Assignment 1

GUO, Yuchen No. 20477118

October 5, 2017

1 Data Preprocessing

1.1 Wavelet Transform

1.1.1 Describe the discrete wavelet transform

1. Find the $\frac{p_1+p_2}{\sqrt{2}}$ value of each pair p_1, p_2 of samples. Fill the first half of the array with the values.
2. Find the $\frac{p_1-p_2}{\sqrt{2}}$ value of each pair p_1, p_2 of samples. Fill the second half of the array with the values.
3. Repeat the process on the first half of the array.

1.1.2 Compute the discrete Haar wavelet transform

$$
\begin{aligned}
&[1,4,2,3,-2,-1,2,1] \\
\Rightarrow &[\frac{5}{\sqrt{2}},\frac{5}{\sqrt{2}},-\frac{3}{\sqrt{2}},\frac{3}{\sqrt{2}}],[-\frac{3}{\sqrt{2}},-\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}] \\
\Rightarrow &[5,0],[0,-3] \\
\Rightarrow &[\frac{5}{\sqrt{2}}],[\frac{5}{\sqrt{2}}]
\end{aligned}
$$

1.2 Principal Components Analysis

1.2.1 Calculate the covariance matrix of data as shown in the Table 1.

The code is:

```
X = [-1 -2 -3 1 2 3 1; -1 -1 -2 1 1 2 2; 1 4 -2 1 2 1 4];
covariance = cov(X');
```

The output is:

$$\begin{bmatrix} 4.8095 & 3.2857 & 1.2381 \\ 3.2857 & 2.5714 & 1.4762 \\ 1.2381 & 1.4762 & 4.2857 \end{bmatrix}$$

1.2.2 Calculate eigenvectors and eigenvalues of the covariance matrix.

The code is:

```
[V, D] = eig(covariance);
```

The output is:

$$V = \begin{bmatrix} 0.5484 & 0.4373 & 0.7128 \\ -0.8259 & 0.1499 & 0.5435 \\ 0.1308 & -0.8867 & 0.4434 \end{bmatrix}; D = \begin{bmatrix} 0.1561 & 0 & 0 \\ 0 & 3.4255 & 0 \\ 0 & 0 & 8.0851 \end{bmatrix}$$

So the eigenvectors are: $[0.5484, -0.8259, 0.1308]^T; [0.4373, -0.1499, 0.8867]^T; [0.7128, -0.5435, 0.4434]^T$.

And the corresponding eigenvalues are $0.1561, 3.4255, 8.0851$.

1.2.3 Calculate the proportion of total population variance explained by the first two components.

$$\lambda_1 = 0.1561$$
$$\lambda_2 = 3.4255$$
$$\lambda_3 = 8.0851$$

The proportion is:

$$\frac{\lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = 0.9866$$

2 Pattern Discovery

$$SessionLengthCategory = \begin{cases} SL1 & \text{if } SessionLength \leq 850 \\ SL2 & \text{if } 850 < SessionLength \leq 1433 \\ SL3 & \text{if } SessionLength > 1433 \end{cases}$$

$$NumWebPageCategory = \begin{cases} WP1 & \text{if } NumWebPage \leq 9 \\ WP2 & \text{if } 9 < NumWebPage \leq 13 \\ WP3 & \text{if } 13 < NumWebPage \leq 25 \\ WP4 & \text{if } NumWebPage > 25 \end{cases}$$

Session ID	Country	Session Length	#Web Page	Buy
1	NA	SL2	WP1	Yes
2	A	SL3	WP2	Yes
3	E	SL1	WP4	Yes
4	E	SL2	WP4	No
5	NA	SL1	WP1	No
6	A	SL2	WP3	Yes
7	A	SL3	WP3	Yes
8	A	SL1	WP1	No
9	NA	SL3	WP2	No
10	E	SL2	WP4	Yes

A means Asia, NA means North American, E means Europe.

