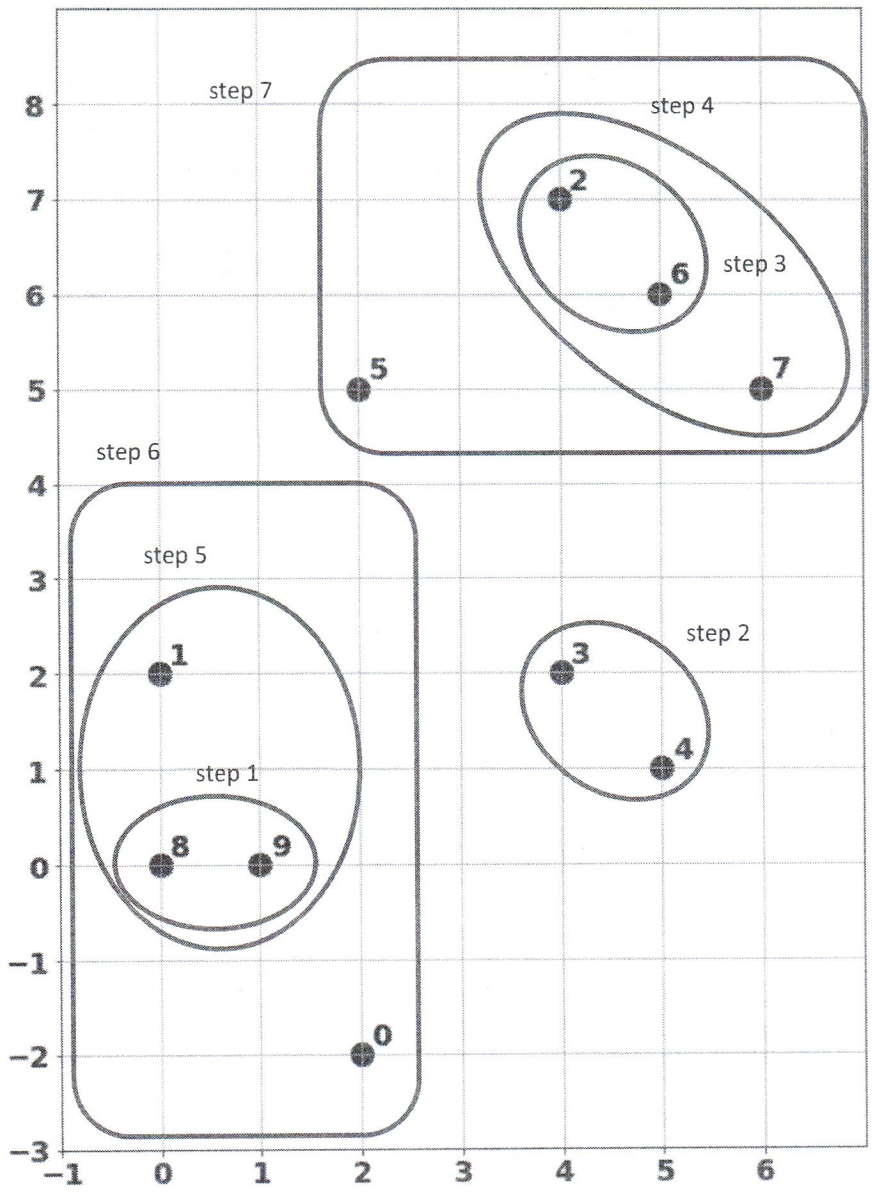
UIC STUDENTS HAVE 1:00h TO SOLVE PROBLEMS 3 AND 4

STUDENTS WHO DID THE COURSE PROJECT HAVE 1:40h TO SOLVE PROBLEMS 1, 2, 3, AND 4

ALL THE OTHER STUDENTS HAVE 2:20h TO SOLVE ALL THE SIX PROBLEMS

Problem 1 (6 points). Given the dataset below, described by the two variables, apply hierarchical clustering using Euclidean distance and single linkage (MIN). Plot the dendrogram and compute WSS when the solution with two clusters is considered.

