German University in Cairo
Faculty of Media Engineering and Technology
Mervat Abu-Elkheir and Ayman Al-Serafi
Tameem Al-Ghazaly, Ghada Mansour, Sarah Samir & Nada Bakeer

CSEN911: Data Mining

Winter Term 2023

Problem Set 1 – Classification

Problem 1

Two document classification systems, X and Y, are being compared. Both are given the same query, applied to a collection of 1500 documents, and should return a set of documents that are relevant to the query. System X returns 400 documents, of which 40 are relevant to the query. System Y returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are in fact 50 documents relevant to the query.

a) What are the values of True Positives, False Positives, True Negatives, False Negatives for system X? ***Hint***: The True Positives are the documents that are retrieved and relevant. The False Positives are the documents that are retrieved but are not relevant.

b) Calculate the precision and recall for the 2 systems, showing the details of your calculations.

Answer

	Relevant (50)	Not relevant (1500 – 50)	Total
Retrieved (400)	40	360	400
Not retrieved (1500 – 400)	10	1090	1100
Total	50	1450	1500

a) **TP** = 40

FP = 400 – 40 = 360

TN = 1100 – 10 = 1090

FN = 50 – 40 = 10

b) **Precision (sys. X) =** $Precision = \frac{TP}{TP+FP} = \frac{TP}{Retrieved}$ = 40/400 = 0.1

Precision (sys. Y) = 15/30 = 0.5

Recall (sys. X) = $Recall = \frac{TP}{TP+FN} = \frac{TP}{Positive(Relevant)}$ = 40/50 = 0.8

Recall (sys. Y) = 15/50 = 0.3

Problem 2

The following table shows the midterm and final exam grades obtained for students in a database course.

Student ID	x Quiz	y Midterm exam	z Final exam
1	8	14	84
2	7	10	63
3	6	14	78
4	9	19	90
5	6	17	75
6	9	16	79
7	5	6	52
8	9	18	74
9	5	13	77
10	8	16	90

Predict the final exam grade of a student who received a 9 on the quiz and a 17 on the midterm exam using the k-nearest neighbor algorithm with $k = 3$. ***Hint***: Use the average of the final exam grades for the k-nearest neighbors for final prediction.

Answer

We measure the distance between the student and each of the 12 students in the dataset:

Distance to 1	$= \sqrt{(9-8)^2 + (17-14)^2} = 3.2$
Distance to 2	$= \sqrt{(9-7)^2 + (17-10)^2} = 7.3$
Distance to 3	$= \sqrt{(9-6)^2 + (17-14)^2} = 4.3$
Distance to 4	$= \sqrt{(9-9)^2 + (17-19)^2} = 2$
Distance to 5	$= \sqrt{(9-6)^2 + (17-17)^2} = 3$
Distance to 6	$= \sqrt{(9-9)^2 + (17-16)^2} = \mathbf{1}$
Distance to 7	$= \sqrt{(9-5)^2 + (17-6)^2} = 11.7$
Distance to 8	$= \sqrt{(9-9)^2 + (17-18)^2} = \mathbf{1}$
Distance to 9	$= \sqrt{(9-5)^2 + (17-13)^2} = 5.7$
Distance to 10	$= \sqrt{(9-8)^2 + (17-16)^2} = \mathbf{1.4}$

Therefore, the students who are the 3 nearest neighbors to the given student are students with **IDs** = **6**, **8**, and **10**. The student's final grade will be the average of their final grades:

$$\boldsymbol{Final = \frac{79 + 74 + 90}{3} = 81}$$

Problem Set 2 – Clustering and Outlier Analysis

Problem 4

The following table shows the midterm and final exam grades obtained for students in a database course.

