CSC 440 HW #4 Yangxin Fan

7.5

Step one: For each frequent item, find its Conditional pattern base and then its Conditional FP-tree.

Reapeat step one on newly created condition FP tree until resulting FP tree is empty or it only has one path. Then we store both frequent and infrequent patterns.

Step two: for each frequent itemset X, we scan infrequent item sets Z, which contain X. Let $W = Z \backslash X$, ie: W contain element in Z but not in X. If W is also a frequent itemset

Step three: Calculate $(P(X|W)+P(W|X))/2$ then whether W and X must be negatively correlated depend on $(P(X|W)+P(W|X))/2$

$= \left(\frac{Sup(Z)}{Sup(W)} + \frac{Sup(Z)}{Sup(X)}\right)/2$, note that $Sup(X)$, $Sup(Y)$ and $Sup(Z)$ are known in previous steps.

7.9. The distance measure Pat_Dist is a valid distance metric. It has following properties:

(a). $Pat\_Dist(P_1, P_2) > 0$, $\forall P_1 \neq P_2$

(b). $Pat\_Dist(P_1, P_2) = 0$, $\forall P_1 = P_2$

(c). $Pat\_Dist(P_1, P_2) = Pat\_Dist(P_2, P_1)$

(d). $Pat\_Dist(P_1, P_2) + Pat\_Dist(P_2, P_3) \geq Pat\_Dist(P_1, P_3)$

Proof of (a): $\forall P_1 \neq P_2$, Since P_1, P_2 are closed patterns, $\therefore 0 \leq \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|} < 1 \Rightarrow$ Pat_Dist$(P_1, P_2) > 0$.

Proof of (b): $\forall P_1 = P_2 \quad \therefore |T(P_1) \cap T(P_2)| = |T(P_1) \cup T(P_2)|$
$= |T(P_1)| = |T(P_2)| \quad \therefore$ Pat_Dist$(P_1, P_2) = 0$

(c) is obvious, since the formula is symmetric.

(d) Since $(T(P_1) \cap T(P_2)) \cup (T(P_1) \cap T(P_3)) \subseteq T(P_1)$

we have

$\because A \cup B = A + B - A \cap B$

$\therefore |T(P_1) \cap T(P_2)| + |T(P_1) \cap T(P_3)| - |T(P_2) \cap T(P_3)|$
$\leq |T(P_1)| \Rightarrow b_1 + c_1 - d_1 \leq a \quad *$

(d) $\Leftrightarrow 1 - \frac{b_1}{a+b_2} + 1 - \frac{c_1}{a+c_2} \geq 1 - \frac{d_1+d_2}{b_1+b_2+c_1+c_2-d_1-d_2}$

$\Leftrightarrow \frac{b_1}{a+b_2} + \frac{c_1}{a+c_2} \leq 1 + \frac{d_1+d_2}{b_1+b_2+c_1+c_2-d_1-d_2}$

$\because 1 + \frac{d_1+d_2}{b_1+b_2+c_1+c_2-d_1-d_2} \geq 1 + \frac{d_1}{b_1+b_2+c_1+c_2-d_1}$

$(d_2 \geq 0) \geq 1 + \frac{d_1}{b_2+c_2+a} \quad (*) = \frac{a+b_2+c_2+d_1}{a+b_2+c_2}$

$\geq \frac{b_1+c_1+b_2+c_2}{a+b_2+c_2} \quad (*) = \frac{b_1+c_2}{a+b_2+c_2} + \frac{b_2+c_1}{a+b_2+c_2}$

$\geq \frac{b_1}{a+b_2} + \frac{c_1}{a+c_2} \quad (\begin{matrix} a+b_2 \geq b_1, c_2 \geq 0 \\ a+c_2 \geq c_1, c_1 \geq 0 \end{matrix})$.

Hence the (d) holds.

7.10: In order to compress a large set of patterns to a small representatives, we can apply clustering method, we use the concept of δ-cluster. Given a transaction database, a minimum support min-sup, the cluster quality measure δ, the pattern compression problem is to find a set of representative patterns R such that for each frequent pattern $P_0 \in R$ (w.r.t min-sup), there's a representative pattern $P_r \in R$, which covers P_0 and $|R|$ is minimized.

8.3: If we directly prune decision tree and convert to rules, we would remove the subtree entirely. But if we prune after converting decision tree to rules, we may remove any precondition of it. The latter results in less information loss.