2.1 Show the major steps to find the frequent patterns using Apriori of the transactions.

The following two pictures Figure 2.1, Figure 2.2 shows the two steps of finding frequent patterns.
The frequent patterns are:
{A}, {E}, {NA}, {SL1}, {SL2}, {SL3}, {WP1}, {WP4}, {Yes}, {No}, {A, Yes}, {E, WP4}, {SL2, Yes}.

2.2 Show the major steps to find the frequent patterns using FP-Growth of the transactions.

Figure 2.3 to Figure 2.8 shows the steps to find fp using FP-Growth.
So frequent items are: {A}, {E}, {NA}, {SL1}, {SL2}, {SL3}, {WP1}, {WP4}, {Yes}, {No}, {Yes, A}, {E, WP4}, {Yes, SL2}.

2.3 Based on the frequent patterns you get, which are closed frequent patterns? Which are max frequent patterns?

As Figure 2.9 shows:

1-Itemset	Sup
{NA}	3
{A}	4
{E}	3
{SL1}	3
{SL2}	4
{SL3}	3
{WP1}	3
{WP2}	2
{WP3}	2
{WP4}	3
{Yes}	6
{No}	4

Figure 2.1: Find 1-itemset frequent patterns

Closed frequent pattern: {A}, {E}, {NA}, {SL1}, {SL2}, {SL3}, {WP1}, {WP4}, {Yes}, {No}, {Yes, A}, {E, WP4}, {Yes, SL2}.
Maximal frequent pattern: {E}, {NA}, {SL1}, {SL3}, {WP1}, {WP4}, {No}, {Yes, A}, {E, WP4}, {Yes, SL2}.

2-Itemset	Sup
{NA, SL1}	1
{NA, SL2}	1
{NA, SL3}	1
{A, SL1}	1
{A, SL2}	1
{A, SL3}	2
{E, SL1}	1
{E, SL2}	2
{E, SL3}	0
{NA, WP1}	2
{NA, WP4}	0
{A, WP1}	1
{A, WP4}	0
{E, WP1}	0
{E, WP4}	3
{NA, Yes}	1
{NA, No}	2
{A, Yes}	3
{A, No}	1
{E, Yes}	2
{E, No}	1
{SL1, WP1}	2
{SL1, WP4}	1
{SL2, WP1}	1
{SL2, WP4}	2
{SL3, WP1}	0
{SL3, WP4}	0
{SL1, Yes}	1
{SL1, No}	2
{SL2, Yes}	3

2-Itemset(continue)	Sup
{SL2, No}	1
{SL3, Yes}	2
{SL3, No}	1
{WP1, Yes}	1
{WP1, No}	2
{WP4, Yes}	2
{WP4, No}	1

Figure 2.2: Find 2-itemset frequent patterns using Apriori

Item	Sup
{NA}	3
{A}	4
{E}	3
{SL1}	3
{SL2}	4
{SL3}	3
{WP1}	3
{WP2}	2
{WP3}	2
{WP4}	3
{Yes}	6
{No}	4

Freq-Item(Ordered)	Sup
{Yes}	6
{No}	4
{A}	4
{SL2}	4
{E}	3
{NA}	3
{SL1}	3
{SL3}	3
{WP1}	3
{WP4}	3

Figure 2.3: Find 2-itemset frequent patterns using FP-tree: Order items

Freq-Item(Ordered)	Sup
{Yes}	6
{No}	4
{A}	4
{SL2}	4
{E}	3
{NA}	3
{SL1}	3
{SL3}	3
{WP1}	3
{WP4}	3

