Q&A session – Week 38: Practice Questions with solutions

Q1: Assume that the largest frequent itemset that can be obtained by Apriori (for a given support threshold) is of *size k*:

A. How many passes does the Apriori algorithm need to execute in the worst case, before it terminates?

Apriori will make the first pass to identify all frequent itemsets of size 1. Then at each i_{th} pass it identifies the frequent itemsets of size i. At the kth pass it will hence find the itemsets of size k. Then it will make an extra pass by creating the candidates for size k+1, but it will eventually eliminate them. Hence, it will need to execute ***k+1 passes****.*

B. What is the smallest number of frequent itemsets that will be generated?

Since we have at least one itemset of size k, it means that Apriori will at least generate all the subsets of that itemset, i.e., 2^k -1.

Note that the i_{th} *pass* of the Apriori algorithm is a complete loop starting from candidate itemset generation, to pruning, support counting, and finally reporting the frequent itemsets of *size i*. We also assume that the *first pass* is completed when the *itemsets of size 1* are reported.

C. Consider the i_{th} *pass* of the Apriori algorithm, and suppose that a set of *n* frequent *i*-itemsets is found. If a transaction does not contain at least *x* frequent *i*-itemsets out of those found during this pass, then it is removed, since it will not contain a frequent itemset in any future pass.
What is the largest possible value of *x* that we can safely use to guarantee that we will never miss any frequent itemset?

Suppose you have found an i-itemset. This means that this itemset contains i items and it may lead to a frequent itemset of the next size (i+1). In order for an itemset of the next size (i+1) to be frequent, all the subsets of size i of that itemset should appear in the list of frequent itemsets found during the i_{th} pass. An itemset of size i+1 contains i+1 itemsets of size i. Hence, we the largest possible value of x is ***i+1****.*

Q2: Consider the following dataset

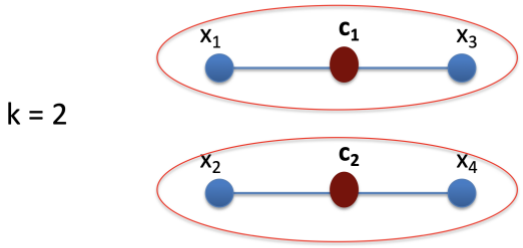
Example ID	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
E1	1	10	100	1000	5
E2	2	20	200	2000	81
E3	3	30	300	3000	33
E4	4	40	400	4000	172
E5	5	50	500	5000	-84

We want to perform dimensionality reduction using *PCA*. What is the *smallest number of principal components* we would need to achieve a reconstruction error of 0? Motivate your answer carefully.

We would need 1 PC on for attributes 1-4 (since these four attributes are perfectly correlated pairwise) and 1 for attribute 5. Hence, we would need two PCs.

Q3: In class, we discussed that *k-means* is sensitive to *initialization*. Provide an illustrative example of this claim. Briefly describe how *k-means++* solves this problem.

You may refer to slides 25-30 of lecture 5 to solve this.



k-means converges immediately!

Q4: Consider a large dataset that contains points in the 2-dimensional space. After running K-means 10 times, we realize that the produced clusterings are not exactly the same. This is expected due to the way K-means is designed. We would, however, like to come up with a final clustering. Describe a simple approach that we could use to decide on the final clustering of the data points using the 10 different clusterings produced by K-means.

One way is to use clustering aggregation. We can then use the BEST greedy approximation algorithm and identify the clustering with the smallest disagreement distance against all other clusterings.

Another way is to compute the Silhouette score (or the Dunn index) for each clustering and report the one with the highest score/index.

Q5: State the *Apriori principle* used by the Apriori algorithm. Does it work only for support or also for confidence?

If an itemset is frequent then all its subsets should be frequent. It only works for support only. For confidence it can be applied only for association rules built from the same itemset.

Q6: Assume we have an association rule:

r: {*buy_apples*, *buy_coffee*} → {*buy_oranges*}

(i) If the association rule has relative support equal to 0.5, that is *supp (r) = 0.5*, what does this imply about the occurrence of the buying apples, coffee, and oranges? **Explain using only <u>one sentence</u>!**

It means that 50% of the time (in 50% of the transactions) these three items are bought together.

(ii) if the rule above has a lift of 5, what does this say about the relationship between buying apples, coffee, and oranges? **Explain using only <u>one sentence</u>!**

It means that if someone buys apples and coffee then the chance of them buying oranges as well is five times more likely than expected (i.e., compared to the case that they are independent).

Recall from the lecture the definition of lift:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

Q7: Suppose you discover the following two association rules for a given set of transactions:

{A}→{B} and **{B}→{A}**

If the confidence of both rules is 0.9, what can you tell about the (relative) supports of items A and B? Motivate your answer carefully.

Recall from the lecture the definition of confidence:

confidence(X→Y) = supp(X, Y) / supp(X)

Math solution: This means that $supp(A,B) / supp(A) = supp(A,B) / supp(B)$

$\Rightarrow supp(A) = supp(B)$

Non-math solution: This means that the chance of seeing B given that we saw A is the same as that of seeing A given that we saw B. This chance is 90%. In other words, A and B occur together in the same transactions 90% of the time, and they have the same support.