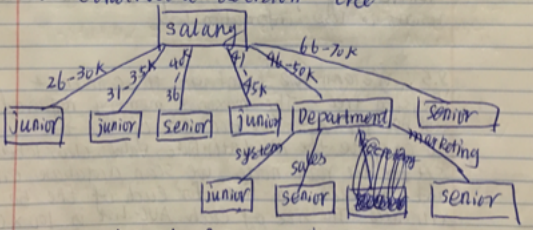
8.5. We can use RainForest to construct decision tree. The maximal memory required will be for AVC-set for root of the tree. To compute the AVC(attribute, value, class label) set for the root node, we scan database once and construct AVC-list for each of the 50 attributes. The size of each AVC-list is $100 \times C$, C is the number of class labels. Then size of AVC-set is $50 \times 100 \times C$, which can be easily fit into 512 MB memory given a reasonable C.

For other nodes, AVC-Sets are done in similar ways but they will be smaller since they will consider less number of attributes. To reduce number of scan for the original data set, we can parallely compute AVC-set for nodes at same level in the tree.

8.7.

(a). Basic decision tree algorithm must be modified to consider the count of generalized data tuple. ①. Count is considered to calculate the most common class among all the tuples. ②. the count of each tuple must be used when calculating attribute selection measure, such as information gain.

(b). Construct a decision tree



(c). According to Bayesian Theorem
Given such an observation data X, posterior probability of a Hypothesis H

	Predict		
	TP	FN	P
actual	FP	TN	N
	P'	N'	

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Let H: be x is senior

$$P(\text{senior}|X) = \frac{P(X|\text{senior})\,P(\text{senior})}{P(X)} \Rightarrow 0$$

$$P(\text{Junior}|X) = \frac{P(X|\text{Junior})\,P(\text{Junior})}{P(X)}$$

$$= \frac{\frac{3}{113}\cdot\frac{113}{165}}{\frac{3}{165}} = 1$$

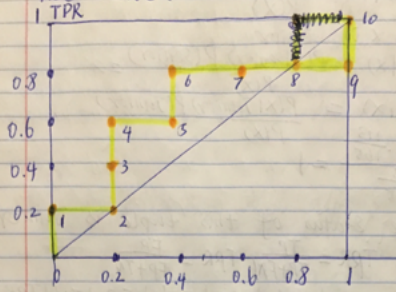
So, the status of this tuple is Junior.

$$TPR = \frac{TP}{TP+FN} \quad FPR = \frac{FP}{FP+TN}$$

8.12

Tuple #	Class	prob	TP	FN	FP	TN	TPR	FPR
1	P	0.95	1	4	0	5	0.2	0
2	n	0.85	1	4	1	4	0.2	0.2
3	P	0.78	2	3	1	4	0.4	0.2
4	P	0.66	3	2	1	4	0.6	0.2
5	n	0.60	3	2	2	3	0.6	0.4
6	P	0.55	4	1	2	3	0.8	0.4
7	n	0.53	4	1	3	2	0.8	0.6
8	n	0.52	4	1	4	1	0.8	0.8
9	n	0.51	4	1	5	0	0.8	1.0
10	P	0.4	5	0	5	0	1.0	1.0

ROC Curve:



8.14:

Let H be Null hypothesis: there's no significant difference between M_1 and M_2

$$t\text{-Statistic} = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{Var(M_1 - M_2)/K}}$$

where $Var(M_1 - M_2) = \frac{1}{K}\sum_{i=1}^{K}[(err(M_1)_i - err(M_2)_i - (\overline{err}(M_1) - \overline{err}(M_2))]^2$

$\overline{err}(M_1) = 27.72$ $\overline{err}(M_2) = 21.27$

$\therefore Var(M_1 - M_2) = \frac{1}{10}(1.65^2 + 11.25^2 + (-8.15)^2 + (-5.45)^2 + 3.85^2 + 14.15^2 + (-0.85)^2 + 0.15^2 + (-1.15)^2 + (-15.45)^2)$

$= 68.1225$

$t = \frac{27.72 - 21.27}{\sqrt{68.1225/10}} \approx 2.47$

For t-distribution with $10 - 1 = 9$ degree of freedom

(two tail)

K for 1% significance level is 3.25

since 2.47 < |3.25| , i.e: $-3.25 < 2.47 < 3.25$

we fail to reject null hypothesis, hence, there's no significant better model among / between M_1 and M_2.