Intro to Unsupervised ML

TM Quest

Overview

What Will we Learn in This Module?

- What does it mean for a model to be unsupervised?
  - What is the different between supervised and unsupervised?
  - Why use unsupervised learning?
  - When use unsupervised learning?
- What is the Kmeans clustering model?
  - How does the Kmeans clustering work?

What is Unsupervised Learning?

Supervised VS. Unsupervised

Supervised Learning

We have features and targets.

Unsupervised Learning

We have features.

Example (Clustering your customer group)

- Want to understand your customers better.
- Divide them into groups based on behavior.
- Can use your better understanding of the different groups to tailor your marketing.

Unsupervised Learning

When to use Unsupervised Learning?

- When the labels are unavailable.
    - Impossible/illegal/hard/expensive to get.
    - To slow to get for the task.
    - Unknown what the labels should be.

Unsupervised Tasks

- Clustering
- Outlier/Anomaly Detection
- Recommender Systems

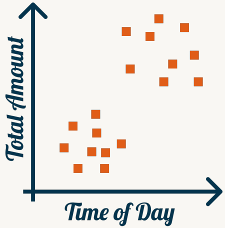
Clustering

Clustering

- Clustering models can both be supervised (e.g., KNN) and unsupervised.
- In unsupervised learning we need to do the clustering without the labels.
- We will learn more about the k-means model later.

Example (Clustering)

- A retailer knows the following:
  - email-address,
  - the time of day,
  - total amount.
- Want to make custom promotions based on this information.



Outlier/Anomaly Detection

Outlier/Anomaly Detection

- Outlier detection is finding outliers in the system.
- Assumes that there are more normal data points than outliers.
- Can benefit from some supervised data.

Example

- Spam filters: Outliers—Spam mails
- Fraud detection: Outliers—Fraudulent transactions
- Find mistakes in the system: Outliers—Mistakes
- Detect cyber attacks in your system: Outliers—Attacks

Recommender System

Recommender System

- Recommender systems are systems that give the user recommendations on what to do next.
- On smaller systems, it is often based on rules rather than machine learning.

Example

- Recommending the next thing to read/watch (YouTube/Netflix/TikTok),
- Recommending additional wares in an online store (Amazon).
- Recommending further information (your bank/state/forum).

K-Means Clustering

K-Means Clustering

Output Clusters

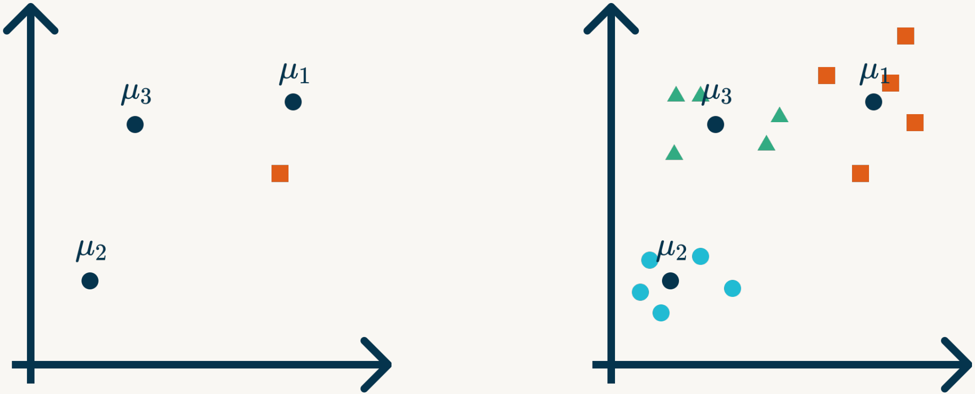
- The algorithm finds the mean points $\mu_1, \dots, \mu_k$.
- Each mean μ_i gives us a corresponding cluster S_i of data points.
  - S_i consists of all data points that are closer to μ_i than any other mean point.
- Additionally, the mean points satisfy

$$\mu_i = \frac{1}{n_i} \sum_{x \in S_i} x,$$

where n_i is the number of points in S_i.

3-Means Clustering

Example



How to Find the Mean Points?

Given a cluster S_i with n_i points, define its mean by

$$\mu_i = \frac{1}{n_i} \sum_{x \in S_i} x.$$

Of all the ways to divide the points into k-clusters $S_1, \ldots, S_k$ the k-means algorithm tries to minimize the quantity

$$\sum_{i=1}^{k} \frac{1}{n_i} \sum_{x \in S_i} d(x, \mu_i)^2 = \sum_{i=1}^{k} \mathrm{Var}(S_i)$$

where $d(x, \mu_i)$ is the distance between x and μ_i.

The k-means algorithm tries to simultaneously minimize how much the clusters spread out.