Midterm Exam

COSC 6335: Data Mining

October 17, 2023

Your Name:

Your student id:

Problem 1 --- K-means/PAM and Clustering in General [19]

Problem 2 --- Density-based Clustering [13]

Problem 3 --- Outlier Detection [5]

Problem 4 --- Density Estimation [11]

Problem 5 --- Basic Statistics and EDA [20]

Problem 6 --- Decision Tree Induction Algorithm [8]

76:

**Grade:**



The exam is “open books and notes” and you have 80 minutes to complete the exam. The exam will count approx. 21% 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 [19]**
2. Assume we apply K-medoids for k=w to a dataset consisting of 5 objects numbered 1,..,4 with the following distance matrix:

0 6 5 2 7🡨object1

0 4 3 5

0 1 2

0 3

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

The current set of representatives is {3,4,5} (objects 3 and 4 and 5); indicate all computations k-medoids (PAM) performs in its next iteration! What clusters are formed and what is their SSE? Does k-medoids get a new set of representatives or does it terminate in the next iteration? [7]

Cluster formed {3} {1,2,4} {5} SSE=2\*\*2 + 3\*\*2=14

New representative Sets

{1,4,5} SSE=

{2,4,5} cluster {2]{1,3,4} {5} SSE= 1\*\*2 + 2\*\*2 best cluster

{1,3,5}

(2,3,5}

{1,3,4}

{2,3,4}

Algorithms runs for one more iteration using the representative set {2,4,5}

Can give partial credit; can give 2 extra points if student did all the computations and correctly.

b) How does K-means form clusters? [2]

By assigning the points in the dataset to the closest cluster centroid.

c) When does K-means terminate? [2]

When the clusterings no longer change

Problem 1 continued

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

e) Compute the Silhouette for the point (3,3) 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]

6+5+1/3=4

Sil ((3,3)= 4-1/4=0.75 [3] If error at most one of 3 points partial credit

Good Silhouette [1]

f) Characterize the shapes of clusters K-means can find! [2]

Convex Polygons!

**2) Density-base Clustering [13]**

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 | 4 | 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[[1]](#footnote-1)=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! [6]

Core points: A, B, C, E, F, G [2]

Outlier H: [1]

Border point: D [1]

Clusters Formed: {A,B,C,D,E,F,G} [2]

No partial credit for wrong answers!

1. Assume we run DBSCAN with MINPOINTS=5 and epsilon=ε=2.2 for a dataset D and we obtain 3 clusters and 5% outliers. Now we run DBSCAN with MINPOINTS=6 and epsilon=ε=2.2; how will the percentage of outliers and number of clusters change; give reasons for your answers. [4]

We get at least 5% outliers [2] Could also say we likely get less core points..

Number of clusters might go up or down[1]: Some clusters might disappear and other might be broken into multiple clusters. [1]

1. How does DENCLUE determine if two objects o and v belong to the same cluster or not? [3]

If o and v are hill climbed to the same density attractor/if o and v reside on the same hill, they will be in the same cluster

3) Outlier Detection [5]

Describe in 4-6 sentences how model-based outlier detection works! [5]

Fit a parametric model to the dataset [2] by using LDE or other methods[1]; use the model to determine the density and use this density as OLS.

4) Density Estimation [11]

a. How do parametric density estimation techniques find the parameters of models they try to fit to a dataset? For example, if we fit a Gaussian Model to a 1D dataset how does this approach choose the value μ and the standard deviation σ? [4]

Maximize the likelihood of the samples in the dataset D [2.5]

Give LDE formula [1.5]

b) Describe in a single natural language sentence without formulas how nonparametric density estimation computes the density for a query point x! [3]

Computes the density of x by adding up the influences of the points in the datasets on x.

c) Assume a dataset O={x1,x2,x3} with data points x1=(1,3), x2=(4,4), x3=(6,4), is given; moreover, assume Manhattan distance[[2]](#footnote-2) is used as the distance function and q1=(5,5) is a query point. Compute fGauss (q1) assuming bandwidth σ=1! [4]

Remark: it is okay to use an expanded formula as your answer; e.g “e12 + e2.5…” as your answer; it is not necessary to report the exact value!

fGauss ((5,5)) = e-36/2 + e-4/2  + e-4/2 = e-32 + 2\* e-2

at most 2.5 points of partial credit if single error!

5) Exploratory Data Analysis and Basic Statistics [20]

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

68% no partial credit

b) Assume you have a dataset with numerical attributes A and B which have a correlation of -0.98; what does this say about the relationship of attributes A and B? [2]

strong linear relationship; if A goes up B goes down and vice versa.

c) What are the characteristics of a good histogram? [2]

it provides a good approximation of the density distribution of the attribute.

d) Boxplots [6]

Assume a boxplot has been created using the following R-code for an attribute x, containing the indicated 12 values:

> x<-c (15,4,3,3,9,10,12,12,12,12,19,29)

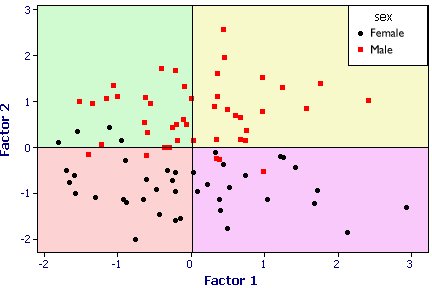
> boxplot(x)

What is the median for the attribute x?[1] What is the IQR for the attribute x?[1.5] What is the position of the upper whisker of the box plot created for attribute x?[1.5] Are there any outliers in the dataset[2] Assume that outliers are values which are 1.5 IQR above the upper box boundary or 1.5 IQR below the lower box boundary. Justify your answers for the last two questions!

Mahin add solution

*Problem 5 continued*

e) Interpret the supervised scatter plot, depicting instances of classes male and female with their respective values of attributes named Factor 1 and Factor 2. Characterize the distribution of the instances of each class in the attribute space. Assess the difficulty of the classification problem distinguishing males from females using the attributes Factor 1 and Factor 2! [8]



No specific solution given!

Should characterize the distribution of the 2 classes [3]

Should discuss decision boundaries [4]

Should clearly say that Factor 2 is much more useful in separating the classes than Factor 1 [2]

At most 8 points.

6. Decision Tree Induction Algorithm [8]

a. How does the decision tree induction algorithms select tests? [3]

Select the test which maximizes the information gain/GINI gain; the test which reduces the GINI/entropy before the split the most.

b. Compute the GINI-gain[[3]](#footnote-3) for the following decision tree split[[4]](#footnote-4) (give the formula and compute the actual value as well)! [5]

(5,3,2) (2,3,0)

(0,0,2)

(3,0,0)

G((0.5,0.3,0.2))- 0.5\*G(0.4,0.6,0)-0-0 [3] at most 1 point if error

Computation of Value [2] at most 1 point if error in computation.

1. The object itself counts towards the number of objects in its ε-radius when determining core points! [↑](#footnote-ref-1)
2. d((x1,y1),(x2,y2))= |x1-x2| + |y1-y2| [↑](#footnote-ref-2)
3. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-3)
4. There are 3 classes, and 10 examples are associated with that node, 5 of which belong to class1, 3 belong to class2 and 2 belong to class3; after the 3-way split the first node contains 5 examples 2 of which belong to class1 and 3 of which belong to class2. [↑](#footnote-ref-4)