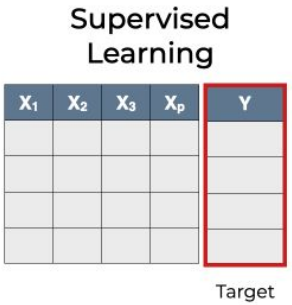
Unsupervised Learning

(Review) Types of machine learning

Supervised

In supervised learning, we have several data points or samples, described using predictor variables or features (X) and a target variable or **label** (Y).

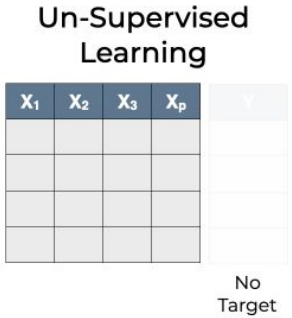


Example

1. Spam/not spam
2. Stock price (actual/prediction)
3. Flower classification

Unsupervised

Uncovering hidden patterns and structures from **unlabeled data**.



Example

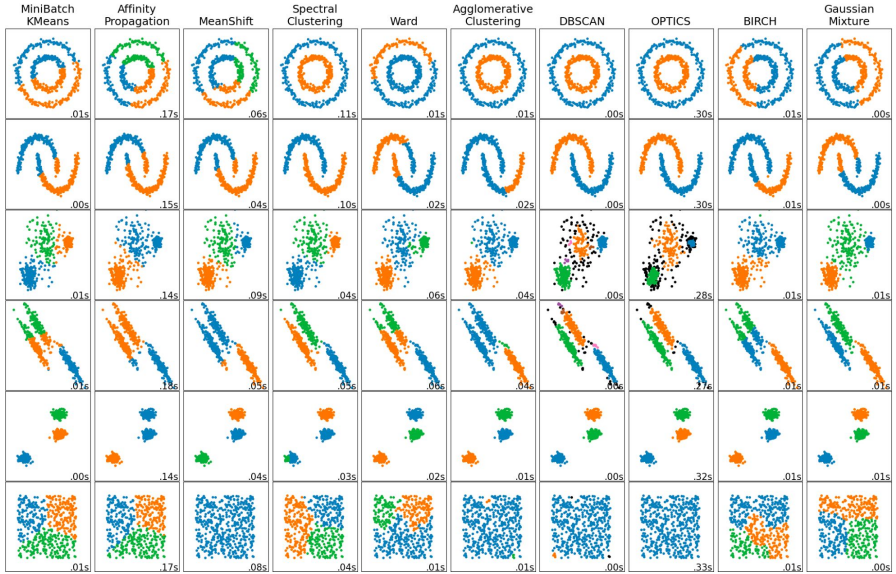
1. Customer segmentation
2. Dimension reduction
3. Feature selection

Labeled vs Unlabeled

Agenda

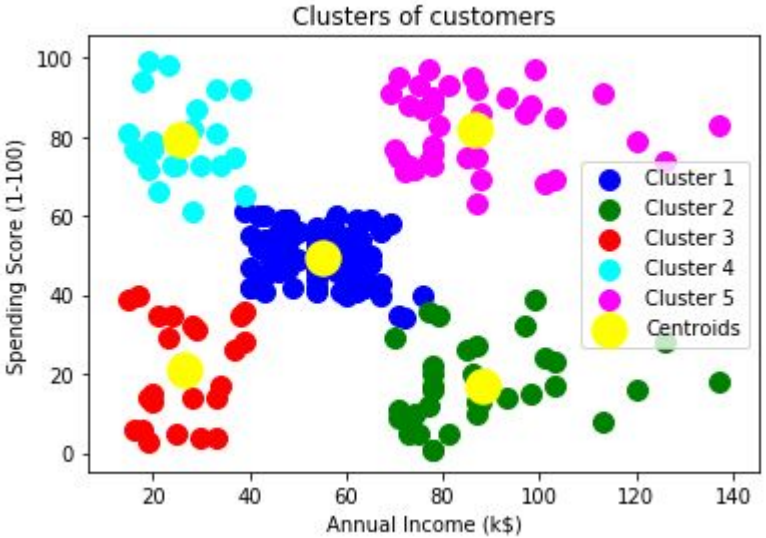
1. Clustering
    a. KMeans
    b. DBSCAN
2. Association rules
    a. Apriori
3. Dimensionality reduction
    a. PCA: Principal Component Analysis
    b. LDA: Linear Discriminant Analysis

