Foundations of Machine Learning
AI2000 and AI5000

FoML-25
Unsupervised Learning - Clustering

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad

July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data-driven Intelligence & Learning Lab

So far in FoML

- Intro to ML and Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Supervised learning
    a. Linear Regression with basis functions
    b. Bias-Variance Decomposition
    c. Decision Theory - three broad classification strategies
    d. Neural Networks

Unsupervised Learning

For today

- Unsupervised Learning
  - Introduction, contrasting with supervised, challenges
- Clustering
  - K-Means

Some of the contents are taken from - [Intro to Statistical Learning]

So far

- Supervised learning techniques
  - p features $X_1, X_2, X_3, \ldots.. X_p$ measured on N observations
  - Response Y also measured on these
  - → goal is to predict Y using $X_1, X_2, X_3, \ldots.. X_p$

Unsupervised learning

- Only have a set of features X_1, X_2, X_3, ….. X_p
- Not interested in prediction (don't have an associated Y)
- → goal is to discover "Interesting things" about the data

Unsupervised learning

- "Interesting things" about the data
  - Is there an informative way to visualize the data?
  - Can we discover 'subgroups' among the variables or samples?

Unsupervised learning

- A diverse set of statistical techniques for answering such questions
    - Clustering
    - Dimensionality Reduction - Principal Component Analysis (PCA)

Unsupervised learning - challenges

- Much more challenging than supervised
- Exercise is 'subjective'
    - No simple goal
    - More like an 'exploratory analysis'
    - No universally accepted method for performance evaluation/validation (no true answer as in the case of supervised setting)

ML problems

	Supervised	Unsupervised
Discrete	Classification	Clustering
Continuous	Regression	Dimensionality Reduction

Clustering

Clustering

- Most widely used technique for exploratory data analysis
  - Computational biologists cluster genes (on the basis of similarities in their expression)
  - Retailers cluster their customers (based on their profiles)
  - Astronomers cluster stars (on the basis of spatial proximity)
  - Textile manufacturers cluster customers into size groups (based on their body type/measurements)

Clustering

- Task of grouping a set of objects, such that
  - Similar objects end up in the same group
  - Dissimilar objects are separated into different groups

Clustering

- Task of grouping a set of objects, such that
  - Similar objects end up in the same group
  - Dissimilar objects are separated into different groups
- Imprecise/ambiguous
  - It's not clear how to come up with a more rigorous definition
  - E.g., 'similarity' is not transitive, where as 'cluster sharing' is

0.5 $A \sim B$ 1

$B \sim C$ 1.1

$\not\Rightarrow A \sim C$ < 0.1

Clustering - Objectives

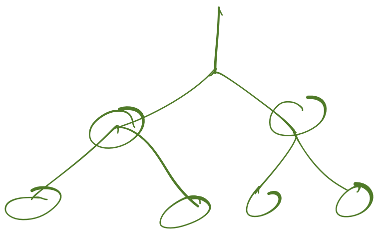
- Discover/Understand the underlying structure of the data
- What subpopulations exist in the data?
    - How many?
    - What are their size?
    - Do the elements in a subpopulation have common properties?
    - Are there outliers in the data?
    - etc.

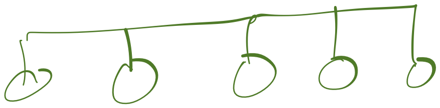
Clustering - Taxonomy

1. Based on the overlap of clusters
    a. Hard clustering - no overlap, complete/single assignment
    b. Soft clustering - strength of association between element and cluster

Clustering - Taxonomy

2. Based on methodology
    a. Flat versus Hierarchical - set of groups vs. taxonomy
    b. Density based versus Distribution based - DBSCAN vs. GMMs





Clustering

- Finding groups of objects such that
  - the objects in a group will be similar (or related) to one another, and
  - different from (or unrelated to) the objects in other groups

