CS6220 Data Mining Techniques – Spring 2015
Assignment 4
Yaxin Huang

Question 1

Explain the difference between the following pairs of data mining techniques:
(a) 'filter' and 'wrapper' methods for feature selection
Filter method rank features according to the correlation or mutual information between features and the label. It has nothing to do with the algorithm that is going to be used. The top features will be selected.

On the other hand, wrapper method takes the learning algorithm, which is going to be used, into account and uses either forward search or backward elimination to choose the best features.

(b) 'bagging' and 'boosting' for ensemble learning
Bagging samples the training data with replacement into N bags. For each bag, an algorithm will learn the bag and output a model for future classification task. When a new data item comes in and needs to be classified, its label will be decided by the majority vote of those N models.

Boosting uses the same training set. Assuming we want to undergo N iterations to get N models. In each iteration, a model will be trained using a learning algorithm on a weighted training set (which initially is even weighted). Using the model got from this iteration, the training set will be labeled and the weights of mislabeled data items will be increased. In the next iterations, same steps will be followed. At last, the classifier will be all the models with weights decided by their accuracy on training set.

Question 2

$$\text{The formula for updating } \pi:$$
$$\pi = N_A / N = \frac{\sum A\text{'s responsibilites}}{3}$$

Data: HHHT HTHH TTTH

Iteration 1:

$$\text{Iteration } 1: \ \pi = 0.5, \mu_A = 0.6, \mu_B = 0.4$$
$$E-step:$$
$$p(A|x_1) \propto p(x_1|A)\,p(A) = \pi \cdot \mu_A^{heads} \cdot (1-\mu_A)^{tails} = 0.5 \cdot 0.6^3 \cdot 0.4^1 = 0.0432$$
$$p(B|x_1) \propto p(x_1|B)\,p(B) = (1-\pi) \cdot \mu_B^{heads} \cdot (1-\mu_B)^{tails} = 0.5 \cdot 0.4^3 \cdot 0.6^1 = 0.0192$$
$$\text{After normalization:} \ \ p(A|x_1) = 0.692307692, p(B|x_1) = 0.307692308$$
$$p(A|x_2) \propto p(x_2|A)\,p(A) = \pi \cdot \mu_A^{heads} \cdot (1-\mu_A)^{tails} = 0.5 \cdot 0.6^3 \cdot 0.4^1 = 0.0432$$
$$p(B|x_2) \propto p(x_2|B)\,p(B) = (1-\pi) \cdot \mu_B^{heads} \cdot (1-\mu_B)^{tails} = 0.5 \cdot 0.4^3 \cdot 0.6^1 = 0.0192$$
$$\text{After normalization:} \ \ p(A|x_2) = 0.692307692, p(B|x_2) = 0.307692308$$
$$p(A|x_3) \propto p(x_3|A)\,p(A) = \pi \cdot \mu_A^{heads} \cdot (1-\mu_A)^{tails} = 0.5 \cdot 0.6^1 \cdot 0.4^3 = 0.0192$$
$$p(B|x_3) \propto p(x_3|B)\,p(B) = (1-\pi) \cdot \mu_B^{heads} \cdot (1-\mu_B)^{tails} = 0.5 \cdot 0.4^1 \cdot 0.6^3 = 0.0432$$
$$\text{After normalization:} \ \ p(A|x_3) = 0.307692308, p(B|x_3) = 0.692307692$$

And we can get a table where each element represent the contribution of heads and tails from coin A or coin B:

	Coin A	Coin B
x1	H=2.076923076, T= 0.692307692	H= 0.923076924, T= 0.30692308
x2	H=2.076923076, T= 0.692307692	H= 0.923076924, T= 0.30692308
x3	H= 0.307692308, T= 0.923076924	H=0.692307692, T= 2.076923076
Total	H=4.46153846 , T= 2.307692308	H=2.53846154, T=2.690769236

$$M-step:$$
$$\pi = N_k / N = \frac{p(A|x_1) + p(A|x_2) + p(A|x_3)}{3} = \frac{0.692307692 + 0.692307692 + 0.307692308}{3} = 0.564102564$$
$$\mu_A = \frac{4.46153846}{4.46153846 + 2.307692308} = 0.659090909$$
$$\mu_B = \frac{2.53846154}{2.53846154 + 2.690769236} = 0.485436893$$