Clustering



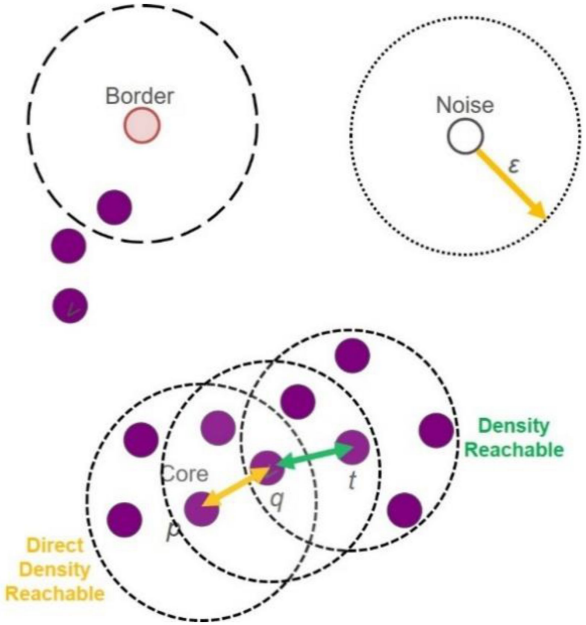
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

Kmeans



- K-means searches for a predetermined number of clusters within an unlabelled dataset by using an iterative method to produce a final clustering based on the number of clusters defined by the user (represented by the variable K).
- In K-means, each cluster is represented by its center (called a “centroid”), which corresponds to the arithmetic mean of data points assigned to the cluster.
- A centroid is a data point that represents the center of the cluster (the mean), and it might not necessarily be a member of the dataset.

DBSCAN

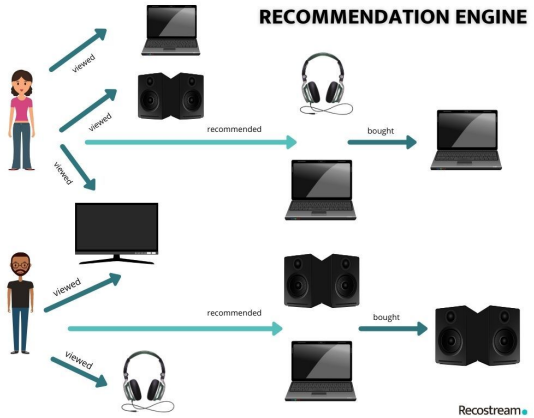


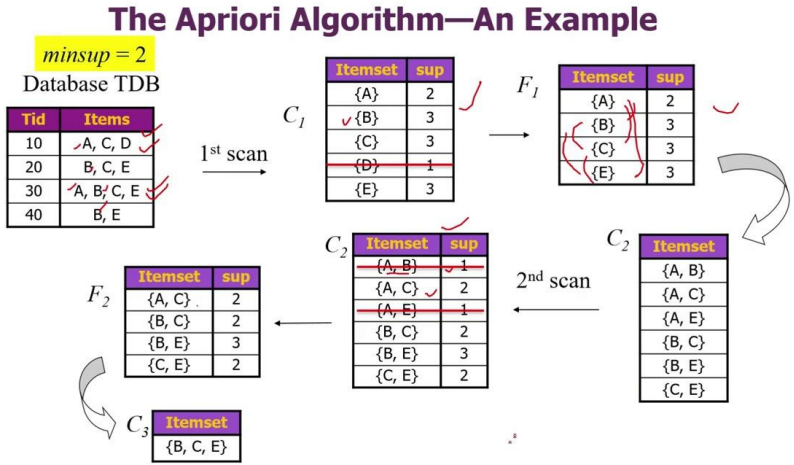
DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density. It groups 'densely grouped' data points into a single cluster.

