Midterm2

COSC 3337 *Data Science I*

November 14, 2023

Your Name:

Your student id:

Problem 1 --- DBSCAN [12]

Problem 2 --- Miscellaneous Questions [17]

Problem 3 --- Clustering [16]

Problem 4 --- Density Estimation [11]

Problem 5 --- Neural Networks [12]

68:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 15% towards the course grade. You can use calculators but the use of computers and cell phones is strictly prohibited.

Write your answers on the exam paper; if you need more space use back of exam paper!

**1) DBSCAN [12]**

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

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F | G | H | I | J |
| A | 0 | 3 | 9 | 8 | 9 | 10 | 8 | 3 | 9 | 6 |
| B |  | 0 | 9 | 7 | 8 | 9 | 8 | 6 | 8 | 4 |
| C |  |  | 0 | 6 | 6 | 6 | 7 | 6 | 7 | 7 |
| D |  |  |  | 0 | 14 | 15 | 7 | 7 | 7 | 8 |
| E |  |  |  |  | 0 | 4 | 2 | 6 | 7 | 9 |
| F |  |  |  |  |  | 0 | 4 | 7 | 8 | 8 |
| G |  |  |  |  |  |  | 0 | 6 | 9 | 7 |
| H |  |  |  |  |  |  |  | 0 | 8 | 6 |
| I |  |  |  |  |  |  |  |  | 0 | 8 |
| J |  |  |  |  |  |  |  |  |  | 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 border points in the clustering result you obtained? Give reason for your answers! [7]

{A,B,H,J} {E,F,G} 4 points at most 1.5 points partial credit for incorrect solutions

Borderpoints: H and J 1.5 points; no partial credit

Outliers: C, D, I 1.5 points no partial credit

b)How does DBSCAN form clusters? [3]

Starting from an unprocessed core points it creates a new cluster and adds points that are density-reachable from this core point to the cluster[2], and continues this procedure for core points that have not been processed yet[1].

Other answers might deserve full or partial credit

c) What is a border point in DBSCAN? [2]

A point which is in the radius of a core point but is not a core point itself.

**2.** Miscellaneous Questions [17]

a. Describe the role and interactions of narrative, visuals and data in data storytelling! [4]

See slide of Data Storytelling lecture on this subject; they should give a short version of what is discussed on this slide!

b What are the characteristics of a “good data storyteller”? [4]

Essay-style questions; create a rubric based on your own preference and then grade using common sense.

c. How does model-based outlier detection work? Limit your answer to at most 4 sentences! [4]

Basic points they should mention: a. fit a model M to the dataset [2] b. using the model D compute the density of each point o in the dataset and use this density as the outlier score [2].

d. How does the APRIORI algorithm compute k+1 itemset candidates from frequent k itemsets[[2]](#footnote-2)? [3]

Find pairs of frequent K items which share the same k-1 prefix and combine the items to form a K+1 item candidates.

Other answers might deserve partial credit.

e. What is the goal of association analysis? [2]

Finding interesting patterns in a dataset [1] based on an intererestingness function. [1]

Other answers might deserve full credit!

**3. Clustering [16]**

a) Assume you cluster a 2-dimensional dataset with K-means. Can K-means obtain spatial clusters with the shape a) of an ellipse b) with the shape of a letter T? Give reasons for your answer! [4]

a. T is not a convex shape, cannot be found [2]

b. If they say ellipse is a convex but not a convex polygon cannot be found is correct; if they say ellipse can be quite precisely approximated and there K-means can obtain this shape, this is also correct.

b) Assume you run K-Medoids/PAM for the following dataset consisting of objects A,…,G whose distance function is given below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F | G |
| A | 0 | 1 | 3 | 4 | 19 | 10 | 12 |
| B |  | 0 | 6 | 5 | 8 | 9 | 11 |
| C |  |  | 0 | 6 | 6 | 6 | 2 |
| D |  |  |  | 0 | 14 | 15 | 9 |
| E |  |  |  |  | 0 | 4 | 8 |
| F |  |  |  |  |  | 0 | 7 |
| G |  |  |  |  |  |  | 0 |