Student ID	x Quiz	y Midterm exam	z Final exam
1	8	14	84
2	7	10	63
3	6	14	78
4	9	19	90
5	6	17	75
6	9	16	79
7	5	6	52
8	9	18	74
9	5	13	77
10	8	16	90

a) If we want to group the students in the table into two clusters, and choose the initial cluster centroids to be the student with $\boldsymbol{ID = 2}$ (for cluster $\boldsymbol{C_1}$) and the student with $\boldsymbol{ID = 9}$ (for cluster $\boldsymbol{C_2}$). Which students will be clustered with student **2** in cluster $\boldsymbol{C_1}$ and which students will be clustered with student **9** in cluster $\boldsymbol{C_2}$? Use only the x and y attributes for clustering and ignore z. Show your solution steps.

b) Comment on the choice of initial centroids. Did this choice result in a good cluster structure? Justify your answer.

c) Assuming you will use the clusters computed in **(a)** to assign new students with either cluster. To which cluster will the student who received a 9 on the quiz and a 17 on the midterm exam be assigned?

Answer

a) Compute distance of each student to students **2** and **9**

Student ID	Distance to 2	Distance to 9
1	$= \sqrt{(8-7)^2 + (14-10)^2} = 4.1$	$= \sqrt{(\mathbf{8-5})^2 + (\mathbf{14-13})^2} = \mathbf{3.2}$
3	$= \sqrt{(6-7)^2 + (14-10)^2} = 4.1$	$= \sqrt{(\mathbf{6-5})^2 + (\mathbf{14-13})^2} = \mathbf{1.4}$
4	$= \sqrt{(9-7)^2 + (19-10)^2} = 9.2$	$= \sqrt{(\mathbf{9-5})^2 + (\mathbf{19-13})^2} = \mathbf{7.2}$
5	$= \sqrt{(6-7)^2 + (17-10)^2} = 7.1$	$= \sqrt{(\mathbf{6-5})^2 + (\mathbf{17-13})^2} = \mathbf{4.1}$
6	$= \sqrt{(9-7)^2 + (16-10)^2} = 6.3$	$= \sqrt{(\mathbf{9-5})^2 + (\mathbf{16-13})^2} = \mathbf{5}$
7	$= \sqrt{(\mathbf{5-7})^2 + (\mathbf{6-10})^2} = \mathbf{4.5}$	$= \sqrt{(5-5)^2 + (6-13)^2} = 7$
8	$= \sqrt{(9-7)^2 + (18-10)^2} = 8.3$	$= \sqrt{(\mathbf{9-5})^2 + (\mathbf{18-13})^2} = \mathbf{6.4}$
10	$= \sqrt{(8-7)^2 + (16-10)^2} = 6.1$	$= \sqrt{(\mathbf{8-5})^2 + (\mathbf{16-13})^2} = \mathbf{4.2}$

Therefore, the cluster structure after the first assignment iteration is:

$C_1 = \{2,7\},\ C_2 = \{9,1,3,4,5,6,8,10\}$

b) The choice of centroids does not reflect the true division of students, since both students 2 and 9 are not high-performing students. Students 1, 4, 8, and 10 should have formed their own cluster since on average they are high-performing students.

c) We measure the distance between the student and the centroids of the cluster. From the table in (a)

Distance to 2	$= \sqrt{(9-7)^2 + (17-10)^2} = 7.3$
Distance to 9	$= \sqrt{(9-5)^2 + (17-13)^2} = 5.7$

We find that the student should be assigned to cluster C_2.

Problem Set 3 – Association Rule Mining

Problem 5

Giving the following database with 5 transactions:

Transaction	List of Items
T1	A, B, C, D, E, F
T2	B, C, D, E, F, G
T3	A, D, E, H
T4	A, D, F, G, I, J
T5	B, D, E, K

Given **minimum support of 2/5** and **minimum confidence of 4/5**:

a) Apply the Apriori algorithm to the dataset of transactions and identify all frequent k-itemsets. Show all of your work. You must show candidates but can cross them off to show the ones that pass the minimum support threshold. If a candidate itemset is pruned because it violates the Apriori property, you must indicate that it fails for this reason and not just because it does not achieve the necessary support count. So, explicitly tag the itemsets that are pruned due to violation of the Apriori property.

b) Find all strong association rules that contain "A" in the left-hand part of the rule.

