Final Exam

COSC 3337 Data Science I

December 6, 2019

Your Name:

Your student id:

Problem 1 --- K-means/PAM and Clustering in General [24]

Problem 2 --- DBSCAN [11]

Problem 3 --- Preprocessing [6]

Problem 4 --- Data Science [10]

Problem 5 --- Classification [17]

:

**Grade:**



The exam is “open books” and you have 115 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 [24]**
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 {2,3} (objects 2 and 3); 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]

A cluster {1,3, 4} {2} is formed with SSE of 5\*\*2+1\*\*2

Next new representative sets {1,3}, {4,3}, {1,2}, and {4,2} are created.

Clusters are formed for those and it turns out that the representative set {4,2} associated with the clustering {1,3,4} and {2} has the lowest SSE of 1\*\*2+2\*\*2=5.

As the new SSE is lower, the algorithm continues for one more iteration.

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), (2,2), (4,4), (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**

centroid C1: (3,3)

centroid C2: (5,5)

New clusters are either C1={(1,1), (2,2), (4,4}} and C2={(5,5), (4,6), (6,4)}

or C1={(1,1), (2,2)}and C2={(4,4},(5,5), (4,6), (6,4)} as (4,4) has the same distance to the two centroids and can therefore be assigned to either cluster.

At most 1 point if they do not have one of the two correct clustering results!

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! [2]

K-means[1] K-mean’s has O(n) complexity instead of PAM’s O(n\*\*2) complexity where n is the number of objects to be clustered[2].

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

Clustering centers on identifying different groups in a dataset whereas outlier detections centers on finding objects that do not belong to any group.

e) Compare k-means with Hierarchical clustering; what are the main differences between the two clustering approaches? [4]

Hierarchical clustering computes a set of clusterings [1] that are organized as a hierarchy[1] where K-means returns a single clustering [1]; HC forms clusters by merging or splitting whereas k-means relies on cluster representatives and form clusters by assigning objects to the closest representative [2]

At most 4 points.

f) One approach for assessing cluster validity uses the correlation between the distance matrix and actual clusters obtained. Explain the key idea of this approach! [3]

The idea is that if objects that belong to the same clusters have small distances there will be correlation with a high absolute value, indicating a “good” clustering, whereas correlations close 0 indicate bad clusters.

g) Give an example of an external index of cluster validity; what do external indices for cluster validity measure? [3]

purity or entropy or Gini [1, naming just one is enough]; measure how well the obtained clustering agrees with externally given classes.

**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 |
| A | 0 | 3 | 3 | 4 | 9 | 10 | 8 |
| B |  | 0 | 3 | 7 | 8 | 9 | 8 |
| C |  |  | 0 | 6 | 12 | 13 | 9 |
| D |  |  |  | 0 | 14 | 15 | 7 |
| E |  |  |  |  | 0 | 4 | 2 |
| F |  |  |  |  |  | 0 | 4 |

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]

Answer: {A, B, C, D} {F, F, G} [5]

Incorrect clustering result; 2 points if very similar to the correct clustering; otherwise, 1 or 0.

There are no noise points/outlier but D is a border point [2] no partial credit for wrong answer.

b)Assume we have two core points o and v that are within each other’s radius—will o and v belong to the same cluster? Now assume, that we have a border point b within a radius of a core point u—will b and u always belong to the same cluster? Give reasons for you answer! [4]

Yes, o is density reachable from v and v is density reachable from o. [1.5]

No if b is also in the radius of another corepoint w, and w is not density reachable from u; in this case b might end up in the cluster formed by growing around w, if we is processed before u. [2.5]

3) Preprocessing [6]

Dimensionality reduction is quite important in many data mining and data science projects. Why do you believe this is the case?

Alleviate the curse of dimensionality [1]: When dimensionality increases, data becomes increasingly sparse in the space that it occupies and definitions of density and distance between points, become less meaningful, and the respective algorithms do no longer work well and the only hope is to reduce the dimensionality of the dataset [3]

Reduce amount of time and memory required by data mining algorithms [1]

Allow data to be more easily visualized [1]

May reduce noise [1]

May reduce irrelevant [0.5] and redundant [0.5] features

At most 6 points!

4) Data Science and Data Storytelling [10]

This is an essay-style question; please use complete sentences when discussing the topic below. Limit you answer to 7-12 sentences!

Assume you are interested to get hired for a well paid Data Science job in industry. Write a short essay discussing what skill set is needed to accomplish this goal!

* Should know R and/or Phyton
* Should have sound software development skills
* Should have some sound knowledge of Statistics
* Should have sound knowledge of the different data analysis tasks; e.g. clustering, classification, similarity assessment
* Should be knowledgeable in data visualization
* Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”
* The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades."
* But what’s even harder is finding people who have those skills *and* are good at communicating the story behind the data.”
* Good team players
* ….

Kind of an essay style question; use your own discretion to grade it!.

5) Classification [17]

a) What does it mean, if an attribute is irrelevant for a classification task? [2]

The attribute is not helping in the classification; the distribution of the attribute does not provide any clues which class should be assigned to a particular example.

b) Ensemble approaches, such as random forests, have been quite successful in obtaining very high accuracies for classification tasks. What is the main challenge of ensemble learning? [2]

To obtain a set of base classifiers that make different kind of errors / are “mostly” independent of each other.

c) Neural Networks use gradient decent to find the “best” set of weights. Please, explain! Limit your answer to 4-6 sentences! [4]

a. starts with a “random” weight vector

b. the gradient of the error function is determined for the current weight vector to determine the direction the reduces the error the most

c. a new weight vector is obtained by taking a small step in the direction chosen in step c from the current weight vector

d. Steps b and c are repeated until some termination condition is reached.

d) Deep learning approaches are quite popular these days; what properties of deep learning approaches contribute to this success? [3+up to 2 extra points}

Can grade this leniently; e.g. things they could mention include:

a. capable to learn new features in their intermediate nodes that are useful for the particular prediction task [3]

b. quite generic approach [1.5]

c, obtain high accuracies [1.5]

d. …

…

e) Assume you have a classification problem involving 2 classes C1 and C2. Given a training example (‘tall’, ‘obese’, ‘high’) giving attribute values for attributes height, body-weight, and blood pressure: how does a naïve Bayesian classifier determine the class label of the example (‘tall’, ‘obese’, ‘high’)? [6]

Compute P(C1|height=tall,body-weight=obese,blood-pressure=high) and

P(C2|height=tall,body-weight=obese,blood-pressure=high) …(give details the 2 probabilities are computed)

Choose the class with the higher probability.

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)