Midterm Exam

COSC 6335: Data Mining

October 21, 2022

Your Name:

Your student id:

Problem 1 --- K-means/PAM and Clustering in General [26]

Problem 2 --- DBSCAN [11]

Problem 3 --- Outlier Detection [6]

Problem 4 --- Similarity Assessment [9]

Problem 5 --- Data Science Basics and EDA [20]

72:

**Grade:**



The exam is “open books and notes” and you have 90 minutes to complete the exam. The exam will count approx. 20% towards the course grade. The use of computers and cell phones is strictly prohibited.

1. **K-Means and K-Medoids/PAM and Clustering in General [26]**
2. Assume we apply K-medoids for k=2 to a dataset consisting of 4 objects numbered 1,..,4 with the following distance matrix:

0 6 5 2 🡨object1

0 4 3

0 1

0 (e.g. the distance between object 2 and 4 is 3)

The current set of representatives is {3,4} (objects 3 and 4); indicate all computations k-medoids (PAM) performs in its next iteration! Does k-medoids get a new set of representatives or does it terminate in the next iteration? [6]

RS={3,4} clusters: {3} (1,3,4} SSE=2\*\*2+3\*\*2

New Representative sets are created

{1,4} …. SSE=3\*\*2+1\*\*2

{2,4} {2} {1,3,4} SSE=2\*\*2+1\*\*2

{1,3} …SSE=3\*\*2+1\*\*2

{1,4} … SSE=3\*\*2+1\*\*2

The SSE decreased and therefore PAN will run for another iteration for the “new” representative set {2,4}

One error: at most 3.5 points; 2 errors at most 1 point.

b) Assume the following dataset is given: (1,1), (2,2) (4,4), (5,5), (4,6), (6,4) . K-Means is used with k=2 to cluster the dataset. Moreover, Manhattan distance is used as the distance function (formula below) to compute distances between centroids and objects in the dataset. Moreover, K-Means’s initial clusters C1 and C2 as follows:

C1: {(1,1), (3,3), (4,4), (6,6)}

C2: {(6,4), (4,6)}

Now K-means is run for a single iteration; what are the new clusters you obtain[[1]](#footnote-1) [4]

**d((x1,x2),(x1’,x2’))= |x1-x1’| + |x2-x2’| Manhattan Distance**

centroid C1= (3.5,3.5}

centroid C2= {5,5}

New Clusters

C1={(1,1), (3,3), (4,4)}

C2={(6,6},(4,6), (6,4)}

One error at most 1.5 points; 2 errors: 0 points

Problem 1 continued

c) What is the main difference between clustering and outlier detection? [2]

Clustering looks for groups of objects and outlier detection looks for objects which do not belong to any group.

d) Compare k-means with Hierarchical clustering; what are the main differences in the way they are forming clusters and in general? [4]

K-Means creates a single clustering and HC creates a hierarchy of object sets; that is, multiple clusterings [2]

HC creates a dendogram by merging the closest clusters[1]; K-means creates clusters by assigning the objects in a dataset to the closest centroid [1]

e) Compute the Silhouette for the point (2,3) for the following clustering that consists of 2 clusters:

{(0,0), (0,1), (2,3)}, {(3,3), (3,4)}; use Manhattan distance for distance computations. Interpret the result! [4]

Average Distance of (2,3) to the points in C2: (1+2)/2=1.5

Average Distance of (2,3) to points in its own cluster: (4+5)/2=4.5

s = (b – a) / max(a,b)

Silhouette((2,3))=(1.5-4.5)/4.5=-2/3

If any errors: at most 1 of 3 points for computing Silhouette((2,3)!

Interpretation: The silhouette coefficient for point (2,3) is very bad as the point should be in the other cluster [1]

f) Assume you apply K-Means for an extremely large dataset. What could be done to obtain a K-means clustering for this dataset more quickly? [6]

A “good” approach use sampling to obtain a much smaller dataset from the original dataset; cluster this dataset with K-means also storing the obtained centroids; next assign the remaining points to the closest centroid of clusters generated in the previous step.

Other ideas might deserve full or partial credit.

