Midterm Exam1 Solution Sketches

COSC 4335 *Data Mining*

October 1, 2018

Your Name:

Your student id:

Problem 1 --- K-means/PAM [14]

Problem 2 --- Data Mining in General [4]

Problem 3 --- DBSCAN [12]

Problem 4 --- Similarity Assessment [9]

Problem 5 --- Data Analysis [19]

:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 16% towards the course grade. The use of computers and cell phones is strictly prohibited.

1. **K-Means and K-Medoids/PAM [14]**
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 2 5 6 🡨object1

0 4 3

0 1

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

The current set of representatives is {1,2}; 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? [7]

Clusters: {1} and {2,3,4}

SSE Error = 25

k-medoids get a new set of representatives {1,3},{1,4},{2,3},{2,4} all of which lead to same clusters with an SSE error of 5 which are: {1.2} [3,4; this clustering is optimal. ]

not mentioned the computations k-medoids perform in first iteration: -2

b) Assume the following dataset is given: (1,1), (3,3), (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), (5,5)}

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**

New Clusters:

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

C2: {(5,5), (4,6), (6,4)}

Incorrect clustering results: -3 points; more than 2 errors: 0 points.

Problem 1 continued

c) Assume you have to run K-means or K-medoids/PAM to cluster a very large dataset and you want to use the “more efficient” algorithm. Which algorithm would you choose? Give reasons for your answer! [3]

K-means as its space complexity is O(n) in contrast to K-medoids whose space complexity is O(n2) when the distance matrix is stored. [3]

2) Data Mining in General [4]

Classification and clustering are important data mining tasks. What are the main differences between the two tasks?

Clustering vs Classification

1. Clustering: find similar groups of object vs Learn to classify examples with **respect to an a priori given class structure** [1]
2. Discover classes vs learn a model that predicts a class [1].
3. Unsupervised vs Supervised [1]
4. Dataset consists of attributes VS, consists of attributes and class labels [1]
5. Similarity assessment which derives a distance function is critical VS. Classifiers are learnt from set of classified examples (with classifiers) [1]

At most 4 points!

**3) DBSCAN [12]**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F |
| A | 0 | 1 | 2 | 4 | 6 | 7 |
| B |  | 0 | 3 | 8 | 9 | 10 |
| C |  |  | 0 | 11 | 12 | 13 |
| D |  |  |  | 0 | 14 | 15 |
| E |  |  |  |  | 0 | 16 |
| F |  |  |  |  |  | 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]

1 Cluster: {A,B,C,D} [4]

Other answers 1 point if close to correct solution; otherwise, 0! e.g. {A, B, C} gets 1 point

Outliers: E & F as they are not core or border points [1.5; one error 0.5]

Borderpoint: D as it is in the neighborhood of core point (A) but has less than 3 points in its ε-neighborhood [1.5]

b) What are the characteristics of a noise point in DBSCAN? [2]

A noise point is any point that is not a core point or a border point. [2]

c) How does DBSCAN form clusters? [3]

1. Select a random point P
2. If P is a core point, retrieve all points density reachable from P with Epsilon and Midpoints
3. If P is a core point, a cluster is formed
4. Visit the next point P that is not in a cluster yet; if there is no such point terminate!
5. Continue with step 2

Verbal descriptions how clusters are formed are also fine…

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’, ‘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’,2100, ‘France’)!

We convert the Oph rating values ‘excellent’, ‘good, ‘medium’, and ‘poor’ to 3:0 and then divide by the numbers by the number of values minus 1; that is by 3 in this case**;** finally we compute the distancesby applying the L-1 norm to the mapped values of the attribute Oph.

Normalize Power-used using Z-score and find distance by using L-1 norm

dCountry\_of\_Citizenship(a,b):= if a=b then 0 else 1

Assign weights 0.4 to Oph, 0.4 to Power-used and 0.2 to Country\_of\_Citizenship

Now:

**d(u,v) = 0.4\*|(u.PowerUsed – v.PowerUsed)/1000| + 0.4\*|φ(u.Oph) – φ(v.Oph)|/3 + 0.2\*dCountry\_of\_Citizenship(u.Country\_of\_Citizenship, v.Country\_of\_Citizenship)**

2 customers:

c1=(111111111, ‘excellent’, 2000, ‘Peru’) and c2=(222222222, ‘good’,2100, ‘France’)

= 0.4|(3 - 2 )/3| + 0.4\*|(2000 - 2100)/1000| + 0.2\*1= 0.133 + 0.04 + 0.2 = 0.373

If distance functions do not make much sense give 2 points or less

Distance functions are not defined properly [-5]

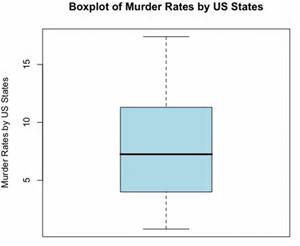
One error [-3]; two errors [-6]; more than 2 errors at most 1 point!

