Solution Sketches Midterm Exam1

COSC 4335 *Data Mining*

March 3, 2016

Your Name:

Your student id:

Problem 1 --- K-means, K-medoids and Clustering in General [16]

Problem 2 --- Similarity Assessment [10]

Problem 3 --- Hierarchical Clustering [8]

Problem 4 --- Data Mining in General [4]

Problem 5 --- Data Analysis [16]

:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 16-19% towards the course grade.

1. **K-Means, K-medoids, and Clustering in General [16]**

a. What are the goal and objectives of clustering? Limit your answer to 2-4 sentences! [3]

The goal of clustering is to partition a dataset/a set of objects into homogenous, non-overlapping groups in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)[2.5]; that is, its goal is to discovery meaningful groupings [1]. Moreover, identifying outliers—objects do not belong to any group—is important [0.5].

At most 3 points; other observations might receive credit!

b. When does K-means terminate? When does K-medoids terminate? [2]

when there is no more change in the set of centroids/clustering[1.5]; when the objective function/SSE function does not improve[1.5]

c. When using K-means, it is recommended to run K-means multiple times—starting with different initial clusters in each run. What is the reason for this? [3]

K-means employs a steepest decent hill climbing strategy with respect to SSE and climbs the nearest valley from the stating position/initial clustering; consequently, using different initializations K-means ends up on different valleys that differ differ with respect to the SSE of the obtained clustering.

d. Assume we use K-medoids to cluster six objects 1,…,6 and k=2 and our current set of representatives is {1,6} and the object distance matrix is:

0 1 2 3 4 5 🡨object 1’s distance

 0 7 7 2 2

 0 3 3 2

 0 3 6

 0 3

 0

 What cluster does K-medoids form for this set of representatives and what is the

 SSE error of the obtained clustering? [3] {1,2,3,4} or {1,2,4} and {5,6} or {3,5,6} and

SSE=1\*\*2+3\*\*2+2\*\*2+3\*\*2=23

No partial credit!

Problem1 Continued

f) Assume the following dataset is given: (1,1), (4,4), (5,5), (7,7), (9,9), (0,6), (6,0). 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: {(1,1), (4,4), (7,7)}

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

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

}

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

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

Centroids and newly formed clusters

C1: (4,4): (0,6) (6,0) (4,4) (5,5) — assigning (0,6) and (6,0) to cluster C2 is also correct!

C2: (3,3): (1,1)

C3: (7,7): (7,7) (9,9)

Centroids of newly formed clusters: C1: (15/4,15/4) C2: (1,1) C3: (8,8)

One error: at most 3 points two errors: 0-1 points

**2. Similarity Assessment [10]**

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]

**3. 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[[1]](#footnote-1) link hierarchical clustering is applied to the dataset? What dendrogram will be returned? [4]

A-E, A-E-C, A-E-C-D. A-E-C-D, A-E-C-D-B

One error: at most 1 point

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_%28graph_theory%29) *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]

4. Data Mining in General [4]

What are the goals and objectives of data mining? Limit yourself to 3-5 sentences!

**Finding something interesting (patterns/associations), models, groupings, summaries,…) in large amounts of data [2.5]. descriptive and predictive techniques, possibly listing what they do [1], use of interesting measures[1], scalable algorithms that can deal with large amounts of data [1]**

At most 4 points; other observations might deserve credit!

5. Data Analysis [18]

a) Assume you have an attribute A with mean value 4 and standard deviation 2; make an estimate of the probability that a value of the attribute A lies in the interval [0,8]—Hint assume that A follows normal distribution! Also give a reason for your answer [3]

If we convert variable A into the into its zscore zA, we obtain

P(0<A<8)=P( 8-4/2 < zA < 0-4/2) = P( -2 < zA <2) = **about 95% see** [**http://en.wikipedia.org/wiki/Standard\_score**](http://en.wikipedia.org/wiki/Standard_score)

b) The following boxplot has been created using the following R-code for an attribute x:

> x<-c(1,2,2,2,4,4,8,9,9,10,18,22)

> boxplot(x)



What is the median for the attribute x? What is the IQR for the attribute x? The lower whisker of the boxplot as at 1; what does this tell you? According the boxplot 18 is not an outlier and 22 as an outlier; why do you believe this is the case? [5]

Median is 6=(4+8)/2 [1]

IQR=9.5-2=7.5 [1]

1 is the lowest value in the dataset that is not an outlier [1] Every value that is 1.5\*IQR above the 75th percentile is an outlier; that is, for the particular boxplots values above 9.5+1.5\*7.5=20.75 and below -9.25 are outliers; consequently, 22 is an outlier and 1 and 18 is not, and the whiskers are therefore at 1 and 18! [2] No partial credit for incorrect answers!

c) Interpret the following histogram that captures the percentage of body-fat in a testgroup [4]:



Bimodal[[2]](#footnote-2), peaks at 27.5 and 37.5 [2.5], one gap/outlier at 45/50[1], somewhat skewed [0.5], somewhat continous, no large jumps in the frequency[1] at most 4 points

d) Interpret the supervised scatter plot depicted below; moreover, assess the difficulty of separating signals and background using attributes x and y based on the scatter plot! [4]



The signal class is centered in a single area (uni-modal) near (2.5,-1) [1] the background class is centered in 2 areas (bi-modal) and focused in two separate regions at (0.5, -2.25) and (2, 0.5) [1] with in=between regions dominated by red examples[0.5]; the classes are separated although some overlap/contamination with example of other classes occurs, particularly near the boundaries [1]; the second region of the background class is very dense[0.5] has a moon-like shape [0.5] and is enclosed by region dominated by signal examples [0.5]. The variable x seems to be more useful in separating the 2 classes; for example, most points less than 1 belong to the background class [1]; however, the variable y is still necessary to correctly classify the examples in moon-like dense are of the background class [0.5].

At most 4 points; slightly varying interpretations might deserve full credit

1. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-1)
2. Triomodal is also correct distinguishing between hills at 20 and 25 is also fine! [↑](#footnote-ref-2)