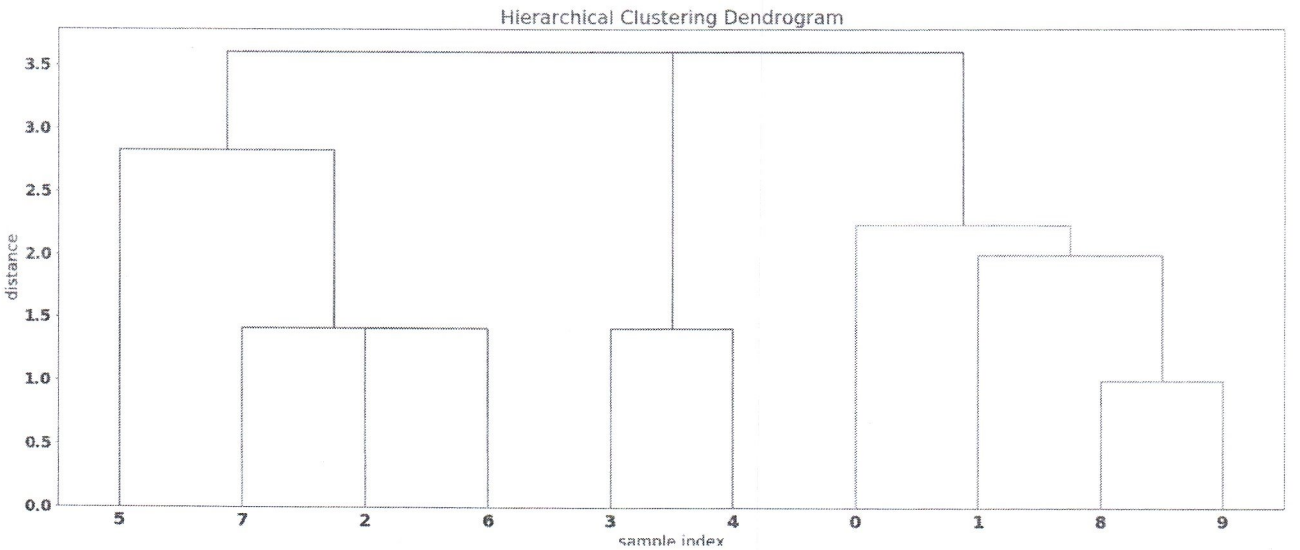
	x0	x1
0	2.0	-2.0
1	-0.0	2.0
2	4.0	7.0
3	4.0	2.0
4	5.0	1.0
5	2.0	5.0
6	5.0	6.0
7	6.0	5.0
8	0.0	-0.0
9	1.0	-0.0



Then the last three clusters are at the same distance, so they are merged together in step 8 and 9.

Problem 1 (continued).

