Assignment 4

11/12/2020

Name: Yusuf Elnady

Problem 1. Boosting

Point	1	2	3	4	5	6	7	8	9	10
$D_1(i)$	1	1	1	-1	-1	-1	-1	-1	1	1

$D_0(i) = 1/10$, as we have 10 points (initial equal weight)

$$\epsilon_t = \sum_{h_t(x_i)\neq y_i}^{M} D_t(i)$$

$$\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$$

H1: X $\leq$ 0.35 then Y = +1 , else Y = -1

Point (9) and (10) are missclassified as -1

$\epsilon_t = 1/10 * 2 = 0.2$, $\alpha_t = 0.693$

Points classified correctly: $D_1(i) = 0.0500$

Points misclassified : $D_1(i) = 0.19997$

Point	1	2	3	4	5	6	7	8	9	10
$D_1(i)$	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500	0.19997	0.19997

H2: X < 0.75 then Y = -1 , else Y = +1

Points (1),(2), and (3) are missclassified as -1, and point (8) is missclassified as +1

$\epsilon_t = 1/10 * 4 = 0.4$, $\alpha_t = 0.203$

Points classified correctly: $D_1(i) = 0.0816$

Points misclassified : $D_1(i) = 0.123$

Point	1	2	3	4	5	6	7	8	9	10
$D_1(i)$	0.123	0.123	0.123	0.0816	0.0816	0.0816	0.0816	0.123	0.0816	0.0816

H3: X $<$ 0.3 or X $\geq$ 0.95 then Y = +1 , else Y = -1

Point (9) and (3) are missclassified as -1

$\epsilon_t = 1/10 * 2 = 0.2$, $\alpha_t = 0.693$

Points classified correctly: $D_1(i) = 0.0500$

Points misclassified : $D_1(i) = 0.19997$

Point	1	2	3	4	5	6	7	8	9	10
$D_1(i)$	0.0500	0.0500	0.19997	0.0500	0.0500	0.0500	0.0500	0.0500	0.19997	0.0500

Problem 2. Cosine Similarity

P2.1

First, we have the support defined as

$$P(X) = \frac{of transactions contains X}{N}$$

, and we have

$$\cos{(X)} = \frac{X}{\sqrt{\prod_i^d P(x_i) * P(x_2) * \cdots * P(x_d)}}$$

, such that for example $\cos(\{a,b\}) = \frac{P(\{a,b\})}{\sqrt{P(a)*P(b)}}$

Now, if we consider all the items in x to be independent, that means

$$\cos(\{a,b\}) = \frac{P(\{a,b\})}{\sqrt{P(a)*P(b)}} = \frac{P(a)*P(b)}{\sqrt{P(a)*P(b)}} = \sqrt{P(a)*P(b)}$$

So, we can generalize the above cosine similarity function to any x_i items in the the itemset X as the following:

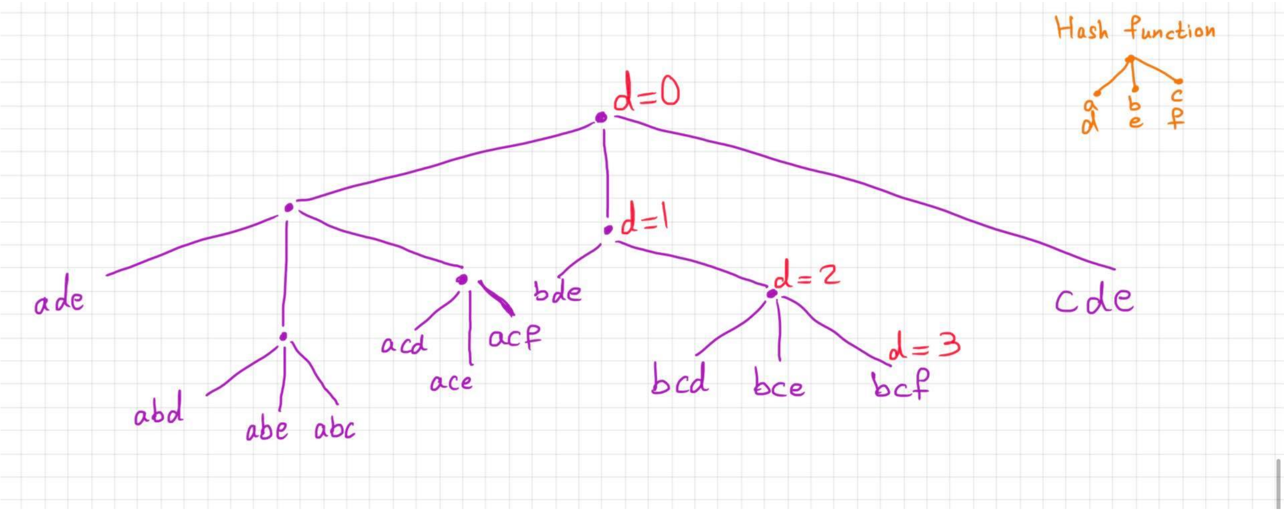
$$\cos(X) = \sqrt{\prod_i^d P(x_i) * P(x_2) * \cdots * P(x_d)}$$

