Midterm Exam A

COSC 6335 Data Mining

October 19, 2010

Your Name:

Your Student id:

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

Problem 2 --- DBSCAN [8]:

Problem 3 --- Region Discovery [5]:

Problem 4 --- Decision Trees & Classification [19]:

Problem 5 --- Data Mining in General [6]:

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

:

**Grade:**

MCAN00744_0000[1]

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. If you run K-means with a different seed, you frequently obtain different clustering results. Why is this case? What can be said about the optimality/compactness of solutions that K-means finds? [4]

The different solutions represent different local minima of the SSE fitness function ⎯which basically measures cluster compactness⎯that K-means minimizes; if different initializations are used search starts at different locations of the SSE-fitness landscape, usually ending up with different minima. K-means is guaranteed to find a local minimum but not the global minimum

1. K-means storage complexity is O(n), whereas the storage complexity of K-Medoids is typically O(n2), because its implementations store the distance matrix. Why is it desirable for K-medoids but not for K-means to store the distance matrix of the objects in the dataset? [3]

K-Means computes only distances between centroids (that are typically not in the dataset) and objects in the dataset, when creating clusters. K-Medoids on the other hand, compares distances between objects in the datasets a lot when assigning objects in the dataset to the closest representative; if the object distance matrix would not be stored, these distances would need to be computed again and again, slowing down k-medoids.

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

Convex Polygons

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

0 1 2 3 4 5 🡨object 1’s distance

0 7 7 2 2

0 3 3 3

0 3 4

0 5

0

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

clusters squared error? [4]

{1,3,4} and {2,4,5,6}

2\*\*2 + 3\*2 +2\*2 + 2\*2=21 or 2\*\*2 + 3\*2 and 2\*2 + 2\*2

e) What are the **main differences** between K-means and K-medoids? [4]

K-means: implicite distance function, uses centroids as representatives, follows a create-model, assign clusters,…

K-medoids: explicate distance function, searches through sets of representatives⎯which are objects in the dataset⎯ by replacing representative through non-representatives,…

**2) DBSCAN [8]**

1. How does DBSCAN form clusters? [4]

See solution from a previous exam/homework

1. Is DBSCAN, in general, capable to discover clusters of non-convex shapes, such as the letter T or V [1]?

Yes

1. Assume we have two core points o and v that are within each other’s radius—will o and v belong to the same cluster? Now assume, that we have a border point b within a radius of a core point u—will b and u always belong to the same cluster? Give reasons for you answer! [3]

Yes, o is density reachable from v and v is density reachable from o.

No if b is also in the radius of another corepoint w, and w is not density reachable from u, b might end up in the cluster formed by growing around w, if we is processed before u.

**3) Region Discovery [5]**

Region discovery algorithms provide plug-in fitness functions; what is the advantage of having clustering algorithms that support plug-in fitness functions?

Clusters can be found based on a domain expert notion of interestingness which has to be captured as a reward based fitness functions. Plug-in fitness functions can be designed to serve for different tasks, such as supervised clustering and co-location mining. Traditional clustering algorithms can only form clusters based on the distance relationships between the objects in the dataset and not based on other information.

**4) Decision Trees/Classification [19]**

1. Compute the GINI-gain[[1]](#footnote-1) for the following decision tree split[4] (just giving the formula is fine!):

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

(1,3,0)

GINIbefore=

GINIafter=

GINI-gain=GINIbefore-GINIafter

**Please ad formulas!**

b) Assume you learn a decision tree for a dataset that solely contains numerical attributes. What can be said about the shape of the decision boundaries that the learnt decision tree model employs? [2]

clusters are rectangles \ use rectangular boundaries

c) Are decision trees capable to learns disjunctive concepts; for example.

IF A>3 OR B>7 THEN class1 ELSE class2?

Give a reason for your answer! [3]

Yes. Reason: Give decision tree with two tests A>3 and B>7 that implements the above rule

d) What is the purpose of decision tree pruning in decision tree learning? Give an example of decision tree pruning technique, and briefly describe how it works [5]

Find the correct amount model complexity, avoiding both under and overfitting [2]

Example of a particular pruning method [3]

e) What roles do test sets play in decision tree learning? What role do training sets play in decision tree learning? [3]

Test sets are used to assess the accuracy of a given decision tree model

training sets are used to learn/derive decision tree models.

f) What are the characteristics of underfitting? [2]

The model is not complex enough, training error and testing error are both high.

5) Data Mining in General [6]

Give a specific example how data mining could help a company to conduct its business! Outline what data mining technique(s) would be used and how they benefit the company! Limit your answer to 5-8 sentences.

No answer given, but there were several issues with this problem:

1. The questions asked for a specific example, but some student give multiple examples, not surprising
2. Some students fail to describe or only very vaguely describe which data mining techniques would be used, and how would they be used, and what the benefit for the company is.
3. Finally, Problem5 ask for an essay-style answer, and students that do not present their answer in a well-written manner lose “some style points”.

6) Exploratory Data Analysis [10]

a) What role do histograms play in visualizing dataset? [2]

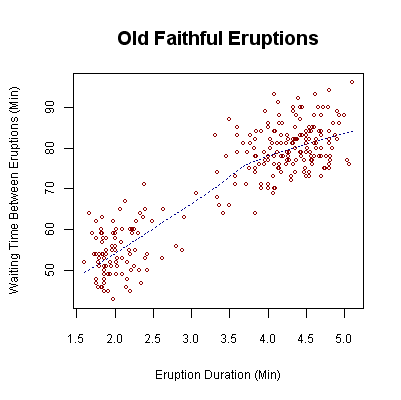
They summarize the distribution of an attribute/the density function of an attribute

b) Assume two attributes have a covariance of 0? What does this tell about the relationship between the two attributes? [2]

There is no linear relationship between the two attributes.

c) Interpret the following scatter plot that describes dependencies between the eruption duration and eruption time of Old Faithful in Yellowstone National Park! [6]

Somewhat liner relationship with minor/medium positive correlation; waiting-time and eruption time are either both low or both high, but there seems to be a lack of data in the[6.5,7.0]×[2.5,3.5] rectangle,…



1. (GINI before the split) minus (GINI after the split) [↑](#footnote-ref-1)