Qualifying Exam for Data Mining

Fall, 2014, Dec. 12, 2014

Open book. Your code: _____

Good luck! (2hrs, 100 points)

PLEASE WRITE YOUR ANSWERS SUCCINCTLY.

Q1. (25 pts) Please answer the following questions related to classification.

1. (12 pts) Suppose that a Naive Bayesian classifier is created using a given training data set. Then, predictions are made on several testing points using this classifier. The predicted positive class probabilities and the actual class labels for these testing points are shown as follows. Please draw the ROC

Record id	Actual class label	Predicted positive class probability $P(+\|record)$
1	-	0.1
2	+	0.3
3	-	0.4
4	+	0.4
5	+	0.6
6	+	0.8

   curve for the above prediction. Please clearly denote your splitting values and show the TP, FP, TN, FN, TPR, FPR values at each splitting value.

2. (13 pts) Please give a CONCRETE example to show that 1NN classifier gets better prediction accuracy than decision tree classifier. Your example should clearly denote the training data and testing data. The total number of training and testing data points should not be more than ten. You should also state the reason for the better performance of 1NN.

Q2. (20 pts) Given a data set with 1-dimensional points $D = \{1, 3, 6, 10, 20, 100\}$, please answer the following questions.

1. (10 pts) (K-means algorithm) Let $K = 2$ and the initial mean values be 1 and 100, what is the clustering results of running one iteration of the K-means algorithm. Please do NOT write down the steps.
2. (10 pts) (K-medoid algorithm) Let $K = 2$ and the initial medoids be 1 and 100. Assume the first step of the first iteration of the K-medoid algorithm tests whether 1 could be replaced with a new medoid with value 3. What is the result of the testing in this step? (I.e., Could 3 replace 1 as a new medoid?) Please justify your answer.

Q3. (35 pts) Given the transactions in Table 1, please answer the following questions.

Transaction id	customer id	transaction date	items bought together
1	c1	11/01/2014	abce
2	c2	11/02/2014	cd
3	c1	11/02/2014	de
4	c3	11/02/2014	bcde
5	c1	11/04/2014	cde
6	c2	11/05/2014	be
7	c3	11/05/2014	ce

Table 1: Transaction database

1. (20 pts) Let $minsup = 50\%$ and $conf = 70\%$. Please run the *apriori* algorithm to calculate all the association rules in the form of $\alpha \rightarrow \beta$ where $|\alpha| = 1 \wedge |\beta| = 1$. The symbol $|A|$ denotes the cardinality of a set A.
2. (15 pts) Let a sequence consists of itemsets that were bought by one customer and are ordered in the ascending order of their transaction times and let $minsup = 50\%$. Please calculate the frequent length-2 sequential patterns that are in the form of $\langle(\alpha)(\beta)\rangle$ where $|\alpha| = 1 \wedge |\beta| = 1$.

Q4. (20 pts) Suppose that we apply two approaches to identify outliers from the data set given in Figure 1: (i) a density-based outlier detection algorithm that utilizes Local outlier factor (LOF) with parameter $k = 3$ and LOF threshold 1.5 and (ii) a distance-based outlier detection algorithm that utilizes k-NN distance with parameter $k = 1$ and distance threshold 3. Please answer the following questions.

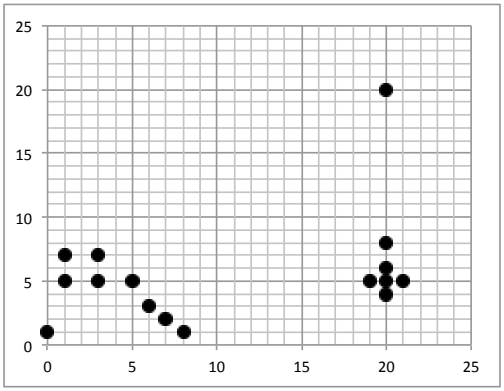

Figure 1: Data set

1. (10 pts) If the given density-based approach is applied, will it identify the object with coordinates (20, 8) as an outlier? Please succinctly show your calculation steps.
2. (10 pts) Assume that the set of outliers that you identified using the given density-based approach is $\mathcal{O}_1$, and the set of outliers that you identified using the given distance-based approach is $\mathcal{O}_2$. Will $\mathcal{O}_1$ and $\mathcal{O}_2$ be the same? (Note: you do not need to calculate $\mathcal{O}_1$ and $\mathcal{O}_2$). If they are the same, please explain the reason. If they are different, please identify one object that is in $\mathcal{O}_2$ but not in $\mathcal{O}_1$ or identify an object that is in $\mathcal{O}_1$ but not in $\mathcal{O}_2$.

END