Iteration 2:

$Iteration\,2:\ \pi=0.564102564, \mu_A=0.659090909, \mu_B=0.485436893$

$E-step:$

$p(A|x_1)\propto p(x_1|A)\,p(A)=\pi\cdot\mu_A^{heads}\cdot(1-\mu_A)^{tails}=0.055059545$

$p(B|x_1)\propto p(x_1|B)\,p(B)=(1-\pi)\cdot\mu_B^{heads}\cdot(1-\mu_B)^{tails}=0.025657911$

$After\ normalization:\ \ p(A|x_1)=0.682126862,\ p(B|x_1)=0.317873138$

$p(A|x_2)\propto p(x_2|A)\,p(A)=\pi\cdot\mu_A^{heads}\cdot(1-\mu_A)^{tails}=0.055059545$

$p(B|x_2)\propto p(x_2|B)\,p(B)=(1-\pi)\cdot\mu_B^{heads}\cdot(1-\mu_B)^{tails}=0.025657911$

$After\ normalization:\ \ p(A|x_2)=0.682126862,\ p(B|x_2)=0.317873138$

$p(A|x_3)\propto p(x_3|A)\,p(A)=\pi\cdot\mu_A^{heads}\cdot(1-\mu_A)^{tails}=0.014730556$

$p(B|x_3)\propto p(x_3|B)\,p(B)=(1-\pi)\cdot\mu_B^{heads}\cdot(1-\mu_B)^{tails}=0.037308414$

$After\ normalization:\ \ p(A|x_3)=0.283067785,\ p(B|x_3)=0.716932214$

And we can get a table where each element represent the contribution of heads and tails from coin A or coin B:

	Coin A	Coin B
x1	H=2.046380586, T=0.682126862	H=0.953619414, T=0.317873138
x2	H=2.046380586, T=0.682126862	H=0.953619414, T=0.317873138
x3	H=0.283067785, T=0.849203355	H=0.716932214, T=2.150796642
Total	H=4.375828957, T=2.213457079	H=2.624171042, T= 2.786542918

$M-step:$

$$\pi=N_k/N=\frac{p(A|x_1)+p(A|x_2)+p(A|x_3)}{3}=\frac{0.682126862+0.682126862+0.283067785}{3}=0.54910717$$

$$\mu_A=\frac{4.375828957}{4.375828957+2.213457079}=0.664082411$$

$$\mu_B=\frac{2.624171042}{2.624171042+2.786542918}=0.484995337$$

Iteration 3:

$$Iteration\,3:\ \pi=0.54910717, \mu_A=0.664082411, \mu_B=0.484995337$$

$$E-step:$$

$$p(A|x_1)\propto p(x_1|A)\,p(A)=\pi\cdot\mu_A^{heads}\cdot(1-\mu_A)^{tails}=0.054020151$$

$$p(B|x_1)\propto p(x_1|B)\,p(B)=(1-\pi)\cdot\mu_B^{heads}\cdot(1-\mu_B)^{tails}=0.026490928$$

$$After\ normalization:\ \ p(A|x_1)=0.670965429,\ p(B|x_1)=0.329034567$$

$$p(A|x_2)\propto p(x_2|A)\,p(A)=\pi\cdot\mu_A^{heads}\cdot(1-\mu_A)^{tails}=0.054020151$$

$$p(B|x_2)\propto p(x_2|B)\,p(B)=(1-\pi)\cdot\mu_B^{heads}\cdot(1-\mu_B)^{tails}=0.026490928$$

$$After\ normalization:\ \ p(A|x_2)=0.670965429,\ p(B|x_2)=0.329034567$$

$$p(A|x_3)\propto p(x_3|A)\,p(A)=\pi\cdot\mu_A^{heads}\cdot(1-\mu_A)^{tails}=0.013822205$$

$$p(B|x_3)\propto p(x_3|B)\,p(B)=(1-\pi)\cdot\mu_B^{heads}\cdot(1-\mu_B)^{tails}=0.02987063$$

$$After\ normalization:\ \ p(A|x_3)=0.316349469,\ p(B|x_3)=0.683650534$$

And we can get a table where each element represent the contribution of heads and tails from coin A or coin B:

