Solution Sketches

Midterm Exam1

COSC 4335 *Data Mining*

March 10, 2015

Your Name:

Your student id:

Problem 1 --- K-means/PAM [9]

Problem 2 --- Data Mining in General [4]

Problem 3 --- DBSCAN [12]

Problem 4 --- Hierarchical Clustering [8]

Problem 5 --- Similarity Assessment [9]

Problem 6 --- Data Analysis [19]

:

**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 and K-Medoids/PAM [9]**
2. If we apply K-means to 2D real-valued dataset; what can be said about the shapes of the clusters K-means is capable for discovering? Can K-means discover clusters which have a shape of the letter ‘S’. [2]

K means is only capable of discovering shapes that are convex polygons [1]

Cannot discover ‘S’ shape because ‘S’ is non convex. [1]

1. What are the main differences between K-means and PAM/K-medoids [3]

K-means is prototype based and it doesn’t use actual object in the dataset as centroids.

PAM uses actual objects in the data set as representatives [2]; PAM uses an explicit objective function that is minimized whereas K-means minimizes an objective functions implicitly without ever call it. [1]

Assume K-Means is used with k=3 to cluster the dataset. Moreover, Manhattan distance is used as the distance function:

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

 to compute distances between centroids and objects in the dataset. Moreover, K-Means’ initial clusters C1, C2, and C3 are as follows:

C1: {(2,2), (8,8}

C2: {(4,6), (8,0)}

C3: {(4,8), (6, 10)}

}

Now K-means is run for a single iteration; what are the new clusters and what are their centroids? [4]

Centroids: New Clusters: New Centroids

C1: (5,5) C1’: {(4,6)} (4,6)

C2: (6,3) C2’: {(2,2), (8,0)} (5,1)

C3: (5,9) C3’: {(8,8), (4,8), (6,10)} (6, 8.67)

Forgot to compute the new centroids (-1); error in centroid computations (-1); incorrect clustering results (-2 points); more than 2 errors: 0 points.

2) Data Mining in General

Classification and clustering are important data mining tasks. What are the main differences between the 2 tasks? [4]

Clustering vs Classification

1. Clustering: find similar groups of object[0.5]/discover classes[1], VS learn a model that predicts a class[1]./Learn to classify examples with **respect to an a priori given class structure** [0.5]
2. Unsupervised vs, Supervised [1]
3. Dataset consists of attributes VS, consists of attributes and class labels [0.5]
4. Similarity assessment which derives a distance function is critical VS. Classifiers are learnt from set of classified examples (with classifiers) [0.5]

Other answer might deserve credit! No more than 4 points total!

**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 | 3 | 6 | 7 | 7 | 8 |
| B |  | 0 | 6 | 6 | 5 | 5 |
| C |  |  | 0 | 2 | 1 | 3 |
| D |  |  |  | 0 | 3 | 5 |
| E |  |  |  |  | 0 | 5 |
| F |  |  |  |  |  | 0 |

1. Assume DBSCAN is run for this dataset with MINPOINTS[[1]](#footnote-1)=3 and epsilon=ε=4

How many clusters will DBSCAN return and how do they look like? [3]

1. **Cluster {C, D, E, F}**

**Other answers 1 point if close to correct solution; otherwise, 0! e.g. {C, D. E} gets 1 point**

b) Which objects are outliers and borderpoints in the clustering result obtained in a? Give a reason for your answers! [3]

A & B are outliers, since they are not core or border points [1.5; one error 0.5]

F is border point because they are within the neighbor of a core point which is C, but has only one other point in its ε-neighborhood [1.5]

c) How does DBSCAN form clusters? [4]

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

Verbal descriptions how clusters are formed are also fine…

d) Now we change epsilon to 10. What will be the clustering result in this case? [2]

If Eps= 10, then there will be one cluster containing all the objects.

**No partial credit here!**

**4) 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 | 1 | 2 | 8 | 9 |
| B |  | 0 | 3 | 6 | 5 |
| C |  |  | 0 | 7 | 10 |
| D |  |  |  | 0 | 4 |
| E |  |  |  |  | 0 |

a) Assume single[[2]](#footnote-2) link hierarchical clustering is applied to the dataset? What dendrogram will be returned? [4]

|  |
| --- |
| A B C D E  |

**If different solution: 1 point if close to correct solution; otherwise 0 points!**

b) What is the goal of hierarchical clustering? How is it different from traditional clustering algorithms, such as K-Means and DBSCAN[4].