**2) DBSCAN [11]**

A dataset consisting of object A, B, C, D, E, F, G with the following distance matrix is given:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F | G | H |
| A | 0 | 3 | 3 | 8 | 9 | 10 | 8 | 6 |
| B |  | 0 | 3 | 7 | 8 | 9 | 8 | 7 |
| C |  |  | 0 | 6 | 6 | 6 | 3 | 6 |
| D |  |  |  | 0 | 14 | 15 | 7 | 7 |
| E |  |  |  |  | 0 | 4 | 2 | 6 |
| F |  |  |  |  |  | 0 | 4 | 7 |
| G |  |  |  |  |  |  | 0 | 6 |
| H |  |  |  |  |  |  |  | 0 |

1. Assume DBSCAN is run for this dataset with MINPOINTS[[2]](#footnote-2)=3 and epsilon=ε=5

How many clusters will DBSCAN return and how do they look like? Which objects are outliers and borderpoints in the clustering result obtained earlier? Give reason for your answers! [7]

Core points: A, B, C, E, F, G

Outliers: D, H

Border points: none [3]

Obtained clustering: {A, B, C, E, F, G} just one cluster [4]

Wrong cluster: only 0 or 1 points

b)How does DBSCAN form clusters? Limit your answer to 3-5 sentences! [4]

solution sketch : you iterate over the core points forming single clusters by obtaining points that are density reachable from the chosen core point and continue processing core points that do not belong any clusters yet until all core points belong to a cluster.

3) Outlier Detection [6] no answer given

a) What is an outlier? [2]

b) Describe in 4-6 sentences how distance-based outlier detection works! [4]

4) Similarity Assessment [9]

Design a distance function to assess the similarity of electricity company customers; each customer is characterized by the following attributes:

1. Ssn
2. Oph (“*on-time payment history*”) which is ordinal attribute with values ‘excellent’, “very good’, ‘good’, ‘medium’, and ‘poor’.
3. Power-used (which is a real number with mean 2000, standard deviation is 1000, its maximum is 10000 and minimum 100)
4. Country\_of\_Citizenship is a nominal attribute

Assume that the attributes Oph and Power-used are of major importance and the attribute Country\_of\_Citizenship is of a minor importance when assessing the similarity between customers. Using your distance function compute the distance between the following 2 customers: c1=(111111111, ‘excellent’, 2000, ‘Peru’) and c2=(222222222, ‘good’, 2500, ‘France’)!

Let ψ be the following function: ψ(excellent)=1, ψ(very good)=3/4, ψ(good)=1/2, ψ(medium)=1/4, ψ(poor)=1.

doph(a,b)= |ψ(a)- ψ(b)|

dp-u(a,b)= |(a-b)/1000|

dccit(a,b)= If a=b then 0 else 1

Let o1=(ssn1,oph1,p-u1,ccit1) and o2=(ssn2, oph2,p-u2,ccit2) be two customers:

d(o1,o2)= (doph (oph1,oph2) + dp-u(p-u1,p-u2) + 0.2\*dccit(ccit1, ccit2))/2.2)

Example: d(c1,c2)= (1/2+1/2+0.2)/2.2=1.2/2.2=0.545

Distance Function 7 points

Example: 2 points

If they do not define a mathematically sound object distance function: at most 2.5 points

Distance function definition: first error: -2 2 errors: 0-1 points for distance function/

5) EDA and Data Science Basics [20]

a) Assume we have a dataset with an attribute A with a mean value 8(μ=8) and standard deviation 4(σ=4). According to the 68–95–99.7 rule, what is the probability that a value of attribute A is between 0 and 16? [2]

95

No partial credit

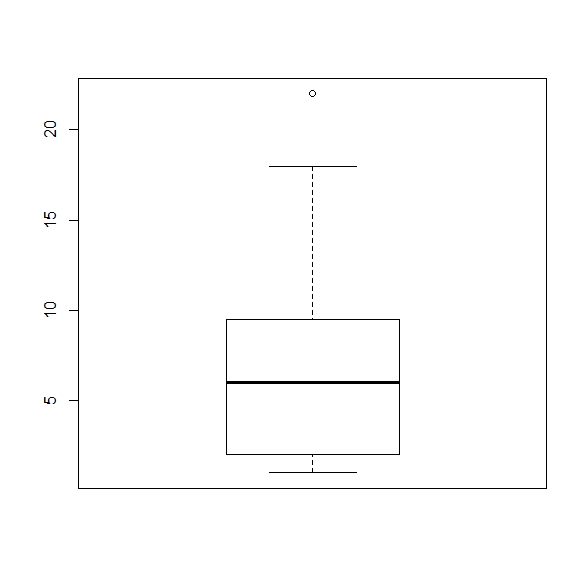
b) Assume you have a dataset with numerical attributes A and B which have a correlation of -0.012; what does this say about the relationship of attributes A and B? [2]

*no linear relationship; if they say ‘no relationship only 1 point”*

c) The following boxplot has been created using the following R-code for an attribute x:

> x<-c(1,2,2,2,4,4,8,9,9,10,18,22)

> boxplot(x)



What is the median for the attribute x? What is the IQR for the attribute x? The lower whisker of the boxplot as at 1; what does this tell you? According to the boxplot, 18 is not an outlier and 22 as an outlier; why do you believe this is the case? [6]

Median=6 [1]