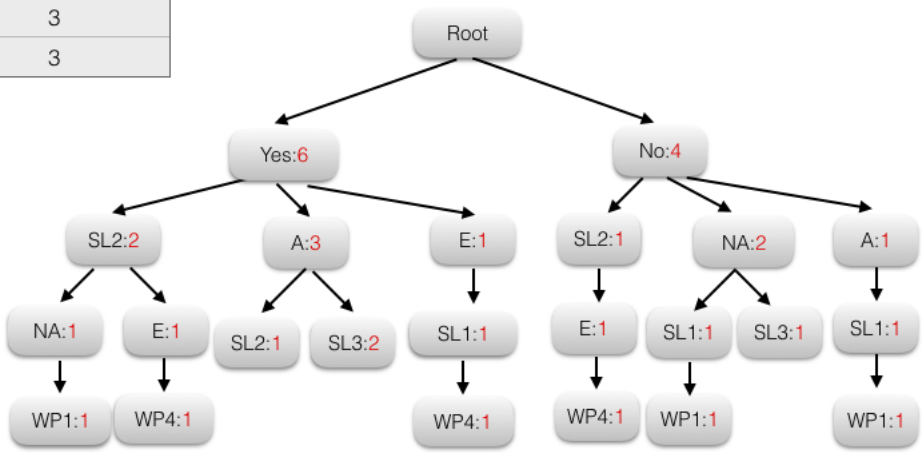

Figure 2.4: Find 2-itemset frequent patterns using FP-tree: Build FP-tree

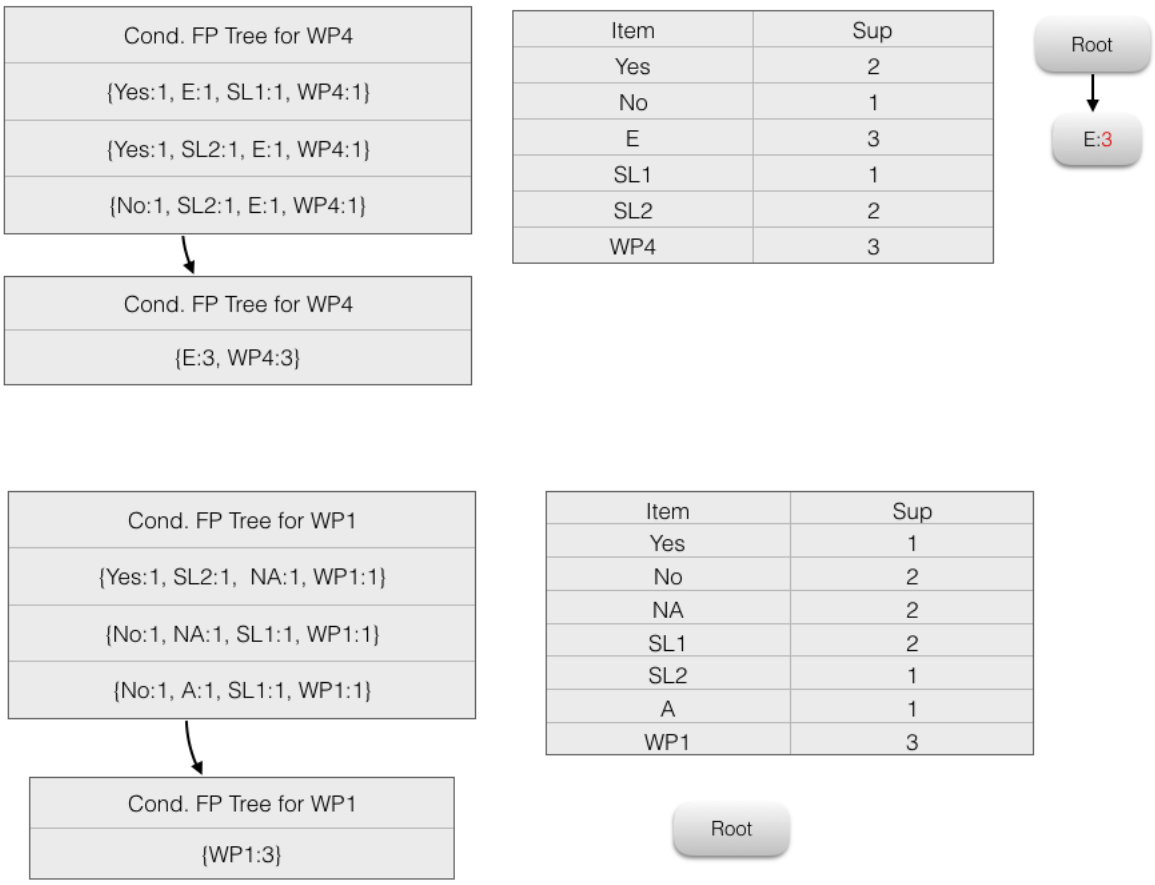

Figure 2.5: Find 2-itemset frequent patterns using FP-tree: Build Cond FP-tree for WP4, WP1