The goal is to identify groups of similar objects incrementally (agglomerative/ divisive), which seeks to build a **hierarchy of clusters**, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. [2.5] do not mention hierarchy [-1.5]

Difference is that Hierarchical clustering computes multiple clusterings whereas K-means and DBSCAN form a single clustering. [1.5]

**5) Similarity Assessment [9]**

Design a distance function to assess the similarity of electricity company customers; each customer is characterized by the following attributes:

1. Ssn
2. Oph (“*on-time payment history*”) which is ordinal attribute with values ‘very good’, ‘good, ‘medium’, ‘poor’ and ‘very poor’.
3. Power-used (which is a real number with mean 100, standard deviation is 10, the maximum 1000 and minimum 20)
4. Gender is an nominal attribute taking values in {male, female}.

Assume that the attributes Oph and Power-used are of major importance and the attribute Gender is of a minor importance when assessing the similarity between customers.

Using your distance function compute the distance between the following 2 customers: c1=(111111111, ‘good’, 150, male) and c2=(222222222, ‘very poor’, 50, female)!

One possible answer: Ignore SSN as it is not important.

Normalize Amount\_spend using Z-score and find distance by L-1 norm

We convert the Oph rating values ‘very good’, ‘good, ‘medium’, ‘poor’, and ‘very poor’ to: 4:0 using a function **φ**

Find distance by taking L-1 norm and dividing by range i.e. 4

Assign weights 0.4 to Oph, 0.4 to Power-used and 0.2 to Gender

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

Now:

**d(u,v) = 0.4\*|(u.PowerUsed)/10 – (v.PowerUsed)/10| + 0.4\*|φ(u.Oph) – φ(v.Age)|/4 + 0.2\*dgender(u.gender, v.gender)**

2 customers:

c1=(111111111, ‘good’, 150, male) and c2=(222222222, ‘very poor’, 50, female)!

= 0.4[( 3 -0 )/4] + 0.4\*[ (150- 100)/10 - (50-100)/10] + 0.2\*1= 0.3+4+0.2=4.5

If distance functions do not make much sense give 2 points or less

Distance functions are not defined properly [-4-7]

One error [-2 to -3]; two errors [-5 to -6]; more than 2 errors at most 1 point!

6) Data Analysis [19]

a) Assume the following boxplot below is given. What does the box of the box plot tell you about murder rates in the US? The line splits the box into two uneven parts with the lower higher box being larger; what does this tell you? [4]



The box shows that the median has murder rate of ~7

It tells that about 50% of murder rate in Us counts between ~(4-11), where 25th percentile having ~4 and the 75th percentile around ~11. [2]

The line splits the box into 2 uneven parts tells that the murder rate is skewed. The higher box being larger means that lower murder rates are concentrated in a narrower range about 5-7

**Grade with “more tolerance” due to the typo!**

b) Assume two attributes A and B have a covariance of 0; what does this tell about the relationship between the two attributes? [2]

This says that A & B do not have a linear relationship.

c) 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]



Factor2 does mostly a good job in separating females and males; there is only overlap close to 0 [3]; Factor1 does a poor job separating the 2 classes[2]/

d) What is the goal of density estimation? [2]

To construct an estimate based on observed data of an unobservable of an unobservable underlying probability density function. That is, the goal is to derive a density functions that best approximates the distribution

e) 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 lies in the interval [2,6]—Hint assume that A follows normal distribution! [2]

P( 2-4/2 < x < 6-4/2) = P( -1 < x <1) = **about 68% see** [**http://en.wikipedia.org/wiki/Standard\_score**](http://en.wikipedia.org/wiki/Standard_score)

Do not mention 68%, only [0.5]; if they just say 68% give them full credit[2]

f) Interpret the following BMI histogram [4]

n

It has no gaps and no outliers

It is skewed to the right (global maxima is at ~26); BMI-frequency drops more quickly for lower BMI values compared to higher BMI-values

It is bi-modal (2 peaks)

The higher BMI 🡪 lower frequency (<12 and > 26)

The type of attribute is positive real numbers.

1. The object itself counts towards the number of objects in its ε-radius when determining core points! [↑](#footnote-ref-1)
2. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-2)