	Coin A	Coin B
x1	H=2.012896287, T= `0.670965429`	`H=0.987103701, T=` 0.329034567
x2	H=2.012896287, T= `0.670965429`	`H=0.987103701, T=` 0.329034567
x3	H= `0.316349469, T=0.949048407`	H= 0.683650534, T=2.050951602
Total	H=4.342142043, T=2.290979265	H=2.657857936, T=2.709020736

$$M-step:$$

$$\pi=N_k/N=\frac{p(A|x_1)+p(A|x_2)+p(A|x_3)}{3}=\frac{0.670965429+0.670965429+0.316349469}{3}=0.552760109$$

$$\mu_A=\frac{4.342142043}{4.342142043+2.290979265}=0.654615202$$

$$\mu_B=\frac{2.657857936}{2.657857936+2.709020736}=0.495233468$$

Question 3

(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?

If we use {a, b, c, d, e, f} to represent the items sold, then the matrix can be viewed as the following table:

Table-1 Transactional Database Converted from the Given Matrix

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

Level One:

First we start with C1. All the items are included in C1. We calculate the support count for every member in C1 as:

Table-2 C1 (L1 is Same as C1)

Item	Support Count
a	5
b	3
c	6
d	5
e	4
f	4

Because the support count threshold is 3 and all items appears equal or larger to 3, all candidates in C1 are also in L1.

Level Two:
From L1 we can join the members and create all combinations of size 2 and they consist of C2. Because every subset of every candidate is in L1, there is no need to prune the candidates. We scan the database again to compute the support count for each candidate in C2:

Table-3 C2 with Support Count

Itemset	Support Count	Itemset	Support Count	Itemset	Support Count
a, b	1	b, c	3	c, e	3
a, c	2	b, d	1	c, f	3
a, d	2	b, e	0	d, e	1
a, e	2	b, f	2	d, f	3
a, f	1	c, d	2	e, f	1

Therefore, removing the itemsets in C2 which do not have enough support count, we can get L2 as the following table shows:

Table-4 L2 with Support Count

Itemset	Support Count
b, c	3
c, e	3
c, f	3
d, f	3

Level Three:
From L2, we can join the itemsets and get the length-3 candidates as:
{c, e, f}

Because {e, f} is not a frequent itemset, we remove {c, e, f} from C3, and the resulting C3 is empty now. Thus, there is no more frequent itemsets to be mined.

Maximal Frequent Itemsets:
The maximal frequent itemsets are shown in the table-5.

Table-5 Maximal Frequent Itemsets

Itemset	Support Count
a	5
b, c	3
c, e	3
c, f	3
d, f	3

(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.
Here I choose {b, c} to compute the association rules and see if they are strong rules:

$$b \Rightarrow c \quad (support = 3/10, confidence = 3/3 = 1)$$
$$c \Rightarrow b \quad (support = 3/10, confidence = 3/6 = 1/2)$$

