Class 7 Unsupervised Learning and K-Means Clustering

Dr. Wei Miao

UCL School of Management

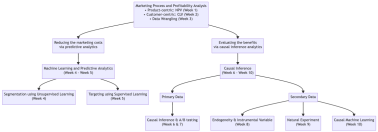
October 23, 2024

Section 1

Overview of Predictive Analytics

Roadmap for Predictive Analytics

- The core of any business decision is **profitability analysis** (BEQ, NPV, CLV). To increase firm profitability,
  1. Increase revenue
  2. Reduce costs (CAC or variable marketing costs)
  3. Reduce customer churn
- In Weeks 4 and 5, we will learn how to utilize **predictive analytics** to improve profitability. Correspondingly,
  1. Develop customers through ML recommender systems
  2. **Reduce costs by targeting more responsive customers**
  3. Predict customer churn and take preventive actions

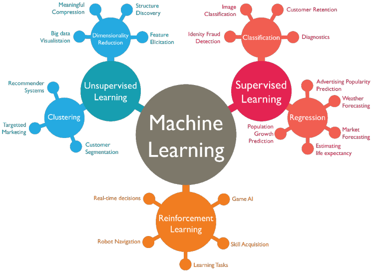


Types of Predictive Analytics

- Unsupervised Learning
  - Only observe X => Want to uncover unknown subgroups
- Supervised Learning
  - Observe both X and Y => Want to predict Y for new data

In Term 2, you will learn predictive analytics models systematically. By then, think about how those techniques can be applied back to these case studies.

Types of Predictive Analytics



Learning Objectives for Today

- Understand the concept of unsupervised learning
- Understand how to apply K-means clustering and find the optimal number of clusters
- How to apply clustering analyses for customer segmentation for M&S

Section 2

K-Means Clustering

K-Means Clustering

- K-means clustering is one of the most commonly used unsupervised machine learning algorithms for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified by the analyst.
- For data scientists: It can classify customers into multiple segments (i.e., clusters), such that customers within the same cluster are as **similar** as possible, whereas customers from different clusters are as **dissimilar** as possible.
- Input: (1) customer characteristics; (2) the number of clusters
- Output: cluster membership of each customer

Similarity and Dissimilarity

- The clustering of observations into groups requires computing the (dis)similarity between each pair of observations. The result of this computation is known as a dissimilarity or distance matrix.
- The choice of similarity measures is a critical step in clustering.
- The most common distance measures are the Euclidean distance (the default for K-means) and the Manhattan distance.

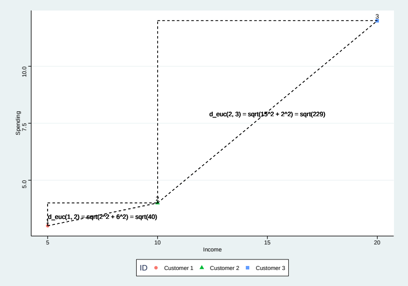
Euclidean Distance

- The most common distance measure is the Euclidean distance.

$$d_{\text{euc}}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Example of Income and Spending for 3 customers
    - $Income = (5, 10, 20)$
    - $Spending = (3, 4, 12)$
- Euclidean distance
    - $d_{\text{euc}}(2, 1) = \sqrt{(5-3)^2 + (10-4)^2} = \sqrt{2^2 + 6^2} = \sqrt{40}$
    - $d_{\text{euc}}(2, 3) = \sqrt{(5-20)^2 + (10-12)^2} = \sqrt{15^2 + 2^2} = \sqrt{229}$

Visualization of Euclidean Distance



Manhattan Distance

- Another common distance measure is the Manhattan distance, which is less commonly used because the absolute value function is not differentiable.

$$d_{\text{man}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- Example of Income and Spending for 3 customers
  - $Income = c(5, 10, 20)$
  - $Spending = c(3, 4, 12)$
- Distance
  - $d_{\text{man}}(2, 1) = |5 - 3| + |10 - 4| = 2 + 6 = 8$
  - $d_{\text{euc}}(2, 3) = |5 - 20| + |10 - 12| = 15 + 2 = 17$

K-Means Clustering: Step 1

(a)

- Raw data points; each dot is a customer
- X and Y axis are customer characteristics, say, income and spending
- Obviously there should be 2 segments
- Let's see how K-means uses a data-driven way to classify customers into 2 segments

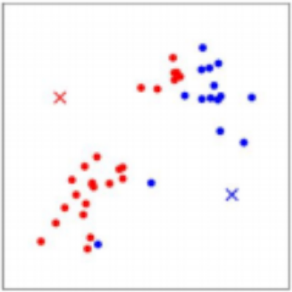
K-Means Clustering: Step 2

(b)

- We specify 2 segments
- K-means initializes the process by **randomly** selecting 2 centroids

Due to this randomness, different starting points may yield varying results. We need to reinitialize the process repeatedly to ensure **robustness** of results.

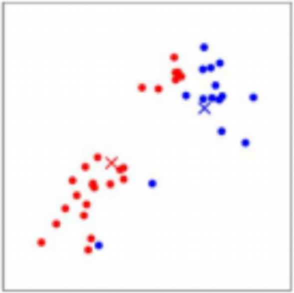
K-Means Clustering: Step 3



(c)

- K-means computes the distance of each customer to the red and blue centroids
- K-means assigns each customer to red segment or blue segment based on which centroid is closer

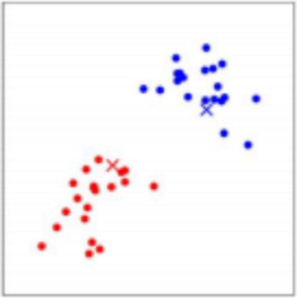
K-Means Clustering: Step 4



(d)

- K-means updates the new centroids of each segment
- The red cross and blue cross in the picture are the new centroids
- We still see some “outliers”, so the algorithm continues

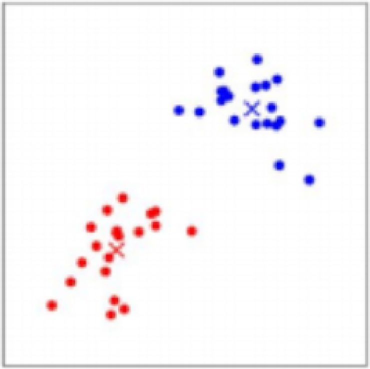
K-Means Clustering: Step 5



(e)

- K-means computes the distance of each customer to the red and blue centroids
- K-means updates each customer to red segment or blue segment based on which centroid is closer
- Now the outliers are correctly assigned each segment

K-Means Clustering: Step 6



(f)

- K-means updates the new centroid from the previous clustering
- K-means computes the distance of each customer to the new centroids
- K-means finds that all customers are correctly assigned to their nearest centroids, so the algorithm does not need to continue
- We say, the algorithm **converges**, and the algorithm stops

After-Class Readings

- More technical details: K-means Cluster Analysis