Clustering

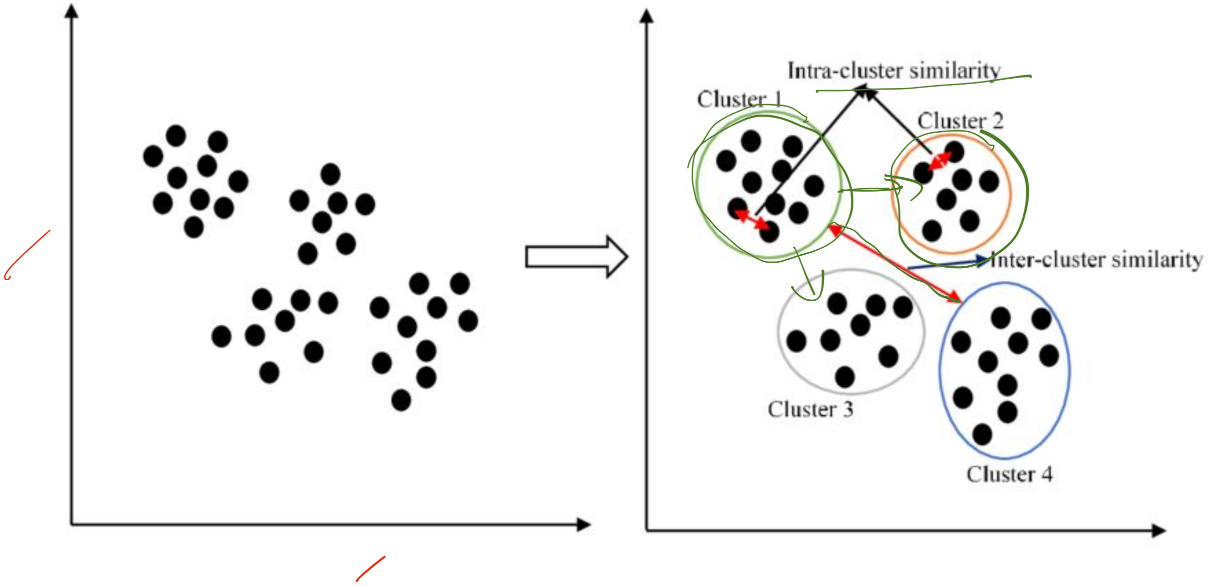


Figure credits - [Absalom El-Shamir Ezugwu]

Clustering methods

- K-Means
- Hierarchical
- GMM
- Evaluation of clustering methods

K-Means

K-Means

- Simple and elegant
- Partitional clustering algorithm
- Non-overlapping (hard) clustering
    - Assigns each element to exactly one cluster
- Must specify the number of clusters - K

K-Means

- Can be posed as an intuitive mathematical problem
- C_i denotes the set of indices of the samples belonging to i-th cluster

training data sample

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}.$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'.$$

K-Means

- Idea - good clustering results in small ‘within cluster variation’ $W(C_k)$
    - Within Cluster Sum of Squares (WCSS)

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}.$$

K-Means

- Need to define - $W(C_k)$
- Most common - Squared Euclidean distance

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

K-Means

- Combining the two equations

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}.$$

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

K-Means

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

WCSS ↓

BCSS ↑

- This minimizes WCSS
  - → Maximizes the 'Between the Clusters Sum of Squares (BCSS)'
  - Why?
  - Total variance in the data is constant - minimizing the WCSS → maximizing BCSS
  - This is related to the 'law of variance' in probability theory

K-Means Algorithm

- Formally, the objective becomes
  - Why/How?

$$\underbrace{\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}_{C_k} = \left\{ 2 \underbrace{\sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2}_{} \right\}$$

WCSS

mean / centroid

$\bar{x}_k$

where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$

Centroid

cluster center of 'k'

K-Means

- Let's find an algorithm to achieve this
- How many different ways of assigning N samples to K clusters?
    - K^N

K-Means Algorithm

clusters index to which the sample belongs

label

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   updating

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   μ_k

   $k = 1$ to K

   Assignment

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

x

It is guaranteed to decrease the objective value!

WCSS

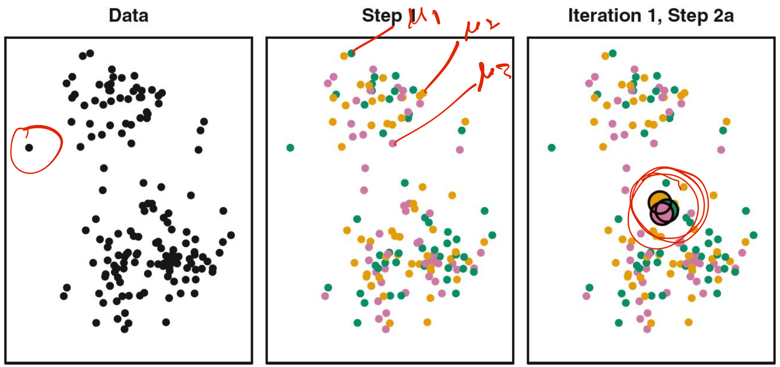
K-Means Algorithm

- With runs, the clustering obtained will continually improve until no change → local optimum is reached
  - Why?

