Qualifying Exam for Data Mining

Spring, 2015

Open book. Your code: ____________________________

Good luck! (2hrs, 100 points)

PLEASE WRITE YOUR ANSWERS SUCCINCTLY.

Q1. (10 pts) Please answer the following concept related questions.

1. (5 pts) Please give an example to show that *Jaccard coefficient* is a more appropriate metric than *Simple Matching Coefficient (SMC)* to measure the similarity of two data points.
2. (5 pts) Please give an example to show that *Recall* is a better metric than *Accuracy* to measure the classification performance of a classifier.

Q2. (20 pts) Assume that one node of an unfinished decision tree consists of data points as shown in Figure 1. The decision tree algorithm needs to decide whether and how this node should be split. Please utilize Gini index to make the decision.

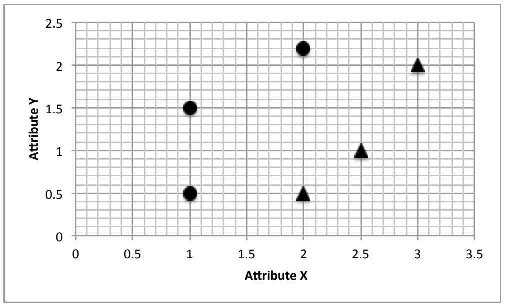

Figure 1: Six two-dimensional points where points (1, 0.5), (1, 1.5) , and (2, 2.2) have class label "+", and points (2,0.5), (2.5,1.0), and (3.0,2.0) have class label "-".

Q3. (25 pts) Given a data set $D = \{abe, cbde, a, af, b\}$ where each point is a string consisting of characters from the set {a,b,c,d,e,f}, please answer the following questions.

1. (12 pts) Let $K = 2$ and the initial mean values be a and b, what is the clustering results of running ONE iteration of the K-means algorithm? Please clearly denote the distance function you used and state the reason for using this distance function.
2. (13 pts) Please draw the dendrogram obtained after clustering points in D using the *Agnes* algorithm. Please use single link to calculate the distance between clusters.

Q4. (20 pts) Given the transactions in Table 1, please answer the following questions.

transaction id	customer id	transaction date	items bought together
1	c1	10/01/2013	abcdf
2	c2	10/02/2013	def
3	c2	10/02/2013	efg
4	c3	10/02/2013	cdef
5	c1	11/01/2013	def
6	c3	11/05/2013	cde

Table 1: Transaction database

- Let $sup = 50\%$ and $conf = 50\%$, please run the FP-Growth algorithm to calculate all the association rules in the form of $\alpha \to \beta$ where $|\alpha| = 1 \wedge |\beta| = 1$. The symbol $|A|$ denotes the cardinality of a set A.

Q5. (25 pts) Given a data set $D = \{0.1, 0.2, 0.3, 0.6, 1\}$

1. (10 pts) Assume that the data points follow a normal distribution, could you identify any outliers from D if a model-based approach is applied? Please succinctly show your calculation steps.
2. (10 pts) Please calculate whether D has any outliers if you apply the distance-based outlier detection approach, $DB(0.3, 20\%)$. If D does not have outliers, please explain. If D has outliers, please write down the outliers.

END