P2.2 If we have for example $\cos(\{a,b\}) = 1/2$, then by adding another item to the itemset as $X = \{a,b,c\}$, then we may have $\cos(\{a,b,c\}) = 1/25$ and so on by increasing the number of items the support will be decreasing or staying the same. Because every $P(x_i)$ is a fraction and the denominator N is fixed, so we can conclude that the cosine is **non-increasing (anti-monotone)**.

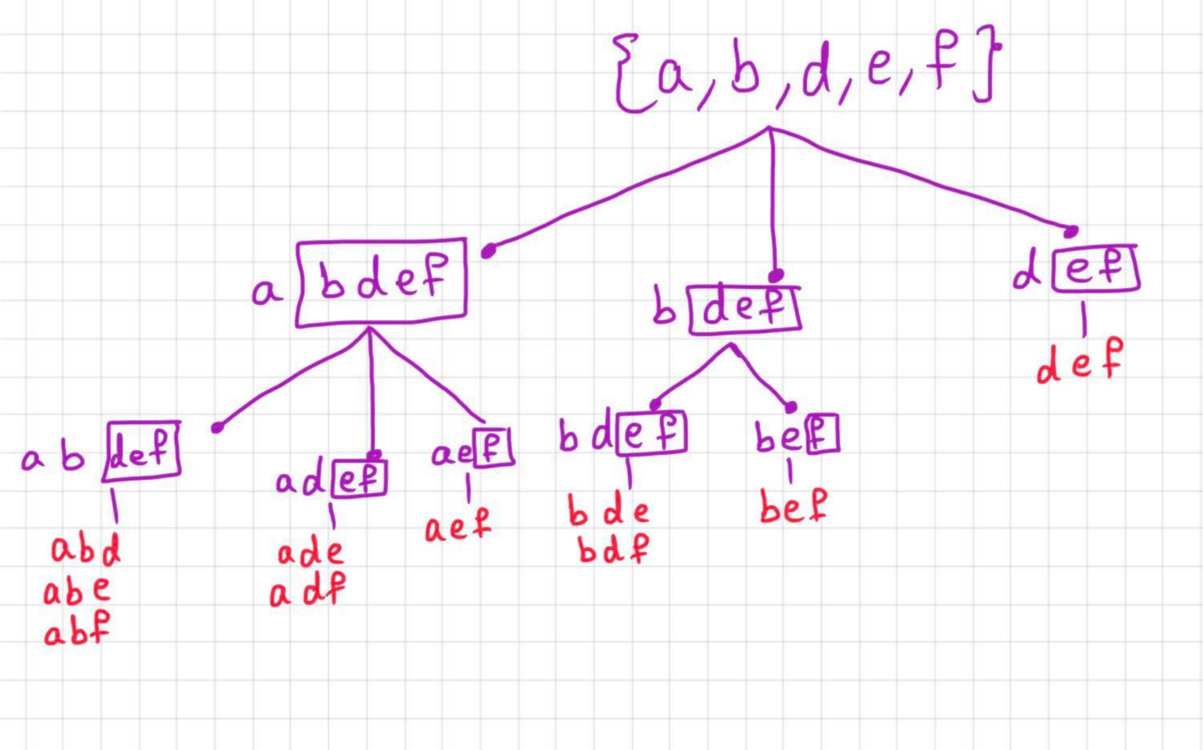
Problem 3. Apriori Algorithm

P3.1

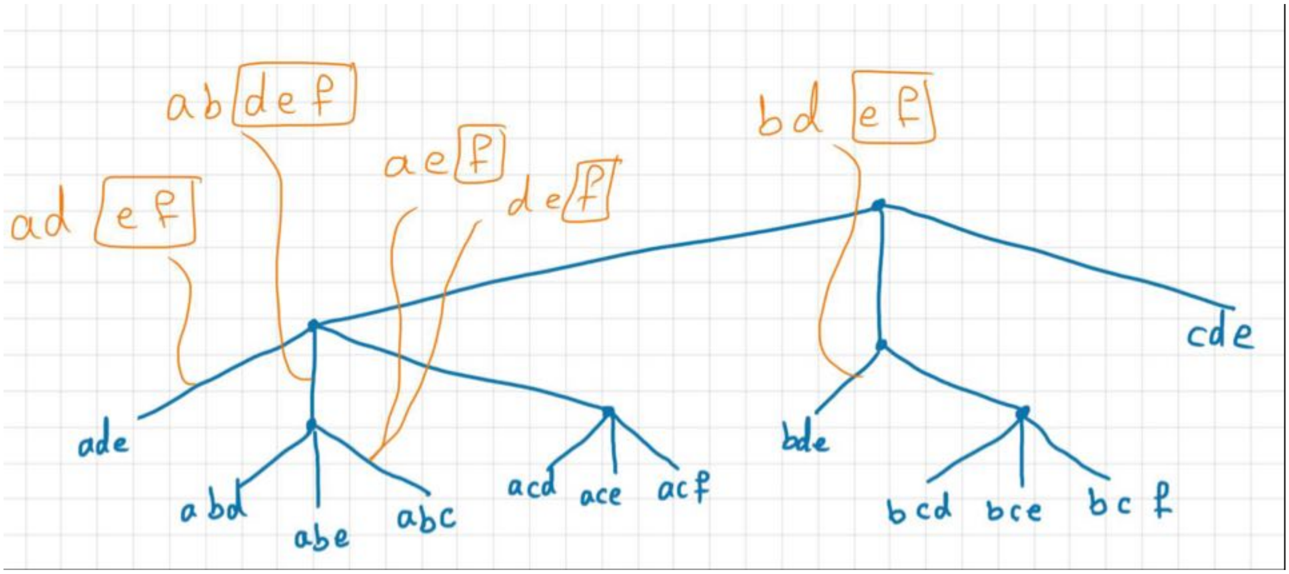
The hash tree:



P3.2 We can enumerate all the subsets of the transaction t=a,b,d,e,f as the following:



And the number of leaf nodes in the hash tree to which the transaction will be hashed into is **5 nodes**.



P3.3 All candidate 4-itemsets that can be generated from the frequent 3-itemsets using the candidate generation procedure for Apriori are: **{a,b,c,d},{a,b,c,e},{a,b,c,f}, {a,b,d,e},{a,c,d,e},{a,c,d,f}, {a,c,e,f}, {b,c,d,e},{b,c,d,f},{b,c,e,f}**

P3.4 The candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm are **{a,b,c,d},{a,b,c,e},{a,b,d,e},{a,c,d,e},{b,c,d,e}**

P3.5

The possible 5-itemsets that we can generate should not contain the item (f) as it doesn't appear in any of the survived 4-itemsets after pruning. That means we end up having only the items (a,b,c,d,e). Therefore the possible 5-itemset is {a,b,c,d,e}, and all of its subsets are frequent then we can say we can generate a frequent 5-itemset.

Problem 4. Maximal and Closed itemsets

P4.1

The **minimum number of maximal frequent itemsets is 0**, by assuming that all of the following are infrequent itemsets.

The **Maximum number of maximal frequent itemsets is 10**, if all itemsets of length-2 are frequent, but ones with length-3 are infrequent all of them.

