CS 5525

Biplav Choudhury (906089826)

19 October 2020

1 Solutions to Assignment 4

1.1 Problem 1

Table 1: Results of hypothesis H1, H2 and H2

X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	-1	-1	-1	-1	-1	1	1
H1	1	1	1	-1	-1	-1	-1	-1	-1	-1
H2	-1	-1	-1	-1	-1	-1	-1	1	1	1
H3	1	1	1	-1	-1	-1	-1	-1	-1	1

Initial weights $w_i = 0.1$

For H1, there are 2 mistakes at X = 0.9 and 1 therefore error $err_1 = 0.2$.
For H2, there are 4 mistakes at 0.1, 0.2, 0.3, 0.8 therefore error $err_2 = 0.4$
For H3, there is 1 mistake at 0.9 therefore error $err_3 = 0.1$

$\alpha = \frac{1}{2} \log \frac{1-err}{err}$

$\alpha_1 = \frac{1}{2} \log \frac{1-0.2}{0.2} = 0.69$, $\alpha_2 = \frac{1}{2} \log \frac{1-0.4}{0.4} = 0.20$, $\alpha_3 = \frac{1}{2} \log \frac{1-0.1}{0.1} = 1.09$

New weights $w = 0.1 \times \exp(\alpha)$ for incorrect points and $w = 0.1 \times \exp(-\alpha)$ for correct points. For H1, the weights for the correct points are 0.05 and weights for the incorrect points are 0.199; for H2, the weights for the correct points are 0.08 and weights for the incorrect points are 0.12; for H3, the weights for the correct points are 0.03 and weights for the incorrect points are 0.297.

1.2 Problem 2

1.2.1 Part 1

If the items in X are independent, $P(X)$ can be written as $P(x_1)(x_2).... \times P(x_d)$. Therefore the cosine measure becomes

$$cosine(X) = \sqrt{P(x_1) \times P(x_2).... \times P(x_d)} \tag{1}$$

1.2.2 Part 1

For $X_1 = (a, b), cosine(X_1) = \sqrt{P(a) \times P(b)}$ and for $X_2 = (a, b, c), cosine(X_2) = \sqrt{P(a) \times P(b) \times P(c)}$. As $P(c) \leq 1, P(X_1) \geq P(X_2)$ which makes it non-monotonic.

1.3 Problem 3

1.3.1 Part 1

The answer to part 1 and part 2 will be provided together in part 2.

1.3.2 Part 2

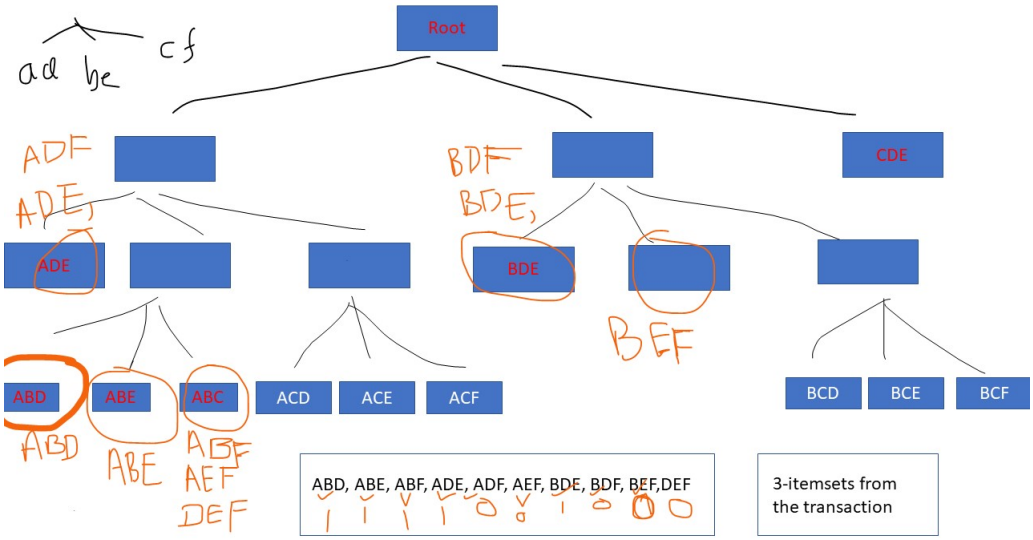

Figure 1: Hash Tree and counting of leaf nodes

As it can be seen from Fig. 1, we visit 5 valid leaf nodes on hashing the transaction.

1.3.3 Part 3

As the given item-sets are frequent, we can construct candidate 4 itemsets using a process similar to the one shown in Lec 9 slide 15 and is shown below -

- $abc \rightarrow abcd, abce, abcf$
- $abd \rightarrow abde, abdf$
- $abe \rightarrow abef$
- $acd \rightarrow acde, acdf$
- $ace \rightarrow acef$
- $acf \rightarrow$ null

- $ade \rightarrow adef$
- $bcd \rightarrow bcde, bcdf$
- $bce \rightarrow bcef$
- $bcf \rightarrow$ null
- $bde \rightarrow bdef$
- $cde \rightarrow cdef$

1.3.4 Part 4

Now to prune the above generated itemsets, we can see if any of the subsets of them are infrequent i.e. not present in the list of itemsets present in the question.

