Christoph F. Eick

Review1 COSC 4335 Spring 2016[[1]](#footnote-1)

1. What is the main difference between ordinal and a nominal attributes?
2. What role does exploratory data analysis play in a data mining project?
3. What does the size of the box of a boxplot measure; what statistical measure is it related to?
4. A R-boxplot (also called Turkey boxplots) of an attribute A has whiskers at 2 and 10; what does this tell you about attribute A? What attribute values are typically considered to be outliers in boxplots?
5. Assume we have a dataset in which the median of the first attribute is twice as large as the mean of the first attribute? What does this tell you about the distribution of the first attribute?
6. What is (are) the characteristic(s) of a good histogram (for an attribute)?
7. Interpret the following 2 histograms and their relationships which describe the male and female age distribution in the US, based on Census Data.



1. Assume you find out that two attributes have a correlation of 0.02; what does this tell you about the relationship of the two attributes? Answer the same question assuming the correlation is -0.98!
2. 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’
3. 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!
4. K-means is probably the most popular clustering algorithm; why do you believe is this the case?
5. Assume the following dataset is given: (2,2), (4,4), (5,5), (6,6), (8,8),(9,9), (0,4), (4,0) . K-Means is used with k=4 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, C2, C3, and C4 are as follows:

C1: {(2,2), (4,4), (6,6)}

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

C3: {(5,5), (9,9)}

C4: {(8,8}}

Now K-means is run for a single iteration; what are the new clusters and what are their centroids?[[2]](#footnote-2) [5]

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

1. 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!

14. What are the characteristics of a border point in DBSCAN?

15. How does DBSCAN form clusters?

16. Assume I run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and I obtain 4 clusters and 5% of the objects in the dataset are classified as outliers. Now I run DBSCAN with MinPoints=8 and epsilon=0.1. How do expect the clustering results to change?

17. Hierarchical Clustering algorithm creates dendrograms; what is a dendogram? How are the clustering results K-means creates differerent from those of hierarchichal clustering algorithms?

Tutorial Exercises(see <http://www2.cs.uh.edu/~ceick/UDM/Clustering_Exercises1.pdf>

18. Tutorial Exercise 3

19. Tutorial Exercise 5

1. Some of the problems will be discussed in the lecture on March 1, 2016. [↑](#footnote-ref-1)
2. If there are any ties, break them whatever way you want! [↑](#footnote-ref-2)