Dendrogram



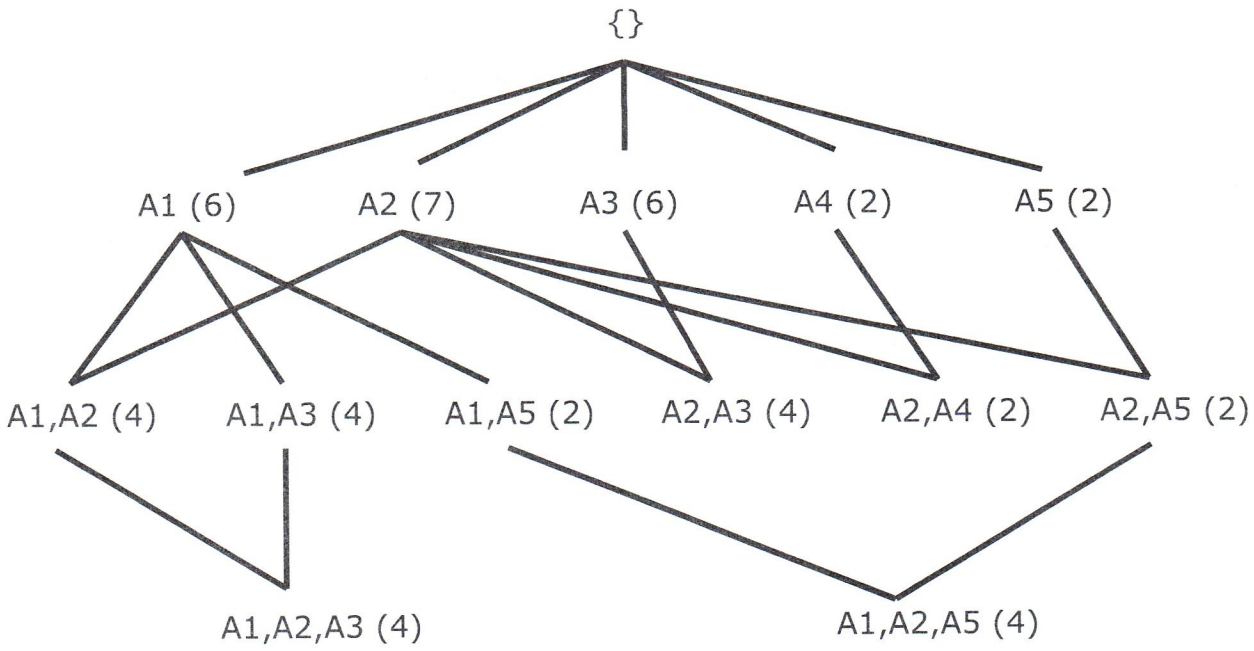
WSS of the Two Cluster Solution

Depending on the final merge we either consider the two clusters [5,7,2,6] and [3,4,0,1,8,9] or [5,7,2,6,3,4] and [0,1,8,9]

Problem 2 (6 points). Given the database below containing 9 transaction. Given a minimum support of 2/9 and a minimum confidence of 7/9:

- Apply A-priori to compute the frequent patterns
- Compute the association rules with only one item in the tail

T1	A1, A2, A5
T2	A2, A4
T3	A2, A3
T4	A1, A2, A4
T5	A1, A3
T6	A2, A3
T7	A1, A3
T8	A1, A2, A3, A5
T9	A1, A2, A3



Problem 2 (continued).

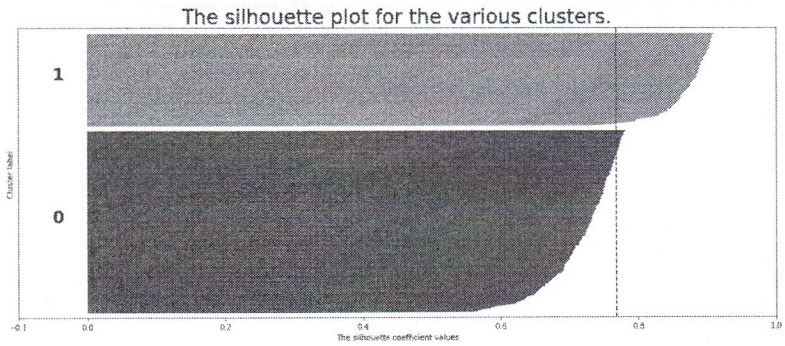
Frequent Itemsets	
1-Items	A1, A2, A3, A4, A5
2-Items	{A1,A2} {A1,A3} {A1,A5} {A2,A3} {A2,A4} {A2,A5}
3-Items	{A1,A2,A3} {A1,A2,A5}
4-Items	
5-Items	