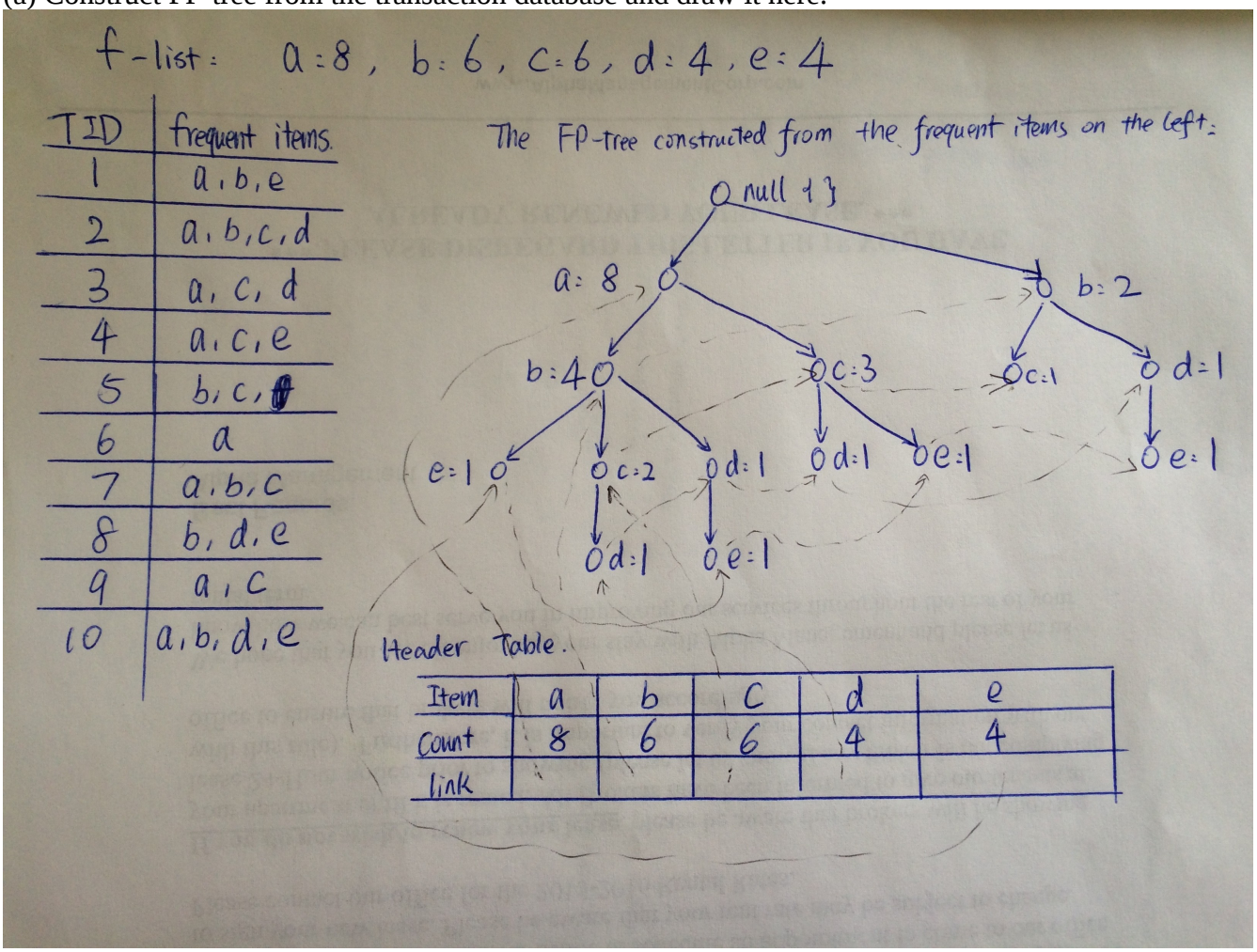
When we compute the confidence of a rule, we use the support count stored with L_k and the following formula:

$$confidence(X \Rightarrow Y) = P(Y|X) = \frac{support\ count\ of\ X \cup Y}{support\ count\ of\ X}$$

So both $b \Rightarrow c$ *and* $c \Rightarrow b$ *are strong association rules.*

