Christoph F. Eick

Review1 COSC 4335 Fall 2018[[1]](#footnote-1)

September 24 and 26, 2018

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?

*The difference between the 25thand 75th quantile, also called IQR, of the attribute; the size of the box is used as an estimator of the standard deviation of the attribute.*

1. 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?
2. Interpret the supervised scatter plot depicted below; moreover, assess the difficulty of separating males from females using Factor 1 / Factor 2 based on the scatter plot! [5]



1. What is (are) the characteristic(s) of a good histogram (for an attribute)?

*It captures the most important characteristics of the underlying density function*

1. 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!

*0.02:= no linear relationships exists between the two attributes—but other relationships might exist; 0.98:=a strong linear relationship exists—if the value of one attribute goes up the value of the other goes down*

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’

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!
2. K-means is probably the most popular clustering algorithm; why do you believe is this the case?

12. 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**

13. 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

14. DBSCAN

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

**b)** Assume you run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and I obtain 4 clusters and 2% of the objects in the dataset are classified as outliers/noise points. Now you run DBSCAN with MinPoints=8 and epsilon=0.1. How do expect the clustering results to change? [4]

c) How does DBSCAN form clusters?

15. Similarity Assessment

Design a distance function to assess the similarity of gradute students; each customer is characterized by the following attributes:

1. Ssn
2. qud (“*quality of undergraduate degree*”) which is ordinal attribute with values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’, ‘very poor’.
3. gpa (which is a real number with mean 2.8 standard deviation is 0.8, and maximum 4.0 and minimum 2.1)
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’, 2.9, male) and c2=(222222222, ‘very poor’, 3.7, female)!

1. A subset of the problems 1-8 and 15 will be discussed on September 24, and a subset of problems 9-14 will be discussed on the lecture on September 26. [↑](#footnote-ref-1)
2. If there are any ties, break them whatever way you want! [↑](#footnote-ref-2)