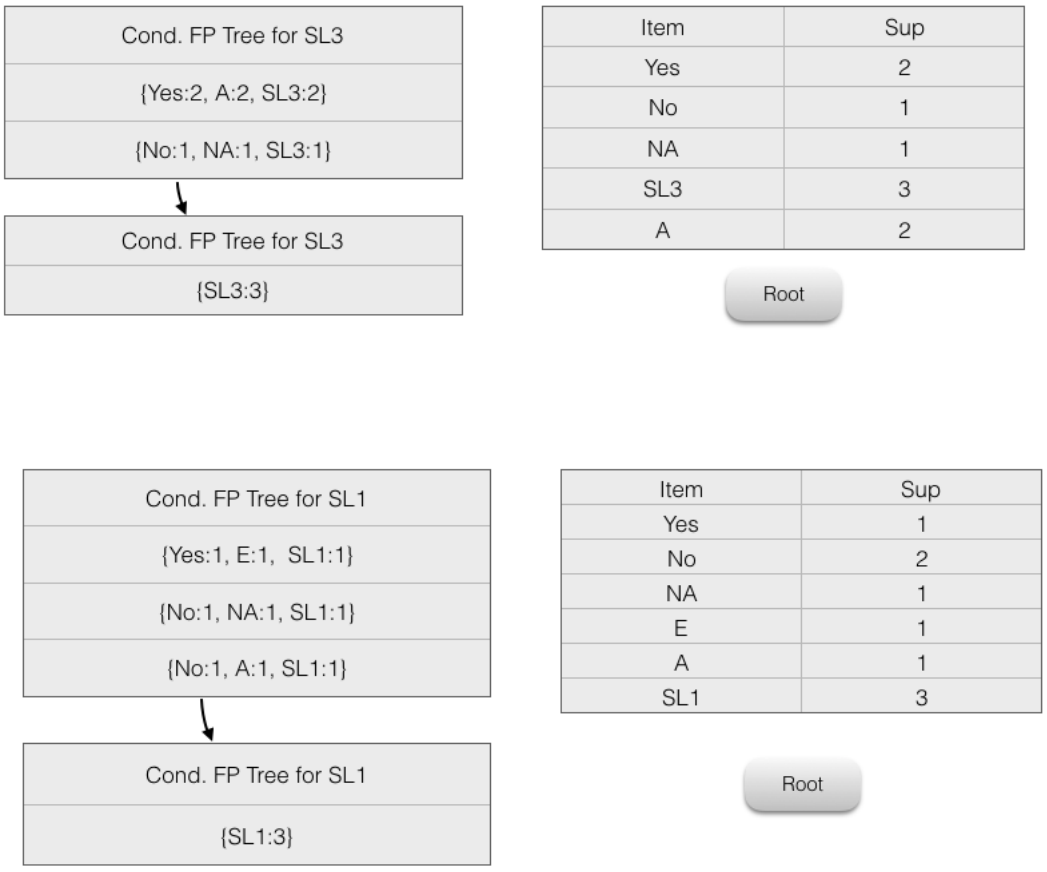

Figure 2.6: Find 2-itemset frequent patterns using FP-tree: Build Cond FP-tree for SL3, SL1