Question 4

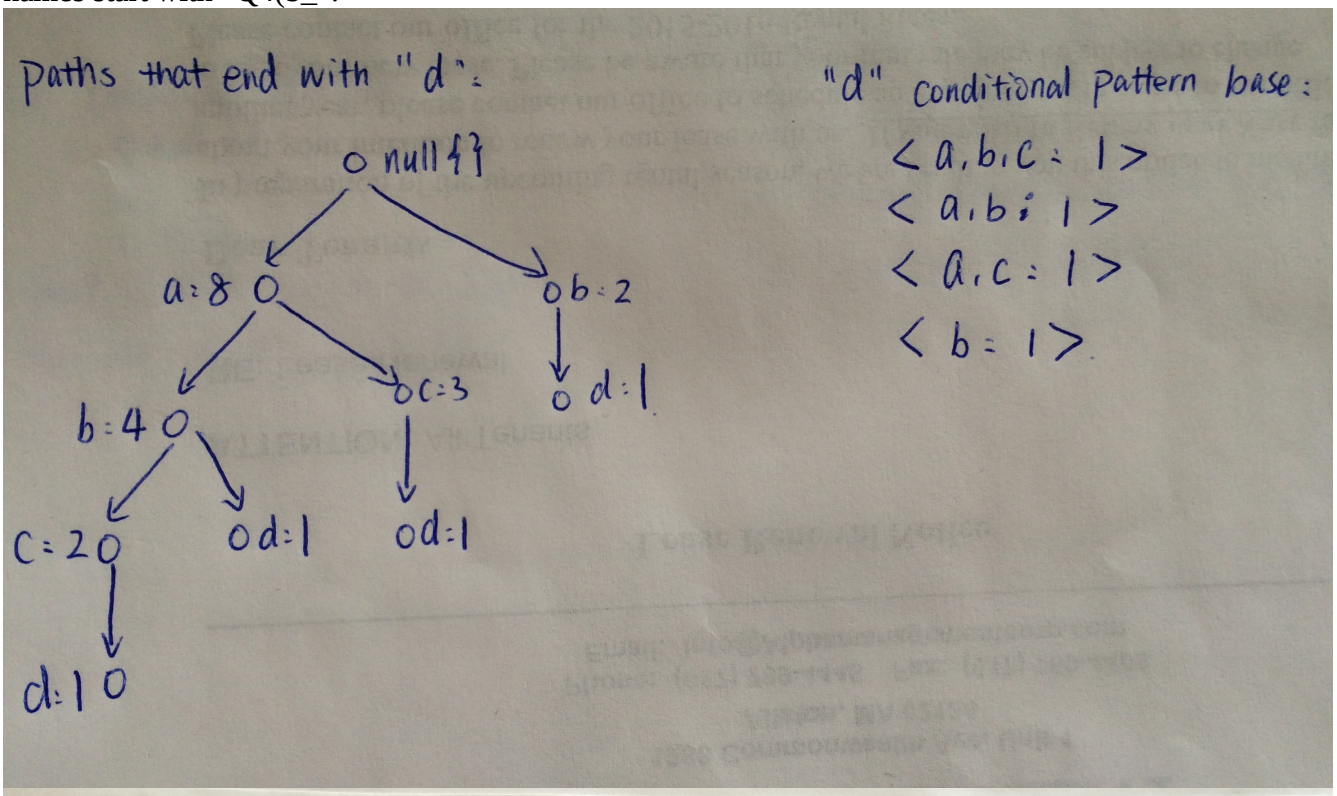
Given the transaction database, let the min support = 2, answer the following questions.
(a) Construct FP-tree from the transaction database and draw it here.

