Assignment 7 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. Let X, Y be two itemsets, and let $supp(X)$ denote the support of itemset X. Then the confidence of the rule $X \rightarrow Y$, denoted by $conf(X \rightarrow Y)$ is

   (a) $\frac{supp(X)}{supp(Y)}$

   (b) $\frac{supp(Y)}{supp(X)}$

   (c) $\frac{supp(X \cup Y)}{supp(X)}$

   (d) $\frac{supp(X \cup Y)}{supp(Y)}$

   (e) $\frac{supp(X \cap Y)}{supp(X)}$

   Sol. (c)
   Confidence measures the probability of seeing items in the consequent (RHS) of the rule given that we have observed items in the antecedent (LHS) of the rule in a transaction.

2. In identifying frequent itemsets in a transactional database, we find the following to be the frequent 3-itemsets: {B, D, E}, {C, E, F}, {B, C, D}, {A, B, E}, {D, E, F}, {A, C, F}, {A, C, E}, {A, B, C}, {A, C, D}, {C, D, E}, {C, D, F}, {A, D, E}. Which among the following 4-itemsets can possibly be frequent?

   (a) {A, B, C, D}

   (b) {A, B, D, E}

   (c) {A, C, E, F}

   (d) {C, D, E, F}

   Sol. (d)
   By the apriori property, only itemset {C, D, E, F} can possibly be frequent since all of its subsets of size 3 are listed as frequent. The other 4-itemsets cannot be frequent since not all of their subsets of size 3 are frequent. For example, for the first option, the itemset {A, B, D} is not frequent.

3. Let X, Y be two itemsets, $supp(X)$ denote the support of itemset X and $conf(X \rightarrow Y)$ denote the confidence of the rule $X \rightarrow Y$, denoted by $conf(X \rightarrow Y)$. Then lift of the rule, denoted by $lift(x \rightarrow Y$ is

   (a) $\frac{supp(X)}{supp(Y)}$

(b) $\frac{supp(X) \times supp(Y)}{supp(Y)}$

(c) $\frac{supp(X \cup Y)}{supp(X)}$

(d) $\frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$

(e) $\frac{supp(X \cap Y)}{supp(X) \times supp(Y)}$

Sol. (d)
The lift of a rule can be thought of as the ratio of the observed support to the support that would be expected if the antecedent and consequent were independent.

4. Consider the following transactional data.

Transaction ID	Items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Assuming that the minimum support is 2, what is the number of frequent 2-itemsets (i.e., frequent items sets of size 2)?

(a) 2
(b) 4
(c) 6
(d) 8

Sol. (c)
Candidate 1-itemsets:

itemset	support
{A}	6
{B}	7
{C}	6
{D}	2
{E}	2

Frequent 1-itemsets:

itemset	support
{A}	6
{B}	7
{C}	6
{D}	2
{E}	2

Candidate 2-itemsets:

itemset	support
{A, B}	4
{A, C}	4
{A, D}	1
{A, E}	2
{B, C}	4
{B, D}	2
{B, E}	2
{C, D}	0
{C, E}	1
{D, E}	0

Frequent 2-itemsets:

itemset	support
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, D}	2
{B, E}	2

5. For the same data as above, what are the number of candidate 3-itemsets and frequent 3-itemsets respectively?

   (a) 1, 1

   (b) 2, 2

   (c) 2, 1

   (d) 3. 2

   Sol. (b)
   Candidate 3-itemsets:

itemset	support
{A, B, C}	2
{A, B, E}	2

Frequent 3-itemsets:

itemset	support
{A, B, C}	2
{A, B, E}	2

6. Continuing with the same data, how many association rules can be derived from the frequent itemset {A, B, E}? (Note: for a frequent itemset X, consider only rules of the form S -¿ (X-S), where S is a non-empty subset of X.)

(a) 3
(b) 6
(c) 7
(d) 8

Sol. (b)
{A} → {B, E}
{B} → {A, E}
{E} → {A, B}
{A, B} → {E}
{A, E} → {B}
{B, E} → {A}

7. For the same frequent itemset as mentioned above, which among the following rules have a minimum confidence of 60%?

   (a) $A \wedge B \implies E$
   (b) $A \wedge E \implies B$
   (c) $E \implies A \wedge B$
   (d) $A \implies B \wedge E$

   Sol. (b), (c)
   The confidence values for the above four rules are respectively, 2/4, 2/2, 2/2, and 2/6. Hence, only rules in (b) and (c) have the minimum required confidence.

8. Suppose we are given a large text document and the aim is to count the words of different lengths, i.e., our output will be of the form - x words of length 1, y words of length 2, and so on. Assuming a map-reduce approach to solving this problem, which among the following key-value outputs would you prefer for the map phase? (Hint: consider the solution for the reduce part asked in the next question as well to ensure a complete algorithm to solve the problem.)

   (a) key - word, value - length (of corresponding word)
   (b) key - word, value - 1
   (c) key - length (of corresponding word), value - word
   (d) key - 1, value - word

   Sol. (c)
   Given a word, in the map phase we create a key-value pair, where the key is the length of the word and the value is the word itself.

9. For the above question, what would be the appropriate processing action in the reduce phase?

   (a) for each key which is a word, compute the sum of the values corresponding to this key
   (b) for each key which is a number, compute the lengths of the words in the corresponding list of values

(c) for each key which is a number, count the number of words in the corresponding list of values

Sol. (c)
In the reduce phase, all words of the same size will be available in the same reduce node. Thus, in each reduce node, counting the number of words in the list of values will give us the number of words of a particular length observed in the original document.

10. Let d_1 and d_2 be two distances according to some distance measure d. A function f is said to be (d_1, d_2, p_1, p_2)-sensitive if

    (a) if $d(x, y) \leq d_1$, then the probability that $f(x) = f(y)$ is at least p_1

    (b) if $d(x, y) \geq d_2$, then the probability that $f(x) = f(y)$ is at most p_2

    where $d(\cdot, \cdot)$ is a distance function. Given such a (d_1, d_2, p_1, p_2)-sensitive function, a better function (for use in locality sensitive hashing) would be one with

    (a) an increased value of p_1

    (b) a decreased value of p_1

    (c) an increased the value of p_2

    (d) a decreased the value of p_2

    Sol. (a), (d)
    Compared to the original function, if we can increase the value of p_1, we can ensure that if two points are close enough ($d(x, y) \leq d_1$), then the probability of a collision is higher. This is a desirable property when performing locality sensitive hashing. Similarly, if p_2 can be reduced it would indicate that given some separation between two points ($d(x, y) \geq d_2$), the probability of still observing a collision (which is undesirable) is reduced.