The following candidate 4 itemsets are pruned and the reason is shown below -

- $abcf \rightarrow abf$ is infrequent
- $abdf \rightarrow abf$ is infrequent
- $abef \rightarrow abf$ is infrequent
- $acdf \rightarrow cdf$ is infrequent
- $acef \rightarrow cef$ is infrequent
- $adef \rightarrow adf$ is infrequent
- $bcdf \rightarrow bdf$ is infrequent
- $bcef \rightarrow cef$ is infrequent
- $bdef \rightarrow def$ is infrequent
- $cdef \rightarrow def$ is infrequent

So the surviving candidate 4 itemsets are -
$abcd, abce, abde, acde, bcde$.

1.3.5 Part 5

Using a similar process where using the frequent 3 itemset we generated a frequent 4 itemset, is we have the transaction information on which of them are actually frequent, we can generate a frequent 5 itemset.

1.4 Problem 4

1.4.1 Part 1

If the minimum support is such that all of the itemsets in the transaction space are frequent, there is no maximal frequent itemset so the minimum number of maximal frequent frequent itemset is 0.
If the min support is such that either (a) all of the 1 and 2 itemsets are frequent and 3,4,5 itemsets are infrequent, the number of maximal frequent itemset is 10; or (b) all of the 1, 2 and 3 itemsets are frequent and 4,5 itemsets are infrequent, the number of maximal frequent itemset is 10; and this is the maximum possible number of maximal frequent is 10.

1.4.2 Part 2

For closed itemsets the idea is that all the immediate super set of any closed itemset should have lesser support. For the maximum case - if all of the itemsets appear the same number of times in the transaction, none of them will be closed so the minimum closed itemset is 0.

On the other hand if there is a case where the the support goes on reducing from 1-itemset to 4-itemset, such that A,B,C,D,E has a support of i, then AB,AC,AD,AE,BC.... has a support of $j < i$ and the support goes on decreasing as the size of the k-itemset becomes bigger, then each of the itemset becomes closed except the last item ABCDE as it doesn't have any superset. In such a case, the number of closed superset is 5+10+10+5 = 30 which is the maximum. The options didn't have 20 so possibly null set is included so the answer is 31.

1.4.3 Part 3

If any transaction that has AB will always contain C and E, it means that at least AB is not closed. Additionally, ABE and ABC are also not closed as ABC and ABE cannot appear in the itemset without ABCE; so both ABC and ABE are also not closed.

So in total, ABC, ABE and AB are not closed.

1.4.4 Part 4

Support of B, BCD being same means that list of transaction just contains BCD and not BC, BD and B separately. This also implies that support of B and (BC, BD) are same and support of BC and BD is same as BCD. So B, BC and BD are not closed.

1.5 Problem 5

1.5.1 b $\rightarrow c$

Table 2: b $\rightarrow c$

	c	$\hat{c}$	
b	3	4	7
$\hat{b}$	2	1	3
	5	5	

Support = $\frac{3}{10} = 0.3$, Confidence = $\frac{3}{7} = 0.428$

1.5.2 a $\rightarrow d$

Table 3: a $\rightarrow d$

	d	$\hat{d}$	
a	4	1	5
$\hat{a}$	5	0	5
	9	1	

Support = $\frac{4}{10} = 0.4$, Confidence = $\frac{4}{5} = 0.8$

1.5.3 b $\rightarrow d$

Table 4: b $\rightarrow d$

	d	$\hat{d}$	
b	6	1	7
$\hat{b}$	3	0	3
	9	1	

Support = $\frac{6}{10} = 0.6$, Confidence = $\frac{6}{7} = 0.8571$

1.5.4 e $\rightarrow c$

Table 5: e $\rightarrow c$

	c	$\hat{c}$	
e	2	4	6
$\hat{e}$	3	1	4
	5	5	

Support $= \frac{2}{10} = 0.2$, Confidence $= \frac{2}{6} = 0.3333$

1.5.5 c $\rightarrow a$

Table 6: c $\rightarrow a$

	a	$\hat{a}$	
c	2	3	5
$\hat{c}$	3	2	5
	5	5	

Support $= \frac{2}{10} = 0.2$, Confidence $= \frac{2}{5} = 0.4$.
In order of decreasing support, $b \rightarrow d, a \rightarrow d, b \rightarrow c, e \rightarrow c = c \rightarrow a$.
In order of decreasing confidence, $b \rightarrow d, a \rightarrow d, b \rightarrow c, c \rightarrow a, e \rightarrow c$

1.6 Problem 6

I have attached a jupyter notebook for this part. Accuracy is 0.8142 and f1 score is 0.7966