Solution Sketches

Midterm Exam

COSC 6335 *Data Mining*

November 5, 2013

Your Name:

Your student id:

Problem 1 --- K-means/PAM [12]

Problem 2 --- DBSCAN [9]

Problem 3 --- Similarity Assessment [9]

Problem 4 --- Decision Trees/Classification [13]

Problem 5 --- APRIORI [8]

Problem 6 --- Explanatory Data Analysis [4]

Problem 7 --- R-Programming [9]

:

**Grade:**



The exam is “open books” and you have 75 minutes to complete the exam. The exam will count approx. 26% towards the course grade.

1. **K-Means and K-Medoids/PAM [12]**
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 ‘K’. [2]

Convex polygons; no

1. What objective function does K-means minimize[[1]](#footnote-1)? [2]

**The sum of the squared distance of the objects in the dataset to the centroid of the cluster they are assigned to**

1. When does K-means terminate? When does PAM/K-medoids terminate? [2]

**When the clustering does not change; when there is no improvement with respect the objective function PAM minimizes with respect to the (n-k)\*k newly generated clusterings.**

1. Assume 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: {(2,2), (6,6)}

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

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

}

Now K-means is run for a single iteration; what are the new clusters and what are their centroids? [3]

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

**C1 centroid: (4,4) {(2,2), (4,6)} new centroid: (3,4)**

**C2 centroid: (6,3) {(6,6), (8,0)} new centroid: (7,3)**

**C3 centroid: (5,8) {(4,8), (6,8)} centroid: (5,8)**

**Remark: Assigning (6,6) to cluster C3 instead, is also correct!**

1. The following clustering that consists of 2 clusters

{(0,0), (2,2)} and {(3,4), (4,4)} is given. Compute the Silhouette for points (2,2) and (3,4)— use Manhattan distance for distance computations[3].

Silhouette((2,2))= (3.5-4)/4=-1/8

Silhouette((3,4))= (5-1)/5=4/5

**2) DBSCAN [9]**

1. Assume you have two core points a and b, and a is density reachable from b, and b is density reachable from a; what will happen to a and b when DBSCAN clusters the data? [2]

**a and b will be in the same cluster**

1. Assume you run dbscan(iris[3:4], 0.15, 3) in R and obtain.

dbscan Pts=150 MinPts=3 eps=0.15

 0 1 2 3 4 5 6

border 20 2 5 0 3 2 1

seed 0 46 54 3 9 1 4

total 20 48 59 3 12 3 5

What does the displayed result mean with respect to number of clusters, outliers, border points and core points?

Now you run DBSCAM, increasing MinPoints to 5:

dbscan(iris[3:4], 0.15, 5).

How do you expect the clustering results to change? [4]

**6 clusters are returned; 20 flowers are outliers, there are 13 border points and the remaining 117 flowers are core points.**

**There will be more outliers; some clusters will cease to exist or shrink in size; some other clusters might be broken into multiple sub-clusters.**

1. What advantages[[2]](#footnote-2) you see in using DBSCAN over K-means? [3]
* Not sensitive to outliers[0.5]; supports outlier detection [1]
* Can detect clusters of arbitrary shape and is not limited to convex polygons. [1.5]
* Not sensitive to initialization [0.5}
* Not sensitive not noise [0.5]

***At most 3 points!!***

**3) Similarity Assessment [9]**

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

1. Ssn
2. *Cr* (“*credit rating*”) which is ordinal attribute with values ‘very good’, ‘good, ‘medium’, ‘poor’, and ‘very poor’.
3. *Av-bal* (avg account balance, which is a real number with mean 7000, standard deviation is 4000, the maximum 3,000,000 and minimum 20,000)
4. Services (set of bank services the customer uses)

Assume that the attributes Cr and Av-bal are of major importance and the attribute Services is of a medium importance. Using your distance function compute the distance between the following 2 customers: c1=(111111111, good, 7000, {S1,S2}) and c2=(222222222, poor, 1000, {S2,S3,S4})

**We convert the credit rating values ‘very good’, ‘good, ‘medium’, ‘poor’, and ‘very poor’ to: 0:4; then the distance between two customers can be computed as follows:**

**d(u,v)=( |u.Cr-v.Cr)/4 + | (u.Av-bal  u.Av-bal-/4000)+ 0.2\* (1u.Services ∩ v. Services|)/ (|u.Services ∪ v. Services|))/2.2**

**d(c1,c2)= (2/4 + 2 +0.2\*3/4)/2.2=2.65/2.2=1.2**

**4) Decision Trees/Classification [13** **]**

1. Compute the GINI-gain[[3]](#footnote-3) for the following decision tree split (just giving the formula is fine!)[3]:

(12,4,6) (3,3,0)

(9,1,0)

 (0,0,6)

 G(0.6,0.2,0,3) – (6/22\*G(0.5,0.5,0) + 10/22\* G(0.9,0,1,0) + 0)

