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?

*The values of ordinal attributes are ordered; this fact has to be considered when assessing similarity between two attribute values!*

1. What role does exploratory data analysis play in a data mining project?

**create background knowledge about the dataset and the task at hand [1], assess difficulty [1], provide knowledge to help select appropriate tools for the task[1], assess quality of data [1], validate data [1], help to form hypothesis [1], find issues, patterns and errors in data [1]**

1. 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?

*The largest attribute value that is not an outlier is 10, and the smallest attribute value that is not an outlier is 2; all attribute values that are 1.5IQR or more above the 25% quantile or 1.5IQR below the the 75% quantile of the attribute are considered outliers.*

1. 5. 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]



Both the female and male class have a uni-modal distribution; no gaps in data density are visible. Factor2 does mostly a good job in separating females and males; there is only overlap close to 0; Factor1 does a poor job separating the 2 classes. The classification task should not be too difficult as the examples are well separated although there are a few exceptions.

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.



**Both histograms: curves are continuous with no gabs or outliers, and somewhat smooth[1], bimodal with 2 (1??; 0??) not well separated maxima at 5-19 and 35-44 [1.5], values significantly drop beyond age 55[1]🡪skewed distribution**

**Comparison: Curves are somewhat similar until age 55 [1] (although there are more males initially [0.5]); decline in the male curve is significantly steeper---women live longer[1].** Other observations might receive credit; points will be subtracted if you write things which do not make any sense or are false.

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’

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!

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].**

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

Fast; runtime complexity is basically O(n); also saves time by minimizing an “implicit” objective function

Easy to use; no complex parameter values have to be selected…

Its properties are well understood!

Can deal with high-dimensional datasets

The properties of clusters can be “tweaked” by using different kind of distance functions/Kernel approaches

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**

**Centroids:**

**c1: (4, 4)**

**c2: (2, 2)**

**c3: (7, 7)**

**c4: (8, 8)**

**Clusters:**

**C1 = {(4, 4), (5, 5)}**

**C2 = {(2, 2), (0, 4), (4, 0)} assigning (0,4) and (4,0) to cluster C1 is also correct!**

**C3 = {(6, 6)}**

**C4 = {(8, 8), (9, 9)}**

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

 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}[[3]](#footnote-3); 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.*

14. DBSCAN

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

*A border point has fewer than MinPts points within Eps but is in the neighborhood of a core point [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]

*Some clusters might disappear [1]*

*Some bigger cluster might be broken up into several smaller clusters [1]*

*Some clusters might shrink in size*

*There will be a decrease in the number of core points. [1]*

*There will be an increase in the number of outliers. [1]*

c) How does DBSCAN form clusters

No answer given!

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)!

We convert the Oph rating values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’, ‘very poor’ to 5:0 using **φ; then we compute the distance** by taking L-1 norm and dividing by the range, 5 in this case.

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

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

Assign weights 1 to qud, 1 to Power-used and 0.2 to Gender

Now[8]: one error: 2.5-5 two errors: 0-2 distance functions not properly defined: at most 3 points

**d(u,v) = (1\*|(u.gpa)/0.8 – (v.gpa)/0.8| + 1\*|φ(u.qud) – φ(v.qud)|/5 + 0.2\*dgender(u.gender, v.gender)) /2.2**

2 students: c1=(111111111, ‘good’, 2.9, male) and c2=(222222222, ‘very poor’, 3.7, female)!

d(c1,c2)= (1 + 3/5 + 0.2)/2.2= 1.8/22=9/11=0.82 [2]

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)
3. The one that originates from {1, 3, 2} is identical and therefore equally good. [↑](#footnote-ref-3)