P4.2

The **Minimum number of Closed frequent itemsets is 0**,by assuming that all of the following are infrequent itemsets.

The **Maximum number of Closed frequent itemsets is 31**, if we have a transaction database such that each item/node in the appear exactly once in the database. Then, the support for every depth will be the same, and different from the support of the next level.

P4.3

As every subset A,B will always have C,E appear with it in the transaction, that means **ab, abc, abcd, abe, abd, abde** are not closed frequent itemset.

P4.4

If we consider a transactionDB that has the following transactions: {bcd,bcd,bcd,a,c,d,e}, then we know that **b, bc, bd, cd** are not closed frequent itemsets.

Problem 5. Support and Confidence

P5.1

b $\rightarrow c$	c	$\overline{c}$
b	3	4
$\overline{b}$	2	1

Support: 3/10 = 0.3 **Confidence:** 3/7 = 0.429

a $\rightarrow d$	d	$\overline{d}$
a	4	1
$\overline{a}$	5	0

Support: 4/10 = 0.4 **Confidence:** 4/5 = 0.8

b $\rightarrow d$	d	$\overline{d}$
b	6	1
$\overline{b}$	3	0

Support: 6/10 = 0.6 **Confidence:** 6/7 = 0.857

e $\rightarrow c$	c	$\overline{c}$
e	2	4
$\overline{e}$	3	1

Support: 2/10 = 0.2 **Confidence:** 2/6 = 0.333

c $\rightarrow a$	a	$\overline{a}$
c	2	3
$\overline{c}$	3	2

Support: 2/10 = 0.2 **Confidence:** 2/5 = 0.4

P5.2

Ranking according to Support: $b \rightarrow d > a \rightarrow d > b \rightarrow c > e \rightarrow c = c \rightarrow a$

Ranking according to Confidence : b $\rightarrow d > a \rightarrow d > b \rightarrow c > c \rightarrow a > e \rightarrow c$

Problem 6. Text Classification

Please see the attached jupyter notebook.