1. Assume there are 3 classes and 50% of the examples belong to class1, and 25% of the examples belong to class2 and class3, respectively. Compute the entropy of this class distribution, giving the exact number not only the formula! [2]

H(1/2,1/4,1/4)= ½\*log2(2)+ 2\*1/4log2(4)=3/2=1.5

1. The decision tree learning algorithms is a greedy algorithm—what does this mean? [2]

**Seeks the shortest path from the current state to a goal state/makes local decisions[1]; does not backtrack[1]; frequently, does not find the optimal solution[1].**

1. Assume you learn a decision tree for a dataset that only contains numerical attributes (except the class attribute). What can be said about the decision boundaries that decision trees use to separate the classes? [1]

**Axis parallel lines/hyperplanes (of the form att=value where att is one attribute of the dataset and value is a floating point number) where each axis corresponds**

1. Why is pruning important when using decision trees? What is the difference between pre-pruning and post pruning? [4]

**To come up with a decision tree that uses the correct amount of model complexity to avoid under and overfitting. [2]**

**Prepruning: directly prevents a tree from growing too much by using stricter termination conditions for the decision tree induction algorithm**

**Postpruning; Grows a large tree and then reduces it in size by replacing subtrees by leaf nodes.**

5) APRIORI [8]

a) Assume the APRIORI algorithm for frequent item set construction identified the following 7 3-item sets that satisfy a user given support threshold: **abc, abd, abe, bcd, bce, bde, cde;** what initial candidate 4-itemsets are created by the APRIORI algorithm in its next step, and which of those survive subset pruning? [4]

abcd (pruned acd not frequence)

abce (pruned ace not frequent

abde (pruned ade not frequent)

bcde (survives, all its 3-item subsets are frequent!)

b) The sequence mining algorithm, GSP—that was introduced in the lecture —generalizes the APRIORI principle to sequential patterns —what is the APRIORI principle for sequential patterns? In which of its steps does GSP take advantage of the APRIORI principle to save computation time? [4]

**When a sequence is frequent all its subsequences are also frequent [2]**

1. **When creating k+1-item that are solely created by combining frequent k-item sequences (not using infrequent k-item sequences)[1]**
2. **For sequence pruning, when we check if all k-subquences of the k+1-item sequence are frequent.[1]**

6) Exploratory Data Analysis [4]

a) Assume we have a group a females and a group of males and have boxplots concerning their body weight. Comparing the two boxplots the box of the boxplot of the male group is much larger than the box for the female group. What does this tell you? [2]

**There is much more variation with respect to bodyweight in the male group; the variance/spread of body weight is much large for the male group than for the female group. If neither variance nor spread are mentioned at most 0.5 points!**

b) Assume you have an attribute A for whose mean and median are the same. How would be this fact reflected in the boxplot of attribute A? [2]

**The line, representing the median value/50% percentile is in the middle the box and splits the box into 2 equal-sized boxes.**

7) R-Programming [9]

Suppose you are dealing with the Iris dataset containing a set of iris flowers. The dataset is stored in a data frame that has the following structure:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **sepal length** | **sepal width** | **petal length** | **petal width** | **class** |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Setosa |
| 2 | 4.4 | 2.9 | 1.4 | 0.2 | Setosa |
| 3 | 7.0 | 3.2 | 4.7 | 1.4 | Versicolor |
| 7 | **…** | **…** | **…** | **…** | **…** |

Write a function most\_setosa that takes a k-means clustering of the Iris dataset as its input, and returns the number of the cluster that contains highest number of Setosa examples; if there is a tie it returns the number of one cluster of the clusters that are in a tie. most\_setosa has two parameters x and n, where x is the object cluster assignment and n is the number of clusters and it called as follows:

y<-k-means(iris[1:4], 3)

z<- most\_setosa(y$cluster,3)

For example, if 3 clusters are returned by k-means and cluster 1 contains 15 Setosas, and cluster 2 contains 20 Setosas, and cluster 3 contains 15 Setosas, most\_setosa would return 2.

most\_setosa<-function(x,n) {

 nd<-data.frame(x, class=iris[,5])

 setosa\_max<-0

 best\_cluster<-1

 for (i in 1:n) {

 q<-nd[which(x==i & nd$class=='setosa'),]

 a<-length(q[,1])

 if (a>setosa\_max) {

 setosa\_max<-a

 best\_cluster<-i}

 }

 return(best\_cluster)}

#Test Examples

set.seed(11)

cl<-kmeans(iris[1:4],4)

table(cl$cluster, iris[,5])

most\_setosa(cl$cluster,4)

set.seed(11)

cl<-kmeans(iris[1:4],6)

table(cl$cluster, iris[,5])

most\_setosa(cl$cluster,6)

This solution basically combines the cluster assignment with the flower labels in a data frame and then queries it in a loop over the cluster numbers, counting the numbers of Setosas in the cluster, and returns the cluster number for which the query returned the most answers.

1. Be clear! [↑](#footnote-ref-1)
2. We are only interested in the advantages and not the disadvantages! [↑](#footnote-ref-2)
3. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-3)