CS561: Advanced Topics In Database Systems
Spring-2014

Homework 2

Total Points: 50
Release Date: 02/11/2014
Due Date: 02/22/2014 (11:59pm)

Question 1 [20 Points]—Frequent Itemset Mining

Transaction ID	Items
1	A, B, C, D
2	A, C, D, F
3	C, D, E, G, A
4	A, D, F, B
5	B, C, G
6	D, F, G
7	A, B, G
8	C, D, F, G

Given the table above where each row represents a transaction and the items sold in this transaction, answer the following questions:

Q1) Find all frequent itemsets using the Apriori technique (given in class) with support higher than or equal to 30% (support is the percentage of transactions containing the itemset).

Hint: As given in lecture (slide 35), create a table divided into scans, and for each scan (say scan number i) identify what are the candidates itemsets of size i considered in this scan (second column in slide 35), and then report the frequent itemsets of size i along with the support of each one (third column in slide 35, but add the support of each itemset).

Q2) What is the support and confidence of the following association rules (check the slides on how to compute the confidence):

A → BD
BD → AC
A → CD

Question 2 [30 Points]—Programming Assignment

* Choose any programming language of your choice to do this assignment.

Select one of the following algorithms and implement it.

1) **K-Means Algorithm**
    a. Input dataset: random points in 2D space between (0,0) and (100, 100)
        i. Create 1000 points
    b. Input parameter: K (the number of clusters to generate), and MaxIterations = 20
    c. Output: one representative from each cluster, e.g., the center of each cluster

2) **Hierarchical clustering**
    a. Input dataset: random points in 2D space between (0,0) and (100, 100)
        i. Create 100 points
    b. You should build the complete hierarchy

c. Output: Given an input L (which is a level in the hierarchy, where L= 1 means the root level, L= 2 means the second level having 2 clusters, etc.), you should report a representative from each cluster in that level.

3) Naïve Bayes classifier

a. Build a model from a training set
   - **i.** Assume we have 5 class labels, namely {C1, C2, …., C5}, and 10 numeric features, namely {F1, F2, …, F10}. You need to create a training dataset for the classifier that consists of 1,000 records, where the first field is the class label, and then the numeric values for the 10 features.
   - **ii.** Use a range and distribution of your choice for the values in each feature.
   - **iii.** The values in each record are comma separated.
   - **iv.** The model should look like:

Class label	Learned probability	Feature 1	Feature 2	….	Feature 10
C1	% of records with C1	Mean and variance of F1 values having label C1	…		….
C2	% of records with C2	Mean and variance of F1 values having label C2	….		…
…					

b. Classify a new object
   i. You will get a new object (without a label), but you will get all the values for the 10 features. Your output should be the label that the model predicts for this object.

What to Submit

You will submit a single zip file containing the following:

- A text, word doc, or pdf file containing the answer to Question 1
- A code for implementing Question 2. The instructor will meet with each student and ask each one to demonstrate the code, e.g., run it on a sample dataset. So, it is good to have a sample dataset ready (or the code that generates it).

How to Submit

Use blackboard system to submit your file zip file.

Late Policy:

We follow the late policy stated on the course website.