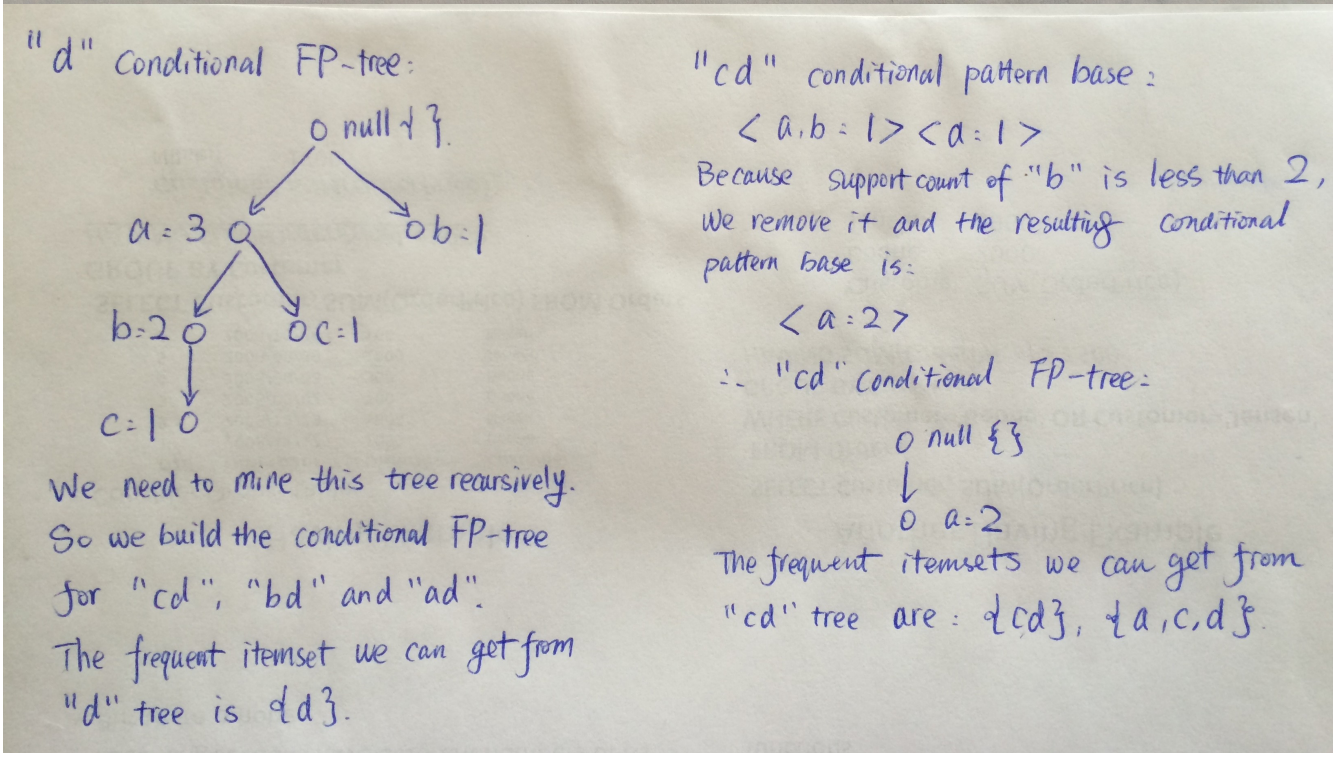


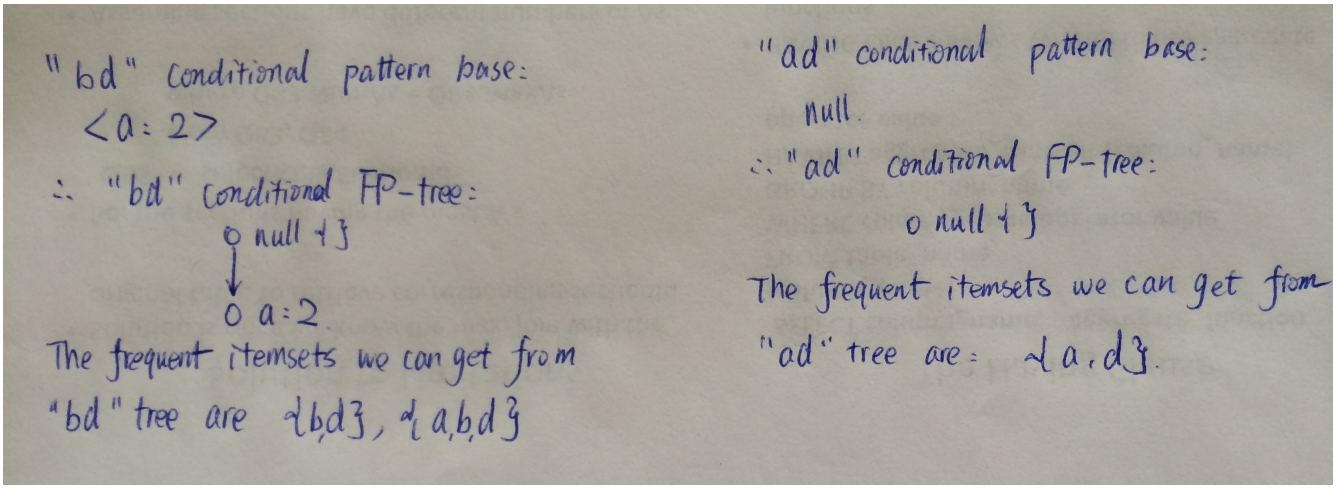
If the image above is not clear, the original file can be found in the same folder. The file name is "Q4(a)_Fptree.jpg".

(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.

If the images below are not clear to view, the original files can be found in the same folder. The file names start with "Q4(b_".







Thus, the frequent itemsets ending with “d” are:
{d}, {c,d}, {a,c,d}, {b,d}, {a,b,d}, {a,d}.