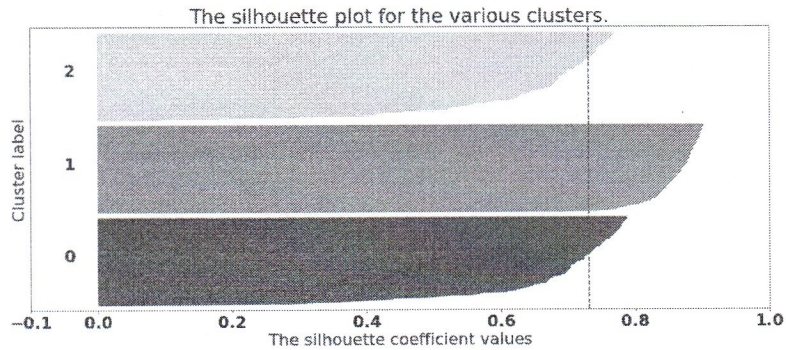
Association Rules
A4 -> A2 Support 0.22 Confidence =1.00
A5 -> A1 Support 0.22 Confidence =1.00
A5 -> A2 Support 0.22 Confidence =1.00
A1,A5 -> A2 Support 0.22 Confidence =1.00
A2,A5 -> A1 Support 0.22 Confidence =1.00

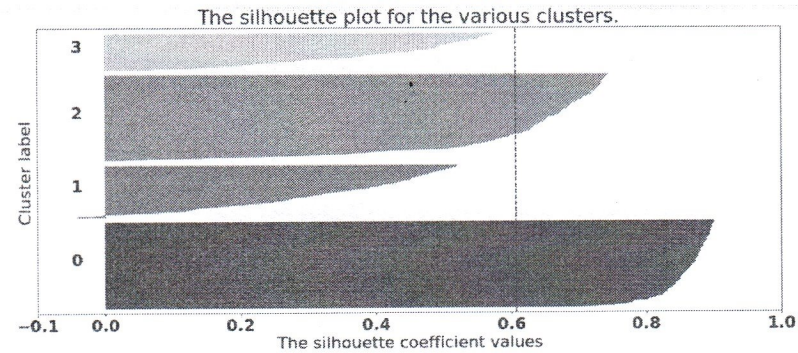
Problem 3 (6 points). You applied k-means with different values of k (2, 3, 4, 5) to a data set. To compare the solutions you first compute the following average silhouette score:

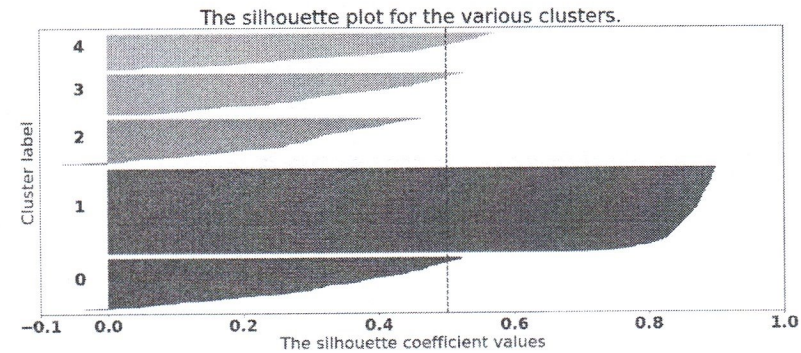
# Clusters	Average Silhouette Score
2	0.77
3	0.73
4	0.60
5	0.50

And then further analyzed the results creating the silhouette plots below.









Problem 3 (continued).

Question #1: Given the average silhouette score, what value of k would you choose?

This problem is taken from the python notebook and has been discussed in class. Check the recordings for further discussions.

Given the average silhouette score, we select k=2 since it is the higher score. But as stated during the course, the score tells only one side of the story.

Question #2: Discuss each silhouette plot adequately commenting why the corresponding value of k should be chosen or discarded

Comment for Silhouette Plot k=2

We have to clusters one completely above the average silhouette score and another very large cluster with almost all the points below the average silhouette score. Good candidate.

Comment for Silhouette Plot k=3

We have three clusters with almost the same number of data points with data points above the average silhouette score. Better candidate than the previous one.

Comment for Silhouette Plot k=4

Two clusters are way below the average silhouette score, so not a good candidate.

Comment for Silhouette Plot k=5

Three clusters have very few points above the average which in this case is very low so it seems that these clusters are now well defined. Not a good candidate, probably the solution with k=3 is the best one at this point.