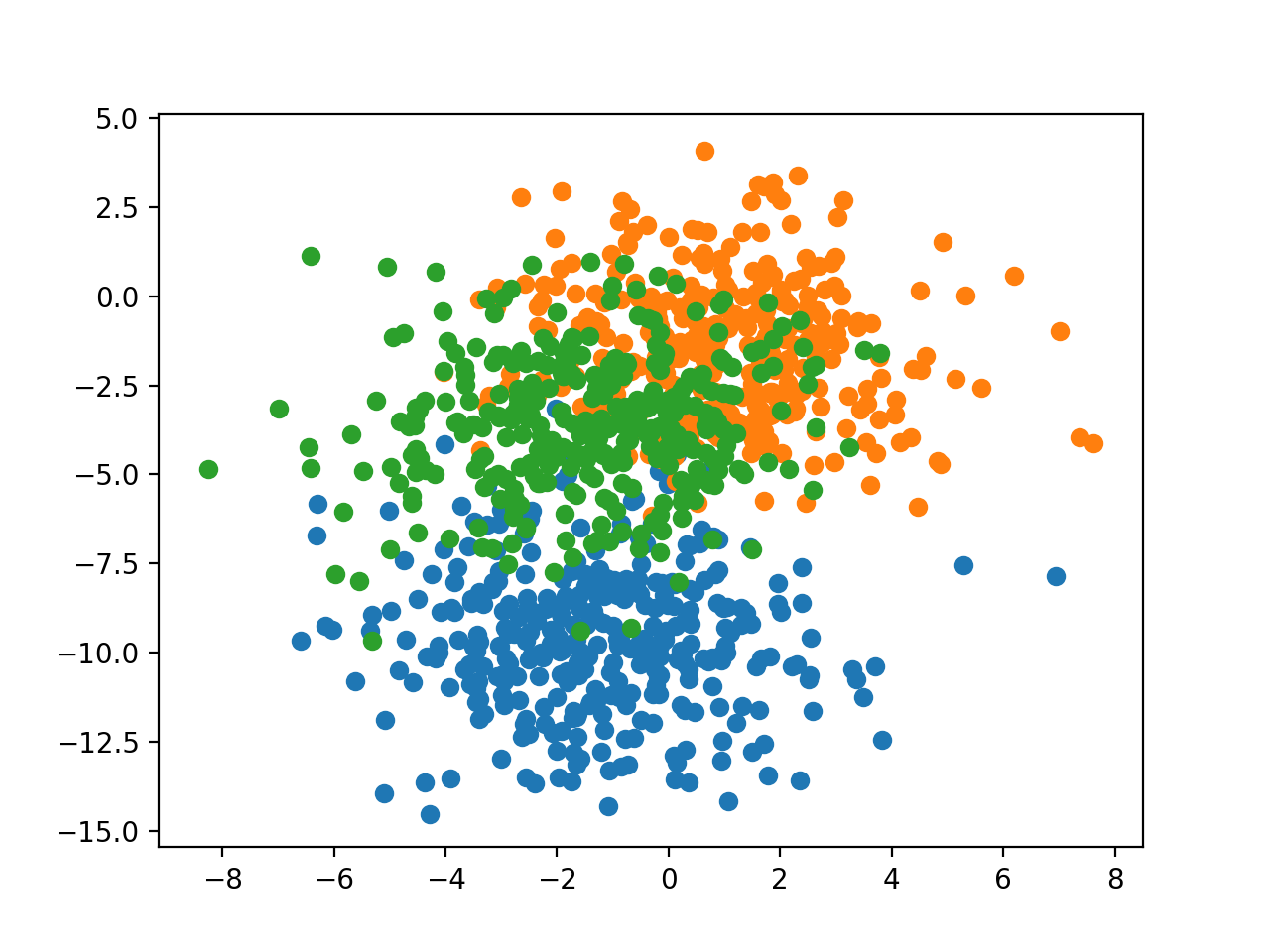
IQR=7.5 [1.5]

1 is the lowest value in the dataset that is not an outlier [1]

No partial credit for the first 3 answers!

9.5+1.5\*7.5=20.75 as 22 is above it is an outlier and as 18 is below it is the large point that is not an outlier. [2.5]

d) Interpret the supervised scatter plot depicting instances of classes in orange, blue and green color with their respective values of attributes named x and y; with attribute x taking values in [-8.4,+8.4] and attribute y taking values in [-15,+5]. Assume in your discussions that the three classes are called ‘orange’, ‘green’ and ‘blue’! Characterize the distribution of the instances of each class in the attribute space. Assess the difficulty of the classification problem of predicting classes 0, 1 and 2 using the attributes x and y! [10]

No answer given! 

1. If there are any ties, break them whatever way you want! [↑](#footnote-ref-1)
2. The object itself counts towards the number of objects in its ε-radius when determining core points! [↑](#footnote-ref-2)