Solutions: Assignment 2

Name: Sidhartha Satapathy
Net-id: ss46

Question 1

Part 1.

The number of cuboids in a full data cube is given by (where Li is the number of levels in the hierarchy for a dimension):

$$T = \prod_{i=1}^{n} (L_i + 1)$$

So, In our case there is no hierarchy which means the level is 1,
Hence the answer is 2 to the power 10 which is equal to **1024**
Ans. 1024

Part 2.

Assuming all of the values are different for each cell then we would have 3 * 2^{10} - 3 + 1 = 3070. The last 3 which we were considering for the apex cuboid would have to be removed as the apex cuboid will have a single cell.
Now from this we also need to remove all the base cells which total to 3.
So we have 3070 - 3 = 3067.
Finally, we need to consider that cuboids of the form (*,*,*,c4…,c10) and its parents like (*,*,*,*…,c10) will have only one cell and we have to subtract 2 for each such cuboid. As the apex has already been considered, we need to remove only 2*127 = 254 from 3067 which leaves us at 2813.
Ans. 2813

Part 3.

If we consider the iceberg condition, count > 2, then we only have to count cells that have a count of 3(max possible), which in our case would be all the cubes of the form (*,*,*,c4…,c10) and its parents like (*,*,*,*…,c10) as all the base cells merge for these and these sums upto 2 to the power 7 which is equal to **128.**
Hence the answer is **128**

Part 4.

Closed cell definition:
A cell c is closed if there exists no cell d, such that d is a descendant of c, and d has the same measure value as c.
Keeping this in mind we have (*,*,*,c4…,c10) as the closed cell with count = 3, (Note: (*,*,*,c4…,c10) and its parents have count = 3) as the other cells with count = 3 have a descendant with the same measure value (count = 3), and this cell has **7** non-star dimensions.
Answer. Hence, the answer is 7

Question 2.

Part 1.

$$T = \prod_{i=1}^{n} (L_i + 1)$$

So, In this case the location dimension has a hierarchy with two levels and the other three dimensions don't have a hierarchy.

By using the formula we have, 2*2*2*3 = **24.**
Ans. 24

Part 2.

We count the number of different tuples for the tuple (Location(city), Category, Price, Rating) as **48.** By different tuple we mean tuple which have a different value for at least one of the values within the tuple, in this case it means different value for at least one of Location(city), Category, Price, Rating.
Ans. 48

Part 3.

After drilling up, we count the number of different tuples for the tuple (Location(state), Category, Price, Rating) as **34.** By different tuple we mean tuple which have a different value for at least one of the values within the tuple, in this case it means different value for at least one of Location(state), Category, Price, Rating.
Ans. 34

Part 4.

We count the number of different tuples for the tuple (*, Category, Price, Rating) as **23.** By different tuple we mean tuple which have a different value for at least one of the values within the tuple, in this case it means different value for at least one of Category, Price, Rating.
Ans. 23

Part 5.

We count the tuples that have the value of Location(state) = “Illinois”, Price=”moderate”, Rating = 3, without caring about category and the number sums upto **2.**
Ans. 2

Part 6.

We count the tuples that have the value of Location(city) = “Chicago”, Category=”food” without caring about others and the number sums upto **2.**
Ans. 2

Question 3:

Note : frequent patterns means a set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set

Support means frequency or the number of occurrences of an itemset. (Sup(X) = number of occurrences of X)

Part a. 1.

With minimum support 20, which means the number of occurrences of an itemset to be greater than equal to 20, the number of frequent patterns are **30.** We count this by considering all possible subsets of items possible and checking the counts. We can use various algorithms to optimize this as well.
Ans. 30

Part a. 2.
We count the number of frequent patterns with length 3 which means patterns which have 3 elements and the number sums upto **8.**
Ans. 8

Part a. 3.

Max Pattern:
A pattern X is a max-pattern if X is frequent and there exists no frequent super-pattern YɔX. First of all we add patterns in an incremental manner with respect to size, that is patterns of size 1 first , then size 2 and so on. So in order to calculate this whenever we add a pattern we check if its subset is already contained and if it does, then we remove the subset. We would never encounter a problem of checking if this patterns superset is already present as that can't be the case given the manner in which we are considering patterns of different sizes.

The number of max patterns with support 20 is 7.

Part b. 1.

With minimum support 10, which means the number of occurrences of an itemset to be greater than equal to 10, the number of frequent patterns are **55.** We count this by considering all possible subsets of items possible and checking the counts. We can use various algorithms to optimize this as well.
Ans. 55

Part b. 2.

We count the number of frequent patterns with length 3 which means patterns which have 3 elements and the number sums upto **20.**
Ans. 20

Part b. 3.

A pattern X is a max-pattern if X is frequent and there exists no frequent super-pattern YɔX. First of all we add patterns in an incremental manner with respect to size, that is patterns of size 1 first , then size 2 and so on. So in order to calculate this whenever we add a pattern we check if its subset is already contained and if it does, then we remove the subset. We would never encounter a problem of checking if this patterns superset is already present as that can't be the case given the manner in which we are considering patterns of different sizes.
The number of max patterns with support 10 is 6.

Part b. 4.
Confidence: The conditional probability that a transaction containing X also contains Y. It is calculated as c =sup(X U Y) / sup(X)
For (C,E) -> A, the confidence is **0.679**

Part b. 5.
For (A,B,C) -> E, the confidence is **0.742**

Source for the images:

Professor Jiawei Han’s Slides.