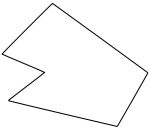
**Solution Sketches December 7, 2021 Review for COSC 3337 Final Exam**

1. What of the following cluster shapes K-means is capable to discover? a) triangles b) clusters inside clusters c) the letter ‘T ‘d) any polygon of 5 points e) the letter ’I’

In general, the shapes k-means can discover are convex polygons; consequently, it can only discover triangles and clusters if the shape of the depicted letter I; it will not be able to discover shapes of concave polygons of 5 points!

[](https://en.wikipedia.org/wiki/File:Simple_polygon.svg)concave polygon

1. What are the characteristics of clusters K-Medoids/K-means are trying to find? What can be said about the optimality of the clusters they find? Both algorithms a sensitive to initialization; explain why this is the case!

**Looking for: compact clusters [1] which minimize the MSE/SSE fitness function[1]**

**Suboptimal, local minima of the fitness function [1]**

**Employ hill climbing procedures which climb up the hill the initial solution belongs to; therefore, using different seeds which are on the foot of different hills (valleys) will lead to obtaining different solutions which usually differ in quality [2].**

**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[[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! [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]

b) How does DBSCAN form clusters? [3]

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

Verbal descriptions how clusters are formed are also fine…

4. PAM/K-medoids

Assume we apply K-medoids for k=3 to a dataset consisting of 5 objects numbered 1,..5 with the following distance matrix:

Distance Matrix:

0 2 4 5 1 🡨object1

0 2 3 3

0 1 5

0 2

0

The current set of representatives is {1,3,4}; indicate all computations k-medoids (PAM)

performs in its next iteration!

*The following cluster is formed: {1,2,5} {3} {4} or {1,5} {2,3} {4} as object 2 has the same distance of 2 to representatives 1 and 3. Let us assume {1,5} {2,3} 5 is selected as the current cluster its SSE is:*

*1\*\*2+2\*\*2=5; in the next iteration six clusters for representative sets {2,3,4}, {5,3,4},{1,2,4},{1,5,4},{1,3,2},{1,3,5} and the cluster with the lowest SSE is selected, which is*

*{1,5} {2} {4,3} which orginated from the representative set {1,2,4}[[2]](#footnote-2); it has a SSE of 1\*\*+1\*\*2=2 and as it is better than the cluster of the previous iteration it becomes the new current cluster and the algorithm continues for at least one more iteration.*

**5. Hierarchical Clustering [8]**

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E |
| A | 0 | 9 | 2 | 3 | 1 |
| B |  | 0 | 8 | 6 | 5 |
| C |  |  | 0 | 7 | 10 |
| D |  |  |  | 0 | 4 |
| E |  |  |  |  | 0 |

a) Assume single[[3]](#footnote-3) link hierarchical clustering is applied to the dataset? What dendrogram will be returned? [4]

NO ANSWER GIVEN; SEE GROUP HOMEWORK CREDIT TASK for a similar problem with solutions.

b) Hierarchical clustering computes dendrograms; what is the dendogram? What is the value of creating dendrograms—what can they be used for? [4]

*A* ***dendrogram*** *is a* [*tree*](https://en.wikipedia.org/wiki/Tree_(graph_theory)) *diagram frequently used to illustrate the arrangement of the clusters produced by* [*hierarchical clustering*](https://en.wikipedia.org/wiki/Hierarchical_clustering)*. Edges of the dendrogram represent split/merge relationships between the nodes of the tree which represent clusters[2]*

Dendograms organize dataset hierarchically---identifying homogeneous groups at differerent levels of granularity---which is important in bio-informatics and social sciences as it allows for discovering meaningful categories/classes in data.[2]

**6) Clustering in General**

a) Compare k-means with Hierarchical clustering; what are the main differences between the two clustering approaches? [5]

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] K-means uses iterations which modify the existing clustering whereas HC directly computes the dendrogram by merging nearest clusters.

At most 4 points.

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

no answer given!.

c) 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.

d) Be able to compute silhouette coefficient

See recent group homework credit specification and lecture slides how this is done.