Answer

a) Support threshold =2/5 → 2 or more transactions

Applying Apriori:

Pass (k)	**Candidate k-itemsets and their support**	**Frequent k-itemsets**
k=1	A(3), B(3), C(2), D(5), E (4), F(3), G(2), ~~H(1), I(1), J(1), K(1)~~	A, B, C, D, E, F, G
k=2	~~{A, B}(1)~~, ~~{A, C}(1)~~, {A, D}(3), {A, E}(2), {A, F}(2), ~~{A, G}(1)~~ {B, C}(2), {B, D}(3), {B, E}(3), {B, F}(2), ~~{B, G}(1)~~ {C, D}(2), {C, E}(2), {C, F}(2), ~~{C, G}(1)~~ {D, E}(4), {D, F}(3), {D, G}(2) {E, F}(2), ~~{E, G}(1)~~ , {F, G}(2)	{A, D}, {A, E}, {A, F} {B, C}, {B, D}, {B, E}, {B, F}, {C, D}, {C, E}, {C, F}, {D, E}, {D, F}, {D, G}, {E, F}, {F, G}
k=3	{A, D, E}(2), {A, D, F}(2), ~~{A, E, F}(1)~~, {B, C, D}(2), {B, C, E}(2), {B, D, E}(3), {B, C, F}(2), {B, D, F}(2), {B, E, F}(2), {C, D,E}(2), {C, D, F}(2), {C, E,F}(2), {D,E,F}(2), ~~{D,E,G}(Apriori)~~, {D,F,G}(2)	{A, D, E}, {A, D, F}, {B, C, D}, {B, C, E}, {B, D, E}, {B, C, F}, {B, D, F}, {B, E, F}, {C, D,E}, {C, D, F}, {C, E,F}, {D,E,F}, {D,F,G}
k=4	~~{A, D, E, F}(Apriori)~~, {B, C, D, E}(2), {B, C, D, F}(2), {B, C, E, F}(2), {B, D, E, F}(2), {C, D, E, F}(2)	{B, C, D, E}, {B, C, D, F}, {B, C, E, F}, {B, D, E, F}, {C, D, E, F}
K=5	{B, C, D, E, F}(2)	{B, C, D, E, F} → #

b) No rules from the two last round would be produced since none contains A. From round k = 3, we have:

A → D, E ➔ confidence = 2/3 < 0.8 confidence threshold

A, D → E ➔ confidence = 2/3 < 0.8 confidence threshold

A, E → D ➔ confidence = 2/2 = 1 ➔ **strong rule**

A → D, F ➔ confidence = 2/3 < 0.8 confidence threshold

A, F → D ➔ confidence = 2/2 = 1 ➔ **strong rule**

A, D → F ➔ confidence = 2/3 < 0.8 confidence threshold

Problem Set 4 – Sentiment Analysis

Problem 6

a) Assume the following likelihoods for each word being part of a positive or negative movie review, and equal prior probabilities for each class.

	positive	negative
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

What class will Naive Bayes assign to the sentence "*I always like foreign films.*"?

b) Train binarized naive Bayes, with add-1 smoothing, on the following document **counts for key sentiment words**, with positive or negative class assigned as noted.

	"*good*"	"*poor*"	"*great*"	Class
Doc1	3	0	3	positive
Doc2	0	1	2	positive
Doc3	1	3	0	negative
Doc4	1	5	2	negative
Doc5	0	2	0	negative

Use binarized naive Bayes model to assign a class (positive or negative) to this sentence:

"*A good, good plot and great characters, but poor acting.*"

What is the final class to the sentence?

