COSC 6335 Data Mining (*Dr. Eick*)

Solution Sketches

Midterm Exam

October 25, 2012

Your Name:

Your Student id:

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

Problem 2 --- DBSCAN [8]

Problem 3 --- Similarity Assessment [9]

Problem 4 --- Decision Trees & Classification [15]

Problem 5 --- APRIORI [9]

Problem 6 --- Exploratory Data Analysis [4]

Problem 7 --- R-Programming [8]

:

**Grade:**



The exam is slightly too long; you are expected to solve about 90% of the problems. The exam is “open books” and you have 80 minutes to complete the exam. The exam will count 26% towards the course grade.

**1) K-Means and K-Medoids/PAM [12]**

1. How do K-means and K-Medoids/PAM form cluster? What is the main difference between K-means and K-medoids with respect to constructing clusters? [3]

**Both algorithms use representatives as cluster models. Each representative corresponds to a cluster, and clusters are formed by assigning objects to the dataset to the closest representative (using a 1-NN query with respect to the set of representatives) [1]**

**K-means uses centroids (of the objects in a cluster) as representatives[1], whereas K-mediods uses the objects in the dataset themselves as representatives[1].**

1. K-means is a very efficient clustering algorithm; why is this the case—what of K-Means’ properties contribute to its efficiency? [3]

**The loop—which assigns objects to clusters and then recomputes centroids until there is no more change—only takes a few iterations and the complexity of each iteration is only O(k\*n), where k is the number of clusters and n is the number of objects in the dataset. [2]**

**K-means minimizes the mean-squared objective function without ever computing the mean-squared error (implicit objective function)[1]**

1. Is K-medoids capable of discovering clusters which have the shape of the letter ‘K’? [1]

**NO**

**d)** Assume the following dataset is given: (2,2), (4,4), (5,5), (6,6), (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: {(2,2), (4,4), (6,6)}

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

C3: {(5,5), (7, 7), (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’|**

**Centroid c1(4,4), c2(3,3), c3(7,7)**

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

**C2 = {(2,2), (6,0), (0,6)}**

**C3 = {(6,6), (7,7), (9,9)}**

**\*(6,0), (0,6) could be also assigned to C1, which is a second correct solution.**

**One Error at most 2 points**

**Two Errors 0 to 1 points.**

**2) DBSCAN [8]**

1. What is a border point in DBSCAN? Assume we have a border point b which is in the radius of 3 different core points c1, c2, and c3. How does DBSCAN deal with this situation—to which cluster or cluster(s) is b assigned or is it treated as an outlier? [3]

**A border point is not a core point, but falls within the radius of a core point [1].**

**b is not considered an outlier, it is assigned to the core point that is processed first[2].**

1. Assume you run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and you obtain 4 clusters and 30% of the objects in the dataset are classified as outliers. Now you run DBSCAN with MinPoints=6 and epsilon=0.15. How do expect the clustering results to change? [5]

**Less outliers[1.5], small clusters might be merged into bigger cluster[2], and new clusters might emerge in regions which did not contain any clusters before[1.5].**

**3) Similarity Assessment [9]**

Design a distance function to assess the similarity of electricity company customers; each customer is characterized by the following attributes[[1]](#footnote-1):

1. Ssn
2. Oph (“*on-time payment history*”) which is ordinal attribute with values ‘very good’, ‘good, ‘medium’, and ‘poor’.
3. Power-used (which is a real number with mean 40, standard deviation is 10, the maximum 300 and minimum 6)
4. Age (which is an integer with mean value 39, standard deviation 15, maximum 103 and minimum 18)

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

**Example Solution—uses z-scores (***there many other correct solutions***):**

1. **We assume that the values of the attribute** *Oph* **are converted to numbers as follows:** ‘very good’=3, ‘good’=2, ‘medium’=1, and ‘poor’=0.
2. **Let u and v be two customers; their distance d is measured as follows:**

**d(u,v) = 1\*|u.Oph  v.Oph|/3**

**+ 1\*|(u.Power-used-40)/10  (v.Power-used-40)/10|**

**+ 0.2\*|(u.Age-39)/15  (v.Age-39)/10|**

**Correct solution, but does not use normalization for the numerical attributes: 5 points; one Major Error: At most 4 points; one Minor Error: At most 6 points;**