**7) Association Rule Mining**

a) What is the anti-monotonicity property, also called APRIORI property, of frequent itemsets? How does the APRIORI algorithm take advantage of this property to create frequent itemsets efficiently? [4]

A⊆B 🡺 support(B)≤support(A) [1.5]

When creating k+1-itemsets from k itemsets [1.5]

For subset pruning of item-set candidates [1]

b) Assume the APRIORI algorithm identified the following six 4-item sets that satisfy a user given support threshold: **abcd, abce, abcf, abde, acde bcdf;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

abcde

abcdf

abcef all three correct [3]; one error 1 points

none survives subset pruning; [1] other answers 0

c) Assume an association rule if drink\_wine then smoke has a confidence of 0.7 and a lift of 1. What does this tell you about the association between drinking wine and smoking? Do you believe this rule is useful in predicting the likelihood of a person being a smoker? [4]

The probability of a wine drinker to smoke is 0.7. [1]

The probability of smoking as well as the probability of a wine drinker smoking are both 0.7; P(Smoke|DrinkWine)/P(Smoke)=1 [1.5}

No[0.5]; drinking wine has not impact on smoking; it does not change the likelihood of somebody smoking [1]

**7) Anomaly/Outlier Detection**

a) What is the goal of outlier/anomaly detection? [2]

The goal is to identify objects in a dataset that significantly differ from the other objects in the dataset.

b) Propose and describe an approach *that uses k-nearest neighbor distances for outlier*/anomaly detection! [5]

No answer given! Covered by group credit presentation

c) What observations are considered as outliers by model-based statistical outlier detection approaches? [2]

Observations whose density is very low / observations that are very unlikely with respect to the statistical model that was fitted to the dataset.

**8) Expect one more essay-style question in the final exam**

**Important: this is an essay: write complete sentences!**

**e.g What skills are important to be hired as a Data Scientist?**

(see slides that discuss this topic)

* 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.”

9) Similarity Assessment

Design a distance function to assess the similarity of graduate students; each graduate student is characterized by the following attributes[7.5]:

1. Ssn
2. qud (“*quality of undergraduate degree*”) which is ordinal attribute with values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’.
3. gpa (which is a real number with mean 2.7 standard deviation is 0.5, and maximum 4.0 and minimum 1.8)
4. gender is an nominal attribute taking values in {male, female}.

Assume that the attributes qud and gpa are of major importance and the attribute gender is of a minor importance when assessing the similarity between students. Using your distance function compute the distance between the following 2 students: c1=(111111111, ‘good’, 2.7, male) and c2=(222222222, ‘poor’, 3.7, female)[2.5]!

We convert the qud values ‘excellent’, ‘very good’, ‘good, ‘fair’, and ‘poor’ using a mapping function **φ** to 4:0 and then divide number of values the ordinal attribute takes 5 minus 1; that is, by 4 in this case**;** finally we compute the distancesby applying the L-1 norm to the mapped values of the attribute qud.

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

dgender(a,b):= if a=b then 0 else 1

Assign weights 1 to qud, 1 to gpa and 0.2 to gender attributes. We obtain:

**d(u,v) = (1\*|(u.gpa – v.gpa)/0.5| + 1\*|φ(u.qud) – φ(v.dud)|/4 + 0.2\*dgender(u.gender, v.gender))/2.2**

For 2 students c1=(111111111, ‘good’, 2.7, male) and c2=(222222222, ‘poor’, 3.7, female) we obtain for their distance.

d(c1,c2)= (2+2/4+0.2)/2.2=2.7/2.2≈1.2

10) Data Storytelling

Expect 1-2 kind of basic questions concerning what was covered in today’s lecture or be prepared for an essay style question…

Reading material: [Data Storytelling: The Essential Data Science Skill Everyone Needs (forbes.com)](https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/?sh=146220b252ad)

1. The object itself counts towards the number of objects in its ε-radius when determining core points! [↑](#footnote-ref-1)
2. The one that originates from {1, 3, 2} is identical and therefore equally good. [↑](#footnote-ref-2)
3. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-3)