Affinity analysis

“The pattern of the prodigal is: rebellion, ruin, repentance, reconciliation, restoration.”—Edwin Louis Cole

Ponder about the last time you made an impulse purchase. Maybe you were waiting in the grocery store checkout lane and bought a pack of candies or gums. Perhaps on a late-night trip for some item and you also picked up a six-pack of beer. You might have even paid for this workshop on a whim on your friend’s recommendation. These impulse purchases are no coincidence, as retailers use sophisticated data science techniques to identify patterns that will drive customer behavior.

In past years, such recommendation systems were based on the subjective intuition of marketing professionals and inventory managers or buyers. However in the recent years, as barcode scanners, computerized inventory systems, and online shopping trends have built a wealth of transactional data, machine learning has been increasingly applied to learn purchasing patterns. The practice is Affinity analysis or commonly known as market basket analysis due to the fact that it has been so frequently applied to supermarket data. Although the technique originated with shopping data, it is useful in various other contexts as well.

By the time you finish this session, you will be able to apply market basket analysis techniques to your own tasks, whatever they may be. Generally, the work involves:

- Using simple performance metrics to find associations in large datasets

- Understanding the peculiarities of transactional data
- Knowing how to identify the useful and actionable patterns

The results of a these are actionable patterns. As such when we apply the technique, you are likely to identify applications to your own work.

To understand this we need to study a type of machine learning known as Association rule/pattern mining.
The classical problem of association pattern mining is defined in the context of supermarket data containing sets of items bought by customers, which are referred to as transactions. The goal is to determine associations between groups of items bought by customers, which can intuitively be viewed as k-way correlations between items. The most popular model for association pattern mining uses the frequencies of sets of items as the quantification of the level of association. The discovered sets of items are referred to as large itemsets, frequent itemsets, or frequent patterns. The association pattern mining problem has a wide variety of applications:

1. **Supermarket data**: The supermarket application was the original motivating scenario in which the association pattern mining problem was proposed. This is also the reason that the term itemset is used to refer to a frequent pattern in the context of super-market items bought by a customer. The determination of frequent itemsets provides useful insights about target marketing and shelf placement of the items.

2. **Text mining**: Because text data is often represented in the bag-of-words model, frequent pattern mining can help in identifying co-occurring terms and keywords. Such co-occurring terms have numerous text-mining applications.

3. **Generalization to dependency-oriented data types**: The original frequent pattern mineing model has been generalized to many dependency-oriented data types, such as time-series data, sequential data, spatial data, and graph data, with a few modifications. Such models are useful in applications such as Web log analysis, software bug detection, and spatiotemporal event detection.

4. **Other major data mining problems**: Frequent pattern mining can be used as a subroutine to provide effective solutions to many data mining problems such as clustering, classification, and outlier analysis.

Because the frequent pattern mining problem was originally proposed in the context of market basket data, a significant amount of terminology used to describe both the data and the output is borrowed from the supermarket analogy. From an application-neutral perspective, a frequent pattern may be defined as a frequent, defined on the universe of all possible sets.

Frequent itemsets can be used to generate association rules of the form X ⇒ Y , where X and Y are sets of items. A famous example of an association rule, which has now become part of the data mining folklore, is {Beer} ⇒ {Diapers}. This rule suggests that buying beer makes it more likely that diapers will also be bought. Thus, there is a certain directionality to the implication that is quantified as a conditional probability. Association rules are particularly useful for a variety of target market applications. For example, if a shop owner discovers that {Eggs, M ilk} ⇒ {Yogurt} is an association rule, he or she can promote yogurt to customers who often buy eggs and milk. Alternatively, the shop owner may place yogurt on shelves that are located in proximity to eggs and milk.

The frequency-based model for association pattern mining is very popular because of its simplicity. However, the raw frequency of a pattern is not quite the same as the statistical significance of the underlying correlations. Therefore, numerous models for frequent pattern mining have been proposed that are based on statistical significance. This sesion will also explore some of these models.

we are going to explore some of the important terminologies with respect to affinity analysis and also try out the ideas with an example.