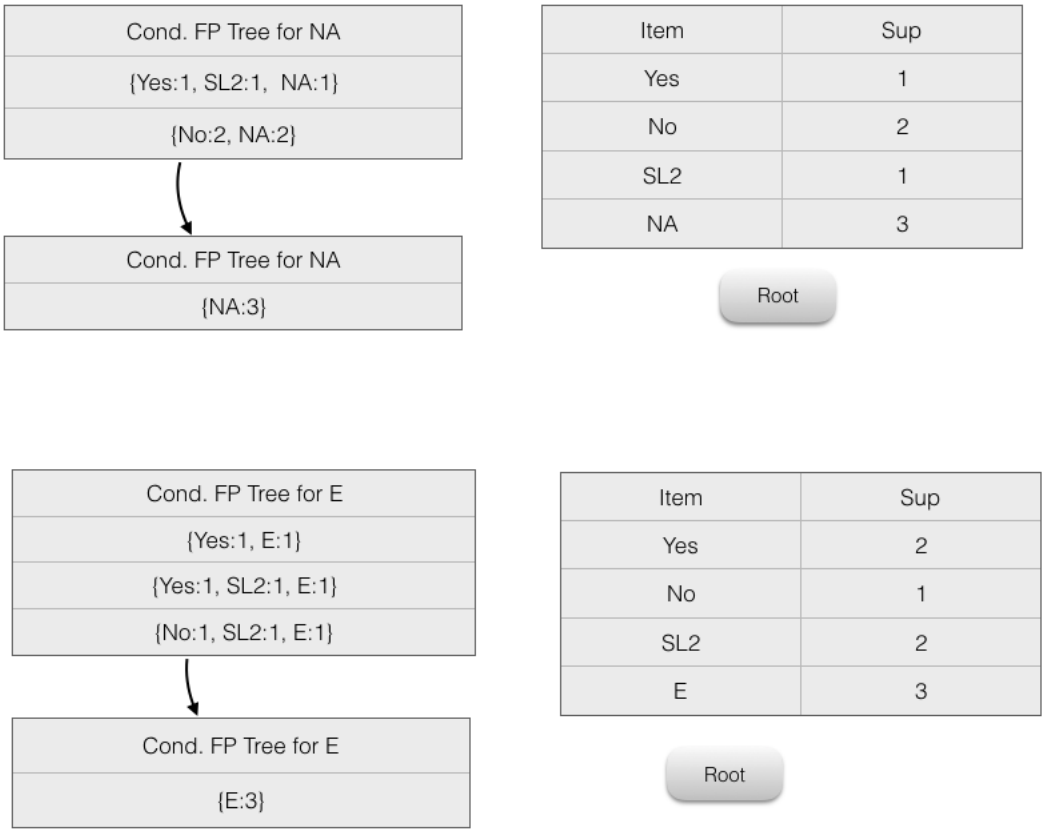

Figure 2.7: Find 2-itemset frequent patterns using FP-tree: Build Cond FP-tree for NA, E

