Machine, Data and Learning (CS7.301)

Spring 2022, IIIT Hyderabad
07 Mar, Monday (Lecture 12)

Taught by Prof. Vikram Pudi

Overview of Data Analytics

The field of data analytics originated from the database community.

Data mining is the automated extraction of interesting patterns from large databases. A pattern may be an association, a clustering, or a classification.

Association Rules

Association rules have two main metrics: *support* and *confidence*. If we have an association rule $X \to Y$, we define

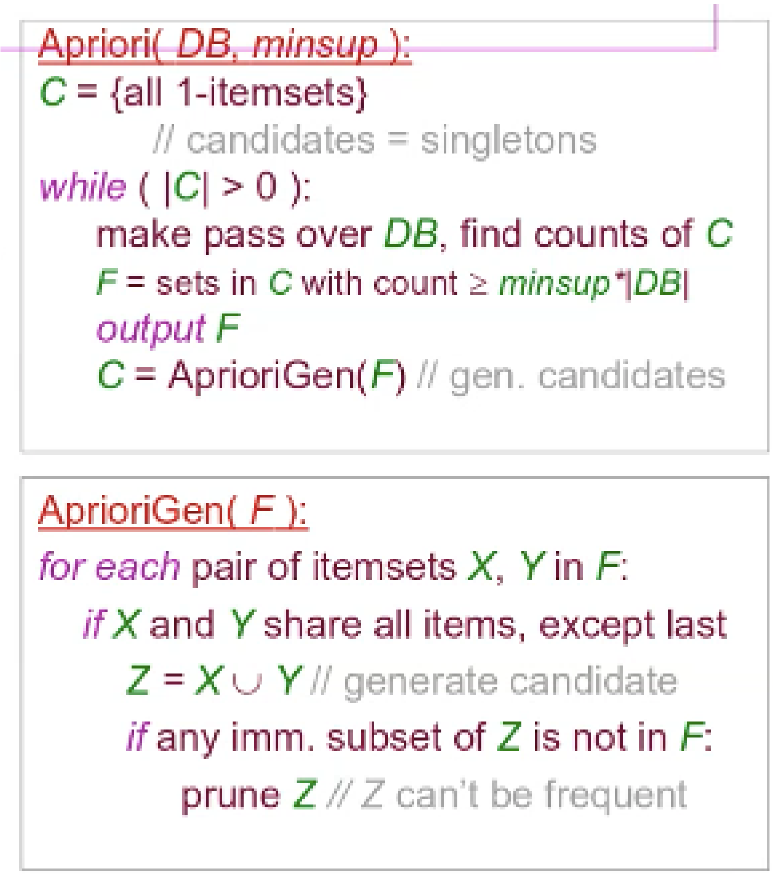
$$\text{support} = \frac{\text{transactions satisfying } X \to Y}{|D|}$$

and

$$\text{confidence} = \frac{\text{support}(X \to Y)}{\text{support}(X)}.$$

An algorithm that helps to find these rules is the *a priori* algorithm. It relies on the idea that an itemset can be frequent only if all its subsets are frequent. It starts by creating a set of singleton itemsets and finding their counts. It filters the set of itemsets by checking for counts greater than the minimum support and combines the resultant elements, iteratively.

There are different types of association rules – boolean, hierarchical, or categorical rules are some of these. They may also be cyclic, constrained or sequential.



```
Apriori( DB, minsup ):
C = {all 1-itemsets}
        // candidates = singletons
while ( |C| > 0 ):
    make pass over DB, find counts of C
    F = sets in C with count ≥ minsup*|DB|
    output F
    C = AprioriGen(F) // gen. candidates
```

```
AprioriGen( F ):
for each pair of itemsets X, Y in F:
   if X and Y share all items, except last
      Z = X ∪ Y // generate candidate
      if any imm. subset of Z is not in F:
          prune Z // Z can't be frequent
```

Figure 1: The *A Priori* Algorithm