Solution Sketches Midterm Exam

COSC 6335 Data Mining (*Dr. Eick*)

October 25, 2011

Your Name:

Your Student id:

Problem 1 --- K-Means/PAM [16]:

Problem 2 --- DBSCAN [8]:

Problem 3 --- Similarity Assessment [8]

Problem 4 --- Decision Trees & Classification [16]:

Problem 5 --- APRIORI [7]:

Problem 6 --- Exploratory Data Analysis [10]:

:

**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 approx. 26% towards the course grade.

**1) K-Means and K-Medoids/PAM[16]**

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! [5]

**Looking for: compact clusters[1] which minimize the MSE/SE 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. Compare K-means and K-Medoids/Pam. What are the main differences between the two algorithms? [5]

**At most 5 points!**

|  |  |
| --- | --- |
| **K-means** | **K-medoids** |
| **Uses centroids which not necessary belong to the dataset as prototypes** | **Uses object in the dataset (representatives) as prototypes [1.5]** |
| **Implicit fitness function; is not called during search** | **Explicit fitness function, is called during search [1]** |
| **More efficient; kind of O(n)** | **Less Efficient O(n\*\*2) [1.5]** |
| **More sensitive to outliers** | **Less sensitive to outliers [1]** |
| **Uses distances between centroids and objects in the dataset for cluster assignment** | **Uses distances between objects in the dataset for cluster assignment; does not use distance matrix [1]** |
| **?!** | **?!** |

1. What can be said about the shapes of cluster that K-medoids is capable to discover? [1]

**Convex Polygons**

**d)** 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? [5]

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

**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)}**

**Grading of Incorrect Solutions:**

**One Error: at most 2.5 points**

**Two Errors: at most 1 point**

**More Errors: 0 points**

**2) DBSCAN [8]**

1. How does DBSCAN form clusters? [4]

**See solution of a previous homework**

1. Assume I run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and I obtain 4 clusters and 5% of the objects in the dataset are classified as outliers. Now I run DBSCAN with MinPoints=8 and epsilon=0.1. How do expect the clustering results to change? [4]

**The graph whose nodes contain core and border points and whose edges connect directly-density-connected points will have less nodes and edges; as a result of that:**

1. **There will be more outliers [2]**
2. **Some clusters that exist when using the first parameter setting will be deleted or split into several smaller sub-clusters [2]**

**3) Similarity Assessment [8]**

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

1. Ssn
2. Items\_Bought (The set of items the bought last month)
3. Age (integer, assume that the mean age is 40, the standard deviation is 10, the maximum age is 96 and the minimum age is 6)
4. Amount\_spend (Average amount spent per purchase in dollars and cents; it has a mean of 40.00 a standard deviation of 30, the minimum is 0.02 and the maximum is 398)

Assume that Items\_Bought and Amount\_Spend are of major importance and Age is of a minor importance when assessing the similarity of the customers. [8]

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

**Let u and v be two customers; their distance d is measured as follows:**

**d(u,v) = 1\*(1-(|(u.Items\_Bought ∩ v.Items\_Bought)|/ |u.Items\_Bought ∪ v.Items\_Bought)|) + 0.2\*|(u.Age-40)/10 - (v.Age-40)/10| + 1\*|(u.Amount\_spend-40)/30 - (v.Amount\_spend-40)/30|**

**Correct solution, but does not use normalization for the numerical attributes: 5 points**

**One Major Error: At most 3.5 points**

**One Minor Error: At most 5.5 points**

**2 Errors: at most 2 points**

**Give 0 points for distance functions which do not make any sense!**

**4) Decision Trees/Classification [16]**

1. Compute the GINI-gain[[2]](#footnote-2) for the following decision tree split (just giving the formula is fine!)[3]:

(5,3,12) (4,0,2)

(1,3,0)

(0,0,10)

**GINI-gain=GINIbefore-GINIafter=G(5/20,3/20,12/20) – 6/20\*G(2/3,0,1/3)-4/20\*G(1/4,3/4,0)-0**

**Partially correct : at most one point!**

1. If the GINI value is 0, what does this mean? [1]

**The objects which are associated with that node are pure—all objects belong to the same class.**

c) Are decision trees capable to learn XOR functions; for example.

If EITHER A OR B THEN class1 ELSE class2? Give a reason for your answer! [3]

**Solution: just give a decision tree which computes this function!**

**Verbal explanations🡪at most 2 points**

d) What are the characteristics of overfitting when learning decision trees? Assume you observe overfitting, what could be done to learn a “better” decision tree? [5]

**overfitting: training error low[0.5], testing error not optimal[0.5], models is too complex—the decision tree has to many nodes[1]**

**what to do to deal with it:**

1. **increase the degree of pruning in the decision tree learning algorithms to obtain smaller decision trees [2]**
2. **increase the number of training examples [1]**

**Other answers might exist which might deserve some credit!**

e) Most machine learning approaches use training sets, test sets and validation sets to derive models. Describe the role each of the three sets plays! [4]

**Training set: used to learn the model [1.5]**

**Test set: used to evaluate the model, particularly its accuracy [1.5]**

**Validation set: used to determine the “best” input parameter(s) for the algorithm which learns the model; e.g. parameters which control the degree of pruning of a decision tree learning algorithm. [2]**

5) APRIORI [7]

a) What is the APRIORI property? Where is it used in the APRIORI algorithm? [3]

Let *i* be an interestingness measure, and X and Y are sets:

**X⊆Y🡺*i*(X)≥*i*(Y) [1]**

1. **When creating k+1-itemsets from k-itemsets [1]**
2. **For subset pruning [1]**

b) Assume the APRIORI algorithm identified the following 7 4-item sets that satisfy a user given support threshold: **abcd, abce, abcf, acde, adef, bcde, and bcef;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

**Candidate 5-itemsets generated are: abcde, abcdf, abcef**

**abcde misses abde**

**abcdf misses bcdf**

**abcef misses acef**

**All three 5-itemsets are pruned!**

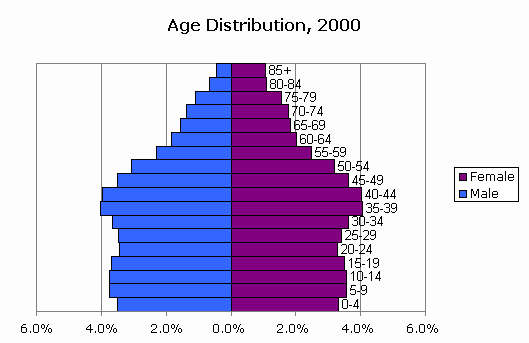
**One error: at most 2.5 points; 2 errors: at most 0.5 points**

6) Exploratory Data Analysis [10]

a) What is the role and purpose of exploratory data analysis in a data mining project? [4]

**a) collect background knowledge for the task at hand b) assess difficulty of the task to be solved c) provide knowledge to help in tool selection d) assess data quality / validate data e) helps in forming hypotheses f) find unexpected issues, errors and patterns in the data.**

b) Interpret the following 2 histograms and their relationships which describe the male and female age distribution in the US, based on Census Data. [6]



**Both histograms: curves are continuous with no gabs, 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. E.g. (111234232, {Coke, 2%-milk, apple}, 42, 3.39) is an example of a customer description. [↑](#footnote-ref-1)
2. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-2)