Answer

a)

P (I always like foreign films|+) = 0.09 × 0.07 × 0.29 × 0.04 × 0.08 = **0.0000058464**

P (I always like foreign films|-) = 0.16 × 0.06 × 0.06 × 0.15 × 0.11 = **0.000009504**

Negative sentiment will be chosen as product for negative sentiment will be larger.

b)

P**(+)** = $\frac{2}{5}$

P**(-)** = $\frac{3}{5}$

vocab = {good, great, poor},

|v| = 3

—

+ve class binarized

P(good|+) = $\frac{1+1}{4+3} = \frac{2}{7}$

P(poor|+) = $\frac{1+1}{4+3} = \frac{2}{7}$

P(great|+) = $\frac{2+1}{4+3} = \frac{3}{7}$

	Good	**Poor**	**Great**	
	1	0	1	
	0	1	1	
SUM	**1**	**1**	**2**	**4**

—

-ve class binarized

$$P(good|-) = \frac{2+1}{6+3} = \frac{3}{9}$$

$$P(poor|-) = \frac{3+1}{6+3} = \frac{4}{9}$$

$$P(great|-) = \frac{1+1}{6+3} = \frac{2}{9}$$

	Good	**Poor**	**Great**	
	1	1	0	
	1	1	1	
	0	1	0	
SUM	**2**	**3**	**1**	**6**

Testing (good)^2 poor great

+ve (binarized) $\Rightarrow \frac{2}{5} \times \frac{2}{7} \times \frac{2}{7} \times \frac{2}{7} \times \frac{3}{7} = \mathbf{0.0041}$

Testing (good)^2 poor great

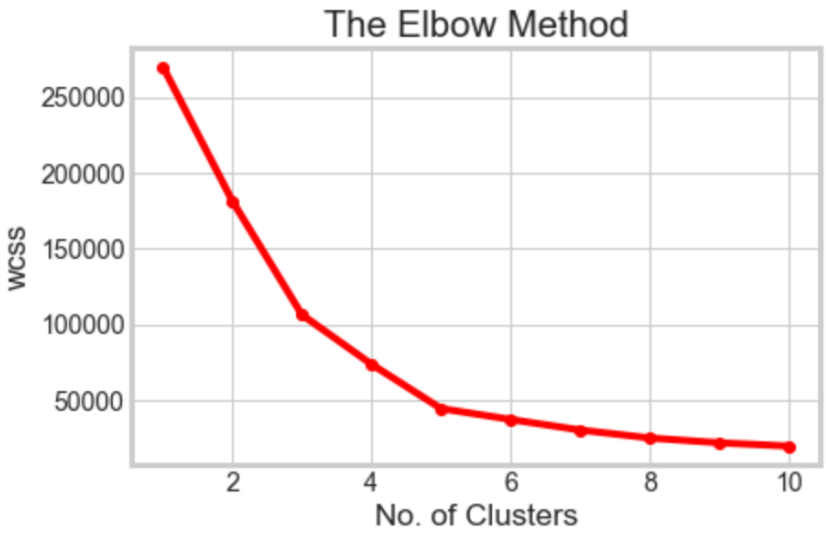
-ve (binarized) $\Rightarrow \frac{3}{5} \times \frac{3}{9} \times \frac{3}{9} \times \frac{4}{9} \times \frac{2}{9} = \mathbf{0.0064}$ → **greater**

Therefore, final class is Negative

Problem Set 5 – Python Questions Sample

Problem 1

Based on the below graph, how many clusters would you chose to run the K-means algorithm with? And Why (give detailed explanation)?



Answer

We will choose 5 clusters, as this is the optimum point which lowers the value of wcss and isn't a high number of clusters which can cause under fitting, as the goal is to minimize the wcss while maintaining relatively low number of clusters, the wcss is the within cluster sum of squares which is the average distances between the cluster's centroid and the points of this cluster.

Problem 2

Write in pseudocode (English): a code that visualizes the top 10 popular Music Genres in terms of the Popularity and the Genres.

