Pattern Mining

A unsupervised learning technique that **finds rules that describe the data**.

Association Rules

The formulation of the association rule problem is as follows:

- Given a **set of transactions** T of **items** I, and a pair of **minimum support** *minsup* and **minimum confidence** *minconf* thresholds, find all the rules $A \rightarrow B$ such that:
  - $A \cap B = \emptyset$;
  - $A \cup B = I$;
  - $sup(A \rightarrow B) = P(A \cap B)$;
  - $conf(A \rightarrow B) = P(B|A)$.
- The **support** of a rule is the **percentage of transactions** that contain the items in the rule - contains A and B.
- The **confidence** of a rule is the **percentage of transactions** that contain the items in the rule and also contain the consequent - contains B given A.

The **discovery** runs in two steps:

1. Find all the **frequent itemsets**, i.e., all the sets of items that appear in **at least** *minsup* **transactions**;
2. From the frequent itemsets, find all the **association rules** that have **at least** *minconf* **confidence**.

   Anti-monotonicity - if a pattern is frequent, all its subsets are frequent; or if a pattern is infrequent, all its supersets are infrequent.

Maximal and Closed Patterns

- A **(maximal) pattern** is a maximal frequent itemset, this means, not within another pattern;
- A **closed pattern** is a frequent itemset that is not a subset of another frequent itemset with the **same support**.

Sequential Patterns

- A **sequential pattern** is a pattern that appears in a sequence of events **- sequential data**;

Drawbacks

- The number of patterns discovered depends on the minimum support and confidence thresholds, and if they are too low, the number of patterns can be too high and the patterns can be **too specific**;
- On the other hand, if the thresholds are too high, the number of patterns can be too low and the patterns can be **too general**.

Assessment

- **Support** measures the **usefulness** of a pattern;
- **Confidence** measures the **certainty** of a pattern;
- **Lift** measures the **correlation** (**interestingness**) of a pattern:
  - The pattern is **interesting** if $lift > 1$;
    - **The farther from 1, the more interesting the pattern is.**;

$$lift(A \rightarrow B) = \frac{sup(A \rightarrow B)}{sup(A) \times sup(B)} = \frac{P(A \cap B)}{P(A) \times P(B)}$$