Association Rule

Apriori

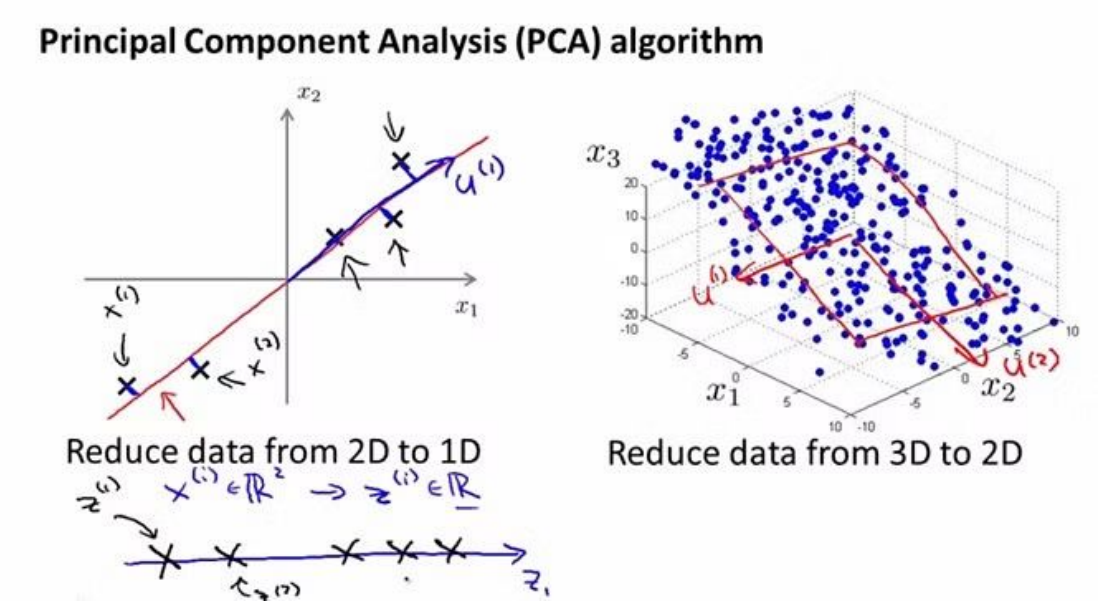
- Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules.
- Apriori algorithm operates on a database containing a huge number of transactions.
- Apriori algorithm helps the customers to buy their products with ease (recommendation system)