Only function is mentioned and no explanation: -4

Not calculated the distance between 2 customers: -2

5) Data Analysis [19]

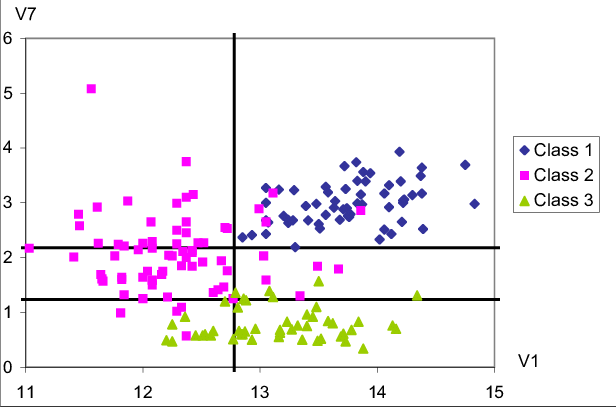
a) Assume the following boxplot is given. What does the box of the box plot tell you about murder rates in the US? The line splits the box into two uneven parts with the higher box being larger; what does this tell you? What do the whiskers in the boxplot indicate? [5]



The 75th percentile is 11 and the 25th percentile is 4; the standard deviation/IQR is 11-4=7 [1.5]; the distribution of the murder rate is skewed [1.5; other answers might deserve full or partial credit]; the largest value that is not an outlier is 17[[3]](#footnote-3) and the smallest value that is not on outlier is 1 [2].

Problem 5 continued

1. Interpret the supervised scatter plot depicted below that consists of instances of 3 classes; moreover, assess the difficulty of separating instances of the 3 classes using attributes V1 and V7 based on the scatter plot! [7]



The distribution of all three classes is unimodal [2] with no major gaps in the data density [0.5]. Attribute V7 is useful to separate the green class from the other 2 classes[1]; if V1 is less than 1.2 objects mostly belong to the green class[0.5]; Attribute V1 is useful to separate the purple class from the blue class[1]; all the example whose V1 value is above 12.8 are blue[0.5]; additionally using attribute V7 requiring that instances whose attribute values for V1 that lie in [1.2,2.2] leads to a even clearer distinction between the two classes [0.5]. Moreover, using the fact that the V1 value is a higher than 12.2 for instances of the green class, leads to a clearer separation of the green and purple class [0.5] By combining those rules [0.5], the classification task should not be too difficult as the examples are well separated although there are a few exceptions.[1]

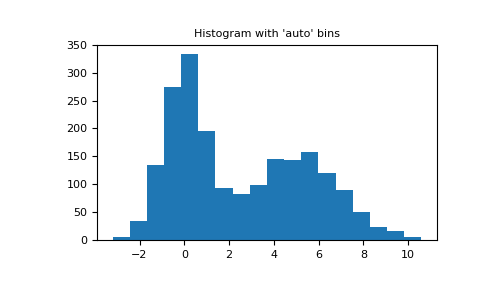
If they describe where the instances of the red, green, and purple are concentrated with respect to the V1/V7 coordinate system: [0.5] for each description of the location of the three classes

However, the can only get 7 points total for whatever they say, but you are allowed to give up to 8 points for “*exceptionally good*” solutions.

c) What does the fact that two attributes A and B have a correlation of -0.02 tell you about the relationship of the two attributes? [2]

There is not linear relation between the two attributes [2; if the do not mention ‘linear’ at most 0.5 points]

d) Interpret the following histogram that gives counts of instances of a class C for the values of an attribute A whose values range between -3.5 and 10.4. [5]



The class distribution is bimodal [1.5] with peaks near 5.8 [0.5] and 0 [0.5], but the first hill is significantly higher [0.5] more than twice as high [0.5] as the second hill; the density between 1.4 and 7.4 does not vary much[1]; there are no gaps[1];.

Other observation might deserve credit. At most 5 points and no extra credit!

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)
3. Saying 18 also deserves full credit [↑](#footnote-ref-3)