Christoph F. Eick

**Review COSC 6335 Fall 2023 for Oct. 10 Midterm Exam**

October 8, 2024

**1) Clustering [17]**

1. What objective function does K-means minimize[[1]](#footnote-1)? [2]

**The sum of the squared distance of the objects in the dataset to the centroid of the cluster they are assigned to**

1. When does K-means terminate? When does PAM/K-medoids terminate? [2]

**When the clustering does not change; when there is no improvement with respect the objective function PAM minimizes.**

1. Assume K-Means is used with k=3 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’ initial clusters C1, C2, and C3 are as follows:

C1: {(2,2), (6,6)}

C2: {(4,6), (8,0)}

C3: {(4,8), (6, 8)}

}

Now K-means is run for a single iteration; what are the new clusters and what are their centroids? [3]

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

**C1 centroid: (4,4) {(2,2), (4,6)} new centroid: (3,4)**

**C2 centroid: (6,3) {(6,6), (8,0)} new centroid: (7,3)**

**C3 centroid: (5,8) {(4,8), (6,8)} centroid: (5,8)**

**Remark: Assigning (6,6) to cluster C3 instead, is also correct!**

d) Compute the Silhouette for the following clustering that consists of 2 clusters:

{(0,0), (0.1), (1,1)}, {(1,2), (4,4)}; use Manhattan distance for distance computations. Compute each point’s silhouette; interpret the results (what do they say about the clustering of the 5 points; the overall clustering?)![6]

(5.5-1.5)/5.5

(4.5-1)/4.5

(3.5-1.5)/3.5

(2-5)/5

(5-7)/7

In general, the silhouette of the first 3 points is good, the silhouette of the 4th point is bad, because this point has been associated with the wrong clustering, and the silhouette for the 5th points is mediocre because the inter-cluster distance is high due to the incorrect assignment of the point (1.2). The quality of the first cluster is decent, whereas the quality of the second cluster and the overall clustering is poor!

e) Assume you apply k-means to a dataset which contains outliers; assume you apply k-means to a 2D-dataset Y={(-100,-100), (0,0), (1,1), (0, 1), (1, 2), (5, 4), (5,5) , (5,6)} with k=2; how do outliers (e.g. the point (-100,-100)) impact the k-means clustering result? Propose an approach that alleviates the strong influence of outliers on clustering results. If your approach would be used to cluster dataset Y; how would the result change? [6]

Leads to a clustering where the outlier forms a single cluster [2]

Method [2.5]

After method is applied points 2-5 form a cluster, and the last 3 objects another cluster [1.5]

2) Outlier Detection

a) Give a brief description of how model-based approaches for outlier detection work.

Fit a statistical model M to the data points of the dataset O; next, the density function dM of the model M is used to assess the likelihood of objects o belonging to O; objects with very values for dM(o) or log(dM(o)) are considered to be outliers in O

b) How do k-nearest neighbor-based outlier detection techniques determine the degree to which “*an object in a dataset is believed to be an outlier*”.

For each object the k-nearest neighbor distance—k is a parameter of the method;—to the other objects in the dataset is computed; objects with very high values for that distance are considered to be outliers

3) Boxplots

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

> x<-c (15,4,2,2,8,8,12,12,12,12,26,29)

> boxplot(x)

What is the median for the attribute x? What is the IQR for the attribute x? What is the position of the lower whisker of the box plot created for attribute x? Are there any low outliers? Assume that outliers are values which are 1.5 IQR above the upper box boundary or 1.5 IQR below the lower box boundary. [5]

Median=10 [1]

IQR=12-2=10 [1.5]

-4 [1.5]

Yes -15 [1]

4. All kind of questions

a) Assume we have a dataset with an attribute A with a mean value 8(μ=8) and standard deviation 2(σ=2). According to the 68–95–99.7 rule, what is the probability that a value of attribute A is between 4 and 12? [2]

95%

No partial credit

d) In Data Science raw data sets are frequently z-scored before applying a particular analysis technique to them; what is the motivation for doing that? [2]

*To make different attributes important or to alleviate the fact that attribute differ in their scale.*

**5. EM**

a) What kind of cluster models does EM use?

For answer see: [Understanding Gaussian Mixture Models: A Comprehensive Guide | by Juan C Olamendy | Medium](https://medium.com/@juanc.olamendy/understanding-gaussian-mixture-models-a-comprehensive-guide-df30af59ced7#:~:text=Gaussian%20Mixture%20Models%20(GMMs)%20play%20a%20pivotal%20role%20in%20achieving)

b) How does EM form clusters? How does a clustering result of EM look like?

The EM algorithms starts with a random mixture of Gaussians model; it then interleaves updating the cluster memberships of the points in the dataset (E-Step) with updating the Gaussian mixture model (M-Step) until there is no or a small amount of change.

It returns a fuzzy clustering, giving a probability for each dataset point belonging to the i-th cluster (i=1,…,k).

**6. DBSCAN**

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 |

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]

Core points: A, B, C are core points

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

7) Density Estimation

How does non-parametric density estimation compute the density of a query point?

The density for a query point is computed by summing up the influences of the points in the dataset to the query point⎯the influence of a point to the query point decreases as the dataset point’s distance to the query point increases.

Remark: Topics covered in GHC presentations are relevant for the Midterm Exam. The following topis are relevant for the midterm exam although not covered in this review:

Interpretation of Supervised and unsupervised scatter plots; comparison of box plots; parametric density estimation including what MLE does; apply kernel density estimation for an example; questions about the data storytelling article.

1. Be clear! [↑](#footnote-ref-1)
2. The object itself counts towards the number of objects in its ε-radius when determining core points! [↑](#footnote-ref-2)