Problem 4 (6 points). Consider the following dataset containing three documents:

Id	Document
1	A time to plant and a time to reap
2	Time for you and time for me
3	Fly Time

- Compute the TF-IDF representation of this data set taking into account that the stopwords ("a", "to", "and", "for") should not be included in the TF-IDF table
- Compute the cosine similarity of the documents represented by the following bags of words:
  - ("plant","you","fly","me","reap","time")
  - ("time","plant","reap","you",me","fly")
  - ("time", "plant", "you", "me", "fly")

Note: Answers and computations should be written in the corresponding boxes.

TF Representation
For this we just use the number of times each word appears in the document.

Id	Time	Plant	Reap	You	Me	Fly		
1	2	1	1	0	0	0		
2	2	0	0	1	1	0		
3	1	0	0	0	0	1		

IDF Values
This is computed as log(M+1)/k where M is the number of document and k in the number of documents where the word appears. We use base 2 to compute the logarithm.

Word	Time	Plant	Reap	You	Me	Fly			
IDF	log(4/3)	log(4)	log(4)	log(4)	log(4)	log(4)			

Using base 2, log(4/3) is 0.42 and log(4) is 2. Using base 10, it would be 0.12 and 0.60

TF-IDF Representation

Id	Time	Plant	Reap	You	Me	Fly		
1	2 * 0.42 = 0.84	1*2 = 2	1*2 = 2	0	0	0		
2	2 * 0.42 = 0.84	0	0	1*2 = 2	1*2 = 2	0		
3	1 * 0.42 = 0.42	0	0	0	0	1*2 = 2		

Problem 4 (Continued).

Cosine Similarity for ("plant","you","fly","me","reap","time")

We compute the cosine similarity between this bag of words and the three documents. Cosine similarity is computed like cosine distance without applying the arcos (see https://youtu.be/X11GUl1ama8 around minute 12:50). For this purpose, we convert the bag of words to a vector using TF-IDF. In this case, frequency is one for every word. Thus, ("plant","you","fly","me","reap","time") becomes, (0.42, 2, 2, 2, 2, 2)

Similarity for document 1, is 0.64; similarity for document 2 is also 0.64; while similarity for document 3 is 0.51

Cosine Similarity for ("time","plant","reap","you", "me","fly")

The computation is the same for the previous case since the order in the bag of word does not change the computation.

Cosine Similarity for ("time", "plant", "you", "me", "fly")

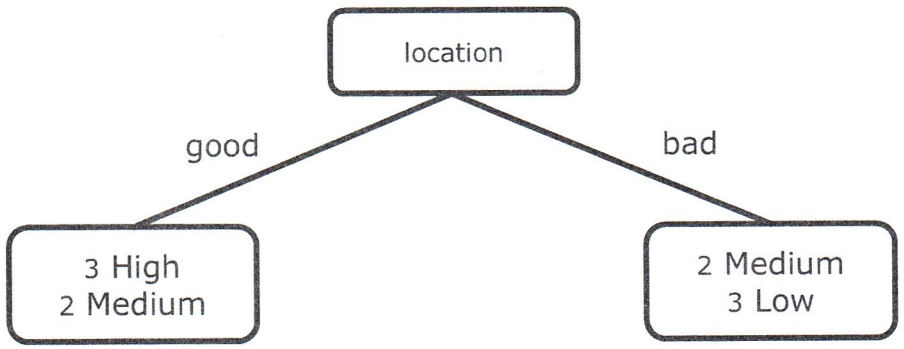
The bag of word is converted to the vector (1.33, 2, 0, 2, 2, 2) and we reapply the similarity computation as before. Similarity for document 1, is 0.46; similarity for document 2 is also 0.70; while similarity for document 3 is 0.57

Problem 5 (6 points). Given the dataset below, where "value" is the class attribute, compute the best decision stump using information gain.

location	size	pets	value
good	small	yes	high
good	big	no	high
good	big	no	high
bad	medium	no	medium
good	medium	only cats	medium
good	small	only cats	medium
bad	medium	yes	medium
bad	small	yes	low
bad	medium	yes	low
bad	small	no	low