Answer

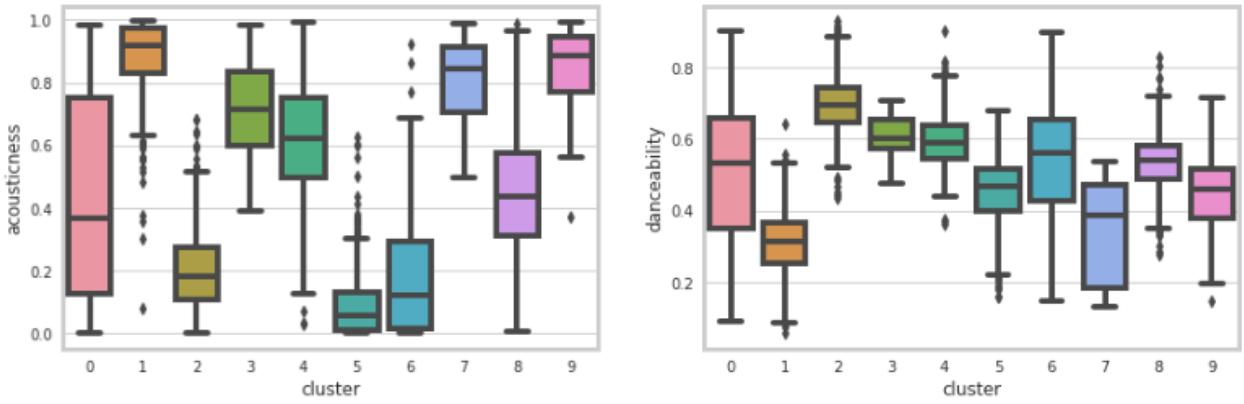
First sorting the data using the sort_values method in a descending order based on the popularity column and taking the top 10 rows using head(10) method then plotting the results in a bar plot with popularity on y axis and genres names on the x axes

OR

First we will get the top 10 rows with the largest popularity using nlargest(10, 'popularity') method. Then we will plot the results (the resulted data frame) in a bar plot with popularity on y axis and genres names on the x axes (*doesn't have to mention the method names or exact syntax. Also, any other correct steps will be accepted*).

Problem 3

Cluster Interpretation: Based on the below graph, give conclusions for Clusters 1 and 2 only.



Answer

Cluster 1 contains the genres which has the highest acousticness and lowest danceability.

Cluster 2 genres has relatively low acousticness and has the highest danceability of all clusters.

Formulas you may need:

Euclidean Distance:

$d(X, Y)=\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}$, where ***X*** and ***Y*** are two data instances and ***p*** are attributes.

Manhattan Distance:

$d(X, Y)= |x_1 - y_1^2| + |x_2 - y_2^2| + ... + |x_p - y_p|^2$, where ***X*** and ***Y*** are two data instances and ***p*** are attributes.

Decision Tree Classification:

$info\ (D) = entropy = -\sum_{x \in X} p(x)\ log_2\ (p\ (x))$

$info_{Attribute}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info\ (D_j)$, where ***j =1…v*** are the attribute values

Association Rules Metrics:

$support(A \Rightarrow B) = P(A \cup B) = \frac{n(A \cup B)}{N}$ $confidence\ (A \Rightarrow B) = P(B|\ A) = \frac{n(A \cup B)}{n(A)}$

Bayes' Theorem:

$P(Y|X) = \frac{P(X|Y)\ P(Y)}{P(X)}$

$P(X|\ Y) = \prod_{k=1}^{n} P(x_k|\ Y) = P(x_1|\ Y) \times P(x_2|\ Y) \times ... \times P(x_n|\ Y)$, where $\boldsymbol{x_k}$ is value of attribute ***k*** for object ***X***.

Naïve Bayes Rule for Sentiment Analysis:

$\hat{P}(w|\ c) = \frac{count(w,c)+1}{count(c)+|V|}$

where ***w*** is a word, ***c*** represents sentiment class, $|\boldsymbol{V}|$ represents vocabulary size in training set