{1 to k}

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 \qquad \bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$$

K-Means - Visual Example



$\{\mu_1, \mu_i, \mu_K\}$

x_1

x_2

x_N

μ_i

$\{x_i\} \rightarrow \mu_i^1 \;\ldots\; \mu_i^K$

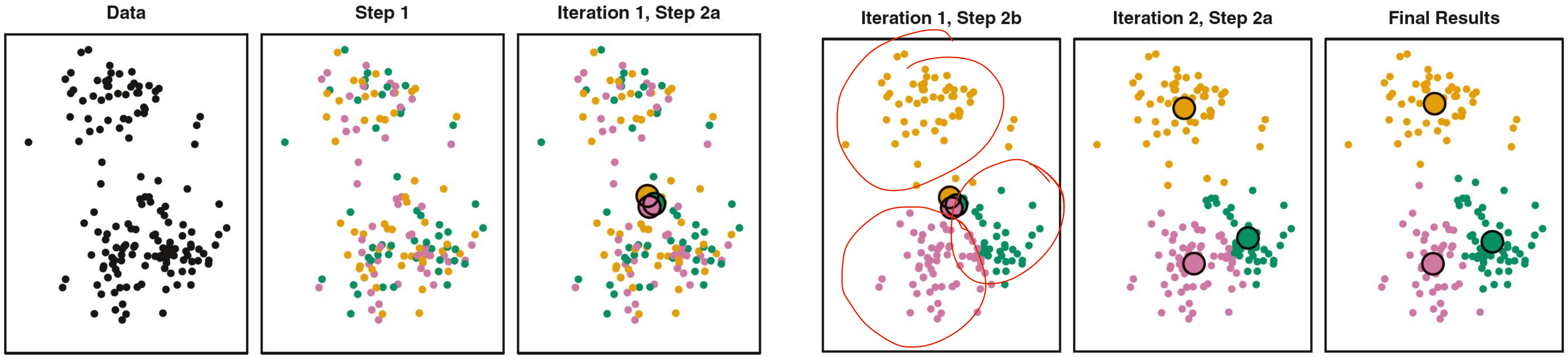
భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

K-Means - Visual Example

$x_i \to \mu_i$

$\mu_i^{t+1} \mu_i^{t}$



update

assign

U

A

K-Means

- Because it finds a local minimum
  - Solution depends on the initial clustering
- Run for multiple initializations → pick the best clustering
  - One with minimal objective function

K-Means

- Need to know the 'K' value
  - Not simple
- Complexity
  - NP-hard problem
  - The heuristic algorithms have a complexity of O(NKdi)
    - i - iterations until convergence

Next class

- Other clustering
  - Hierarchical
  - GMM
- Dimensionality Reduction
  - PCA

Rough Work

$$\mathrm{Var}(X) = \underbrace{E\left[\mathrm{Var}(X/C)\right]}_{\text{WCSS}} + \underbrace{\mathrm{Var}\left(E(X/C)\right)}_{\text{BCSS}}$$

a. updating C_k

b. Assigning labels

$C = \{1, 2, .., k\}$

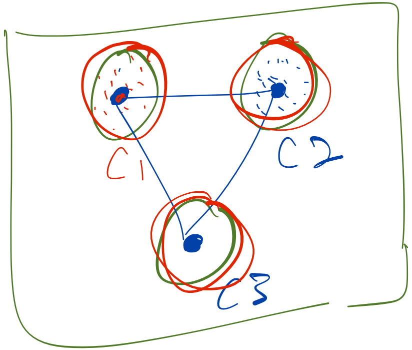
mean of samples in

$x_i \to C_l$

$x_i \to \boxed{C_m}$

$WCSS_m$

$\leq WCSS_l$



WCSS $\downarrow$

BCSS $\uparrow$

Rough Work

$\{x_i\}_1^N$

$\arg\min_{\mu} \sum_{i=1}^{N} \| x_i - \mu \|^2$

$\mu = \text{mean of } \{x_i\}$

$\arg\min \sum_{i=1}^{K} WCSS_i$

$|X_i - c_k|$

$|X_i - C_m|^2$

$\arg\min_{k} |X_i - C_k|^2$