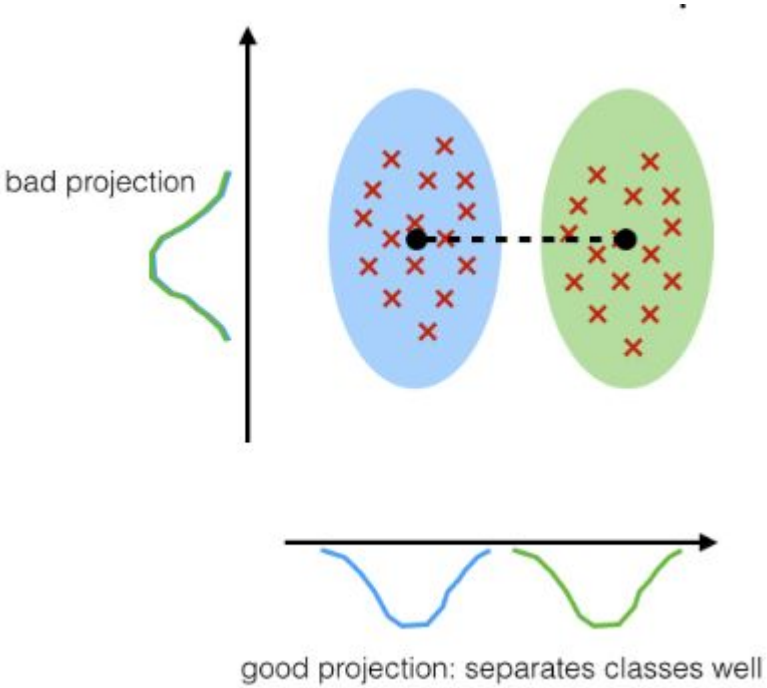
Dimensionality Reduction

PCA: Principal Component Analysis



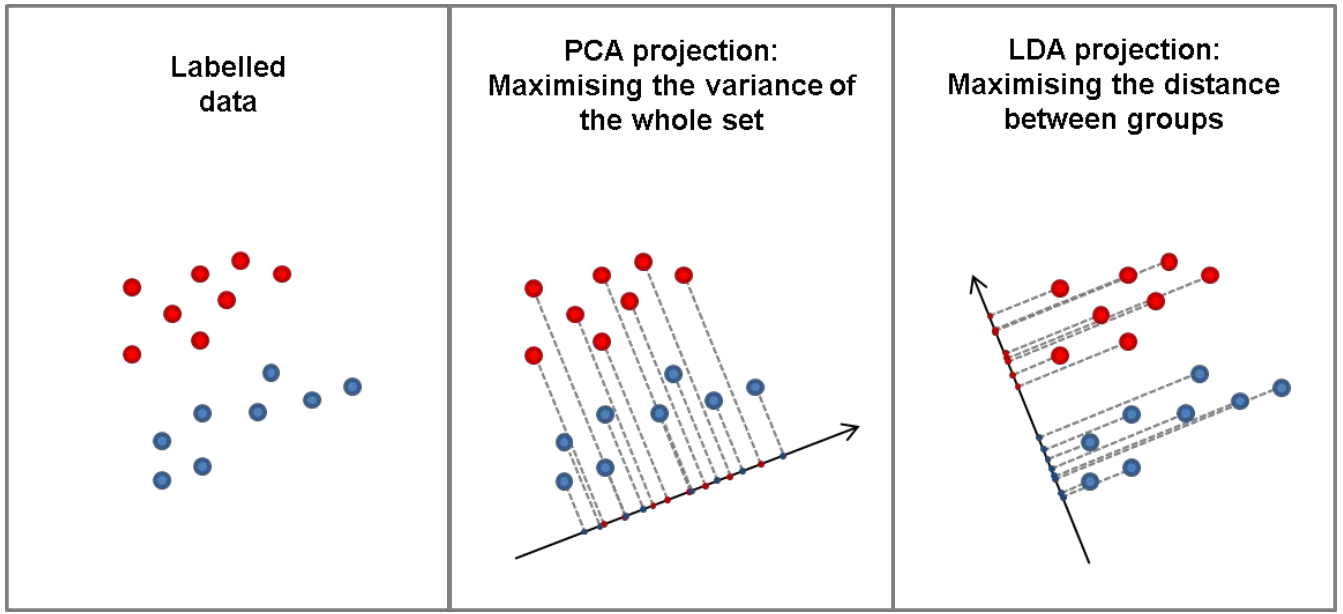
- Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns.
- It does this by transforming the data into fewer dimensions, which act as summaries of features.

LDA: Linear Discriminant Analysis



LDA pick a new dimension that gives maximum separation between means of projected classes and minimum variance within each projected classes.

PCA vs LDA


Labelled data
PCA projection: Maximising the variance of the whole set
LDA projection: Maximising the distance between groups