We run PAM for k=2 and the current representative set is {B, C}. What clusters does PAM create from this representative set and what is the SSE of the obtained clustering? [4]

**{**B, A, D} [2]

**{**C, E, F, G} [2]

SSE= 1\*\*2+5\*\*2+6\*\*2+6\*\*2+2\*\*2

c) Assume you apply single link/min hierarchical clustering to the following dataset, consisting of objects A,…F whose distances are given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F |
| A | 0 | 1 | 3 | 4 | 19 | 10 |
| B |  | 0 | 6 | 5 | 8 | 9 |
| C |  |  | 0 | 2 | 7 | 6 |
| D |  |  |  | 0 | 14 | 11 |
| E |  |  |  |  | 0 | 12 |
| F |  |  |  |  |  | 0 |

What dendrogram is created? [4]

Should give a histogram

Join A and B; Join C and D; Join (A and B) and (C and D); join F with (ABCD); join E with ABCDF

One error at most 1.5 pomts

d) Compute the Silhouette for the point (4,4) for the following clustering which consists of 2 clusters: {(0,0), (0,1), (2,3)} and {(3,3), (4,4)}; use Manhattan distance for distance computations. Interpret the result! [4]

(((3+7+8)/3-2)/6=4/6=2/3 [3] one error at most 1 points

Good [1]

**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 mean value μ and the standard deviation σ? [4]

The idea is to select parameters; e.g. value μ and the standard deviation σ in the case of a Gaussian 1D- distribution---which maximizes the probability of the examples in D: Maximize the sample that is: ∏d∈D P(d|μ,σ)

where P is the density function of the distribution whose parameter need to be selected.

Other solution might deserve credit.

b) Assume a dataset O={x1,x2,x3} with data points x1=(1,2), x2=(5,7), x3=(7,7), is given; moreover, assume Manhattan distance[[3]](#footnote-3) is used as the distance function and q1=(6,6) 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 ((6,6)) = e-81/2 + e-4/2  + e-4/2= e-81/2+ 2\* e-2

Solutions which use the normalized 2D KDE function also deserve full credit.

c) If you change the bandwidth when using non-parametric density estimation, how does this impact the underlying density function? [3]

It changes how quickly/slowly the influence of a point in the dataset on a query points is reduced when the distance increased.

**5) Neural Networks [12]**

a) Neural networks employ gradient decent for weight learning. Describe in 3-4 sentences how this approach works! [4]

Things they should mention: a) start with a randomly chosen weight vectore, b) find the gradient for the current weight vector c) take a (small) step in the direction of the inverse[[4]](#footnote-4) gradient d) repeat steps b and c until some convergence is reached.

Problem 5 continued

b) The step size in neural network learning depends—among other factors—on the gradient of the error function[[5]](#footnote-5); explain why this is important! [4]

If the gradient is large you like to take large steps for faster convergence [2]; on the other hand, if you are close to a local minimum which is characterized by a low gradient you like to take small steps in order not to overshoot the local maximum. [2]

c) Take a look at the sub neural network consisting of nodes A, B, C, and D in the figure below; give a formula that computes the associated error ΔA for a node A. Assume the used activation function is g and its derivative is denoted by g’, and the activation of a node X is denoted by aX and the linear input of a node X is denoted by zX. First provide a general formula; then, replace general variables in the formula by their actual known values. [4]

wA,B=0.2

ΔB=0.4

A B

wC**,A**=1 wD,A=0.5

C D

ΔA depends on: the associated error ΔB in the node B, the weight of the connection between A and B and the derivative g’ of the activation function g for the linear input of node A zA.

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.2\*0.4=g’(0.4)\*0.2\*0.4=g’(0.4)\*0.2\*0.4

1. The object itself counts towards the number of objects in its ε-radius when determining core points! [↑](#footnote-ref-1)
2. E.g. how does it create 4 itemset candidates from frequent 3 itemsets? [↑](#footnote-ref-2)
3. d((x1,y1),(x2,y2))= |x1-x2| + |y1-y2| [↑](#footnote-ref-3)
4. The direction in which the error function decreases the most. [↑](#footnote-ref-4)
5. Also called loss or cost function in some textbooks. [↑](#footnote-ref-5)