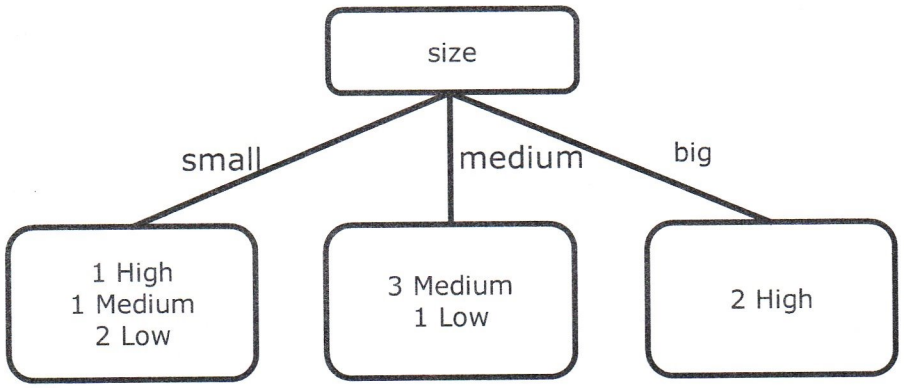
Decision Stump

Info(D) = Entropy([3,4,3]) = 1,57
Gain(Location) = Info(D) – 0.5 Info([3,2,0]) – 0.5 Info([0,2,3]) = Info(D) - Info([3,2,0]) = 1,57 – 0.97 = 0.60



Decision Stump

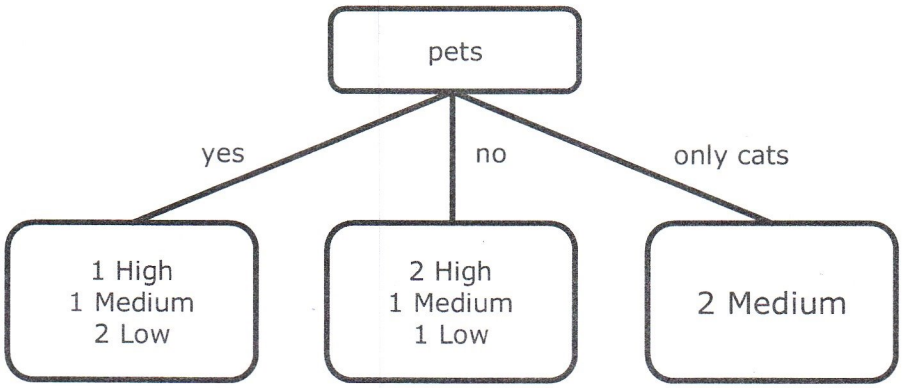
Info(D) = Entropy([3,4,3]) = 1,57
Gain(size) = Info(D) – 0.4 Info([1,1,2]) – 0.4 Info([3,1]) = 0.64



Decision Stump

Gain(pets) = Info(D) - 0.4 Info([1,1,2]) - 0.4 Info([2,1,1]) = 0.37

Note that we already computed Info([1,1,2]) at the previous step.



Decision Stump

What is the best decision stump?

Best decision stump is for attribute size.

Problem 6 (3 points). You are an employee of ClusterThis! A company specialized in the clustering of massive amount of data, described by thousands of variables, using k-means clustering. While you are sitting at your desk your boss storms into the room and tells you that is worried since your competitor (the company ThisIsClustering) is applying a brand-new method that is very successful called Mean Shift Clustering. Your boss is worried and asks you whether, for what you are doing, k-means is still the best option. What would be the pros for switching to Mean Shift Clustering? What would be the cons? Elaborate an answer for your boss.

There are several pros and cons in replacing k-Means with Mean Shift Clustering. From the discussion, it appears there are no issues about k-Means, in fact our boss does not mention existing issues with k-Means. She is only asking whether we should switch. So, k-Means is working and the question is whether it would be a good idea to switch. The only thing we know about our application is that we apply k-Means to massive amounts of data and have thousands of variables and this is a problem for Mean Shift that we know has a higher computational complexity with respect to k-Means ($O(Tn^2)$ instead of $O(Tkn)$ see the slides). So, no, it would not be a good idea to switch.