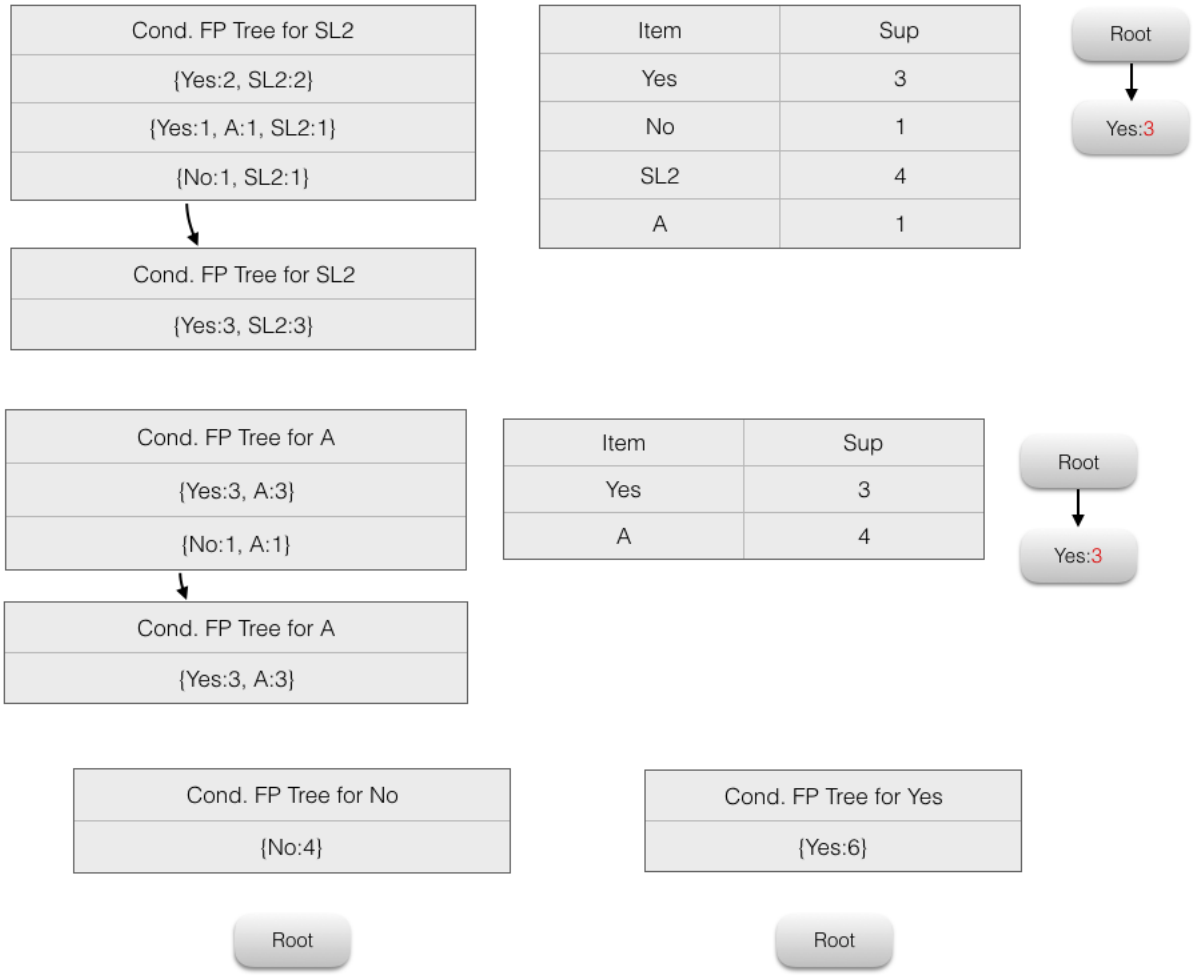

Figure 2.8: Find 2-itemset frequent patterns using FP-tree: Build Cond FP-tree for SL2, A, Yes, No

2-Itemset	Sup
{E, WP4}	3
{A, Yes}	3
{SL2, Yes}	3

1-Itemset	Sup
{NA}	3
{A}	4
{E}	3
{SL1}	3
{SL2}	4
{SL3}	3
{WP1}	3
{WP4}	3
{Yes}	6
{No}	4

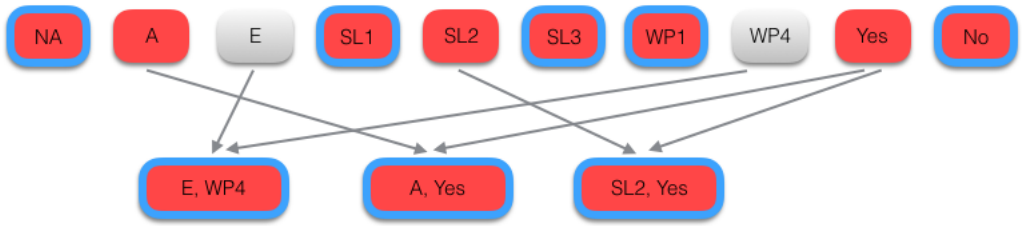

Figure 2.9: Maximal and Closed frequent patterns