**2 Errors: at most 2 points; 0 points if distance function does not make any sense!**

**4) Decision Trees/Classification [15** **]**

1. Compute the GINI-gain[[2]](#footnote-2) for the following decision tree split (just giving the formula is fine!)[5]:

(5,3,12) (4,1,0)

(1,2,0)

 (0,0,12)

**G(5/20, 3/20, 12/20) – (5/20\*G(4/5, 1/5, 0) + 3/20\*G(1/3, 2/3, 0) + 0)**

 **One Error 0-2.5 points; two errors no points!**

1. If the entropy function has a value is 0, what does this mean? Why do decision tree learning algorithms prefer choosing tests which lead to a low entropy? [3]

**The node is pure; all examples that are associated with the node belong to the same class [1.5]. The higher the purity the lower the (training) error[1.5].**

c) What are the characteristics of under-fitting when learning decision trees? What can be done to deal with under-fitting? [3]

**The model is not complex enough, training error and testing error are both high. [1.5]**

**Increase the number of objects in training set [0.5],**

**Reduce pruning when generating the decision tree [1.5]**

**At most 3 points!**

d) Assume you apply the decision tree learning algorithm to a data set which does not contain any inconsistent[[3]](#footnote-3) examples, and you continue growing the decision tree until you have leafs which are pure. What can be said about the decision tree which you obtain following this procedure? [4]

**Decision tree classifies all training examples correctly! [3]**

**Decision tree is quite large! [1]**

**Decision tree likely exhibits overfitting! [1]**

**At most 4 points!**

5) APRIORI [9]

a) Assume the APRIORI algorithm identified the following 6 4-item sets that satisfy a user given support threshold: **abcd, abce, abcf, bcde, bcdf, and bcef;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

**abcde, abcdf, abcef, bcdef**

**All four 5-item sets are pruned**

**One Error at most 2 points.**

**Two Errors at most 0.5 points.**

b) Assume APRIORI identified a frequent item set {A, C, D, E}. How is this item set used for generating candidate rules? How does the rule-generation algorithm determine that a generated rule satisfies the confidence threshold? [5]

**This item set is used to generate all possibly rules which contain all the attributes in {A,B,C,D}; e.g. ABD🡪C. [2]. Next, the algorithm computes the confidence for each rule by dividing the support for {A,B,C,D} by the support of the left hand side of the rule—the support of {A,B,D} in the example [2]. All these support values have already been computed by APRIORI when generating the frequent itemset; they only need to be retrieved when computing confidence [1].**

6) Exploratory Data Analysis [4]

Interpret the following histogram for the body weight in a group of cancer patients!



**Two peaks around body weight 63kg and 78kg [2]**

**Median around 70kg[0.5]**

**No gap** or **two small gabs at 112&118, not significantly skewed[2]**

**At most 4 points; other observations might deserve credit!**

7) R-Programming [8]

Suppose you are dealing with the Iris dataset *Iris.dat* without header information on your hard drive under directory */usr/data*. The dataset is delimited by comma and the first 5 observations are described as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **sepal length** | **sepal width** | **petal length** | **petal width** | **class** |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 3 | 7.0 | 3.2 | 4.7 | 1.4 | Iris-versicolor |
| 4 | 5.0 | 2.0 | 3.5 | 1.0 | Iris-versicolor |
| 5 | 6.3 | 3.3 | 6.0 | 2.5 | Iris-virginica |
| 6 | **…** | **…** | **…** | **…** | **…** |
| 7 | **…** | **…** | **…** | **…** | **…** |

We are only interested in the subset *S* that has the second attribute’s value larger than 3.05 (sepal width > 3.05). Write *R* function that takes the dataset’s full path and a K-means seed as an input and applies the k-Means clustering algorithm to obtain 4 clusters in the subset *S*. Your function should return the result object *O* as the output.

**myFun <- function (path, seed)**

**{**

 **data <- read.table(“path”, sep = “,”)**

 **data <- subset (data, data$V2 > 3.05)**

**set.seed(seed)**

**result <- kmeans(data[-c(5)], 4)**

 **return (result)**

**}**

1. E.g. (111234232, poor, 41.41, 46) is an example of a customer description. [↑](#footnote-ref-1)
2. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-2)
3. Two examples are inconsistent, if they agree in all attribute values but disagree in the class they belong to. [↑](#footnote-ref-3)