Week13-1: Unsupervised Learning: K-Means Clustering

Ekarat Rattagan

July 28, 2025

Unsupervised Learning

Given a training set $= \{x_1, x_2, \ldots, x_n\}$

Application

1. Market segmentation (Clustering): **K-MEANS**
2. Social network analysis
3. Astronomical data analysis

Clustering Approaches

1. **Partition-based clustering**
   - k-means (Mean of data points)
   - k-medoids (Actual data point)
2. **Density-based clustering**
   - DBSCAN

Cost Function

Cost function:

$$J(C_1, C_2, \ldots, C_m, \mu_1, \ldots, \mu_K) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \mu_{C_i}\|_2^2$$

Where:

- C_i: index of cluster $1, 2, \ldots, K$ to which sample x_i is currently assigned.
- μ_K: centroid of cluster K
- μ_{C_i}: centroid of cluster to which sample x_i has been assigned

K-means Algorithm

Dataset: $\{x_1, \ldots, x_5\}$, Let $K = 2$

- $C_1 = 1$, $C_2 = 1$, $C_3 = 2$, $C_4 = 2$, $C_5 = 2$
- $\mu_{C_1} = \mu_1, \mu_{C_2} = \mu_1, \mu_{C_3} = \mu_2, \mu_{C_4} = \mu_2, \mu_{C_5} = \mu_2$

Steps:

1. Initialize K
2. Initialize $\mu_1, \mu_2, \ldots, \mu_K$ (centroids)
3. Repeat until centroids do not change:
    1. For $i = 1$ to N, assign each x_i to closest cluster centroid c_i
    2. For $k = 1$ to K, update μ_k as mean of x_i assigned to cluster k

Issues in K-means

- What is the best K?
- What is the best μ_K?

Initial Centroids

Local Optima

- Different initial centroids lead to different cluster outcomes.
- Might converge to local minima.

Solution 1: Random Initialization

- For $i = 1$ to ∞
  - Randomly initialize K-means
  - Run K-means, get $C_1, \ldots, C_m, \mu_1, \ldots, \mu_K$
  - Compute J
- Pick clusters that gave lowest J
- **Slow & Unstable**, not guaranteed global optimum

Solution 2: K-means++ Initialization

- Take μ_1 uniformly at random from x_i
- Take μ_k with probability:

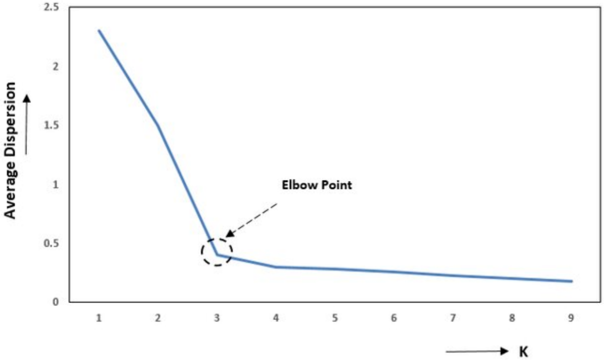
  $$\frac{D(x)^2}{\sum D(x)^2}$$

  where $D(x)$ is shortest distance from data point to closest μ_k already chosen
- Repeat until K centroids are chosen

Elbow Method

Solution to choose best K:

- Plot J vs. number of clusters K
- Find the *elbow point* where the decrease levels off



Silhouette Analysis

$$s(x_i) = \begin{cases} 1 - \frac{a(x_i)}{b(x_i)} & \text{if } a(x_i) < b(x_i) \\ 0 & \text{if } a(x_i) = b(x_i) \\ \frac{b(x_i)}{a(x_i)} - 1 & \text{if } a(x_i) > b(x_i) \end{cases}$$

- $a(x_i)$: average distance to other points in same cluster
- $b(x_i)$: minimum average distance to other clusters

Additional Materials

- https://developers.google.com/machine-learning/clustering
- https://github.com/ekaratnida/Applied-machine-learning/blob/master/Week14-kmeans/K-means.ipynb
- https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf

Reference

- Hamerly,G. and Elkan,C.(2003). Learning the k in k-means. Advances in neural information processing systems, 16.