CS6220 Data Mining Techniques – Spring 2015
Assignment 4

1. Explain the difference between the following pairs of data mining techniques:

   (a) 'filter' and 'wrapper' methods for feature selection

   (b) 'bagging' and 'boosting' for ensemble learning

2. Recall that in class we applied the expectation maximization algorithm to the two coins (A and B) example. Now instead of picking a coin at random (as we did in class), assume that there is a third coin (C) that when flipped dictates which of coin A or B will then be flipped 4 times. If coin C is a head then we flip coin A four times; if coin C is a tail, then we flip coin B four times.

   So now our model consists of three parameters: π, μ_A, μ_B, where π is the probability of getting a head with coin C, μ_A is a probability of getting a head with coin A and μ_B is a probability of getting a head with coin B.

   Given a starting guess of ($\pi = .5, \mu_A = .6, \mu_B = .4$) and the following data:

   HHHT HTHH TTTH

   Your task is to run 3 iterations of the EM algorithm (3 E and M steps) by hand, carefully labeling each step as the E-Step or the M-Step. Show ALL of your work. At the end of each iteration, show the updated values for π, μ_A, μ_B.

   Note that in our example in class, there was no updates for π because we assumed that our generative model picks A or B at random. In your writeup, please give the formula for updating π.

3. Consider a dataset for frequent set mining as in the following table where we have 6 binary features and each row represents a transaction.

   0 0 1 0 1 0
   0 1 1 1 0 1
   1 0 0 0 1 0
   1 1 1 0 0 0
   0 0 0 1 0 0
   1 0 0 1 0 1
   0 0 1 1 1 1
   1 0 1 0 1 0
   1 0 0 1 0 0
   0 1 1 0 0 1

   (a) Illustrate the first three levels of the level-wise algorithm (set sizes 1, 2 and 3) for support threshold of 3 transactions, by identifying candidate sets and calculating their support. What are the maximal frequent sets discovered in the first 3 levels?

(b) Pick one of the maximal sets and check if any of its subsets are association rules with frequency at least 0.3 and confidence at least 0.6. Pleas explain your answer and show your work.

4. Given the following transaction database, let the min_support = 2, answer the following questions.

TID	Items
1	{a,b,e}
2	{a,b,c,d}
3	{a,c,d}
4	{a,c,e}
5	{b,c,f}
6	{a}
7	{a,b,c}
8	{b,d,e}
9	{a,c}
10	{a,b,d,e}

(a) Construct FP-tree from the transaction database and draw it here.

(b) Show d's conditional pattern base (projected database), d's conditional FP-tree, and find frequent patterns based on d's conditional FP-tree.