

Basic Concept

1. Data Cleaning

- **Denoise Data:** Remove noise in datasets, if applicable.
- **Remove Outliers:** Identify and eliminate anomalies.
- **Handle Missing Data:** Address incomplete datasets.
- **Remove Duplicates:** Ensure unique data entries.
- **Categorical Data Encoding:** Convert categorical variables into numerical formats.
- **Data Normalization:** Standardize data across features.

2. Data Exploration

- **Summary Statistics:**
 - **Location:** Mean, median.
 - **Spread:** Range, variance, percentiles.
 - **Frequency:** Mode.
 - **Visualization:**
 - **Histograms:** Represent data distributions.
 - **Box Plots:** Visualize spread and outliers.
-

1. What is Clustering?

Clustering Analysis:

- Group objects such that:
 - **Intra-cluster differences** are small (within a cluster).
 - **Inter-cluster differences** are large (between clusters).

Applications:

1. Understanding:

- Stand-alone tool to gain insights into data distributions.
- Pre-processing step for other algorithms.
- Examples:
 - Grouping related documents for browsing.
 - Grouping genes or proteins with similar functionality.
 - Grouping stocks with similar price fluctuations.

2. Summarization:

- Reduces the size of large datasets.
 - Preserves privacy (e.g., in medical data).
-

General Applications:

1. Cluster Items:

- Better organization.
- Faster searching.

2. Cluster People:

- Patients: Different treatments for different groups (e.g., children vs. elderly).
- Customers: Different groups with different needs, enabling product optimization.

Biological Applications:

1. Cluster Genes:

- Identify co-expressed genes involved in the same pathway.
- Identify differentially expressed genes related to diseases.

2. Cluster Samples/Cells:

- Identify new disease sub-types.
- Identify new cell types.

3. Similarity and Dissimilarity

Similarity:

- **Definition:** Numerical measure of how similar two objects are.
- **Range:** Typically $[0, 1]$.
- **Higher values** indicate more similarity.

Dissimilarity (Distance):

- **Definition:** Numerical measure of how different two objects are.
 - **Range:** Minimum = 0 (completely alike).
 - **Larger values** indicate more dissimilarity.
-

4. Similarity and Distance Measurements

1. Cosine Similarity:

- Measures the cosine of the angle between two vectors d_1 and d_2 :

$$\text{Cosine Similarity} = \frac{d_1 \cdot d_2}{|d_1| \cdot |d_2|}$$

- Example: Applied to text data or high-dimensional data.

2. Correlation:

- Measures the linear relationship between objects.
- Correlation coefficient ranges from -1 to 1:
 - +1: Perfect positive correlation.
 - -1: Perfect negative correlation.

- 0: No correlation.

3. Euclidean Distance:

- Straight-line distance between two points in multi-dimensional space:

$$d(x,y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

- **Normalization** is necessary if scales differ across dimensions.

4. Minkowski Distance:

- Generalization of Euclidean distance:

$$d(x,y) = \left(\sum_{k=1}^n |x_k - y_k|^p \right)^{\frac{1}{p}}$$

- $p=1$: Manhattan distance (City Block distance).
- $p=2$: Euclidean distance.
- $p \rightarrow \infty$: Chebyshev (maximum) distance.

5. Mahalanobis Distance:

- Considers data distribution by incorporating the covariance matrix:

$$d(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$$

- Example: Used in multivariate datasets where features are correlated.

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

↓ Min-max normalization

Person	Height	Weight
P1	0.625	0.875
P2	0	0
P3	0.25	0.375
P4	1	1

5. Hierarchical Clustering

Definition:

- Produces nested clusters organized as a hierarchical tree (dendrogram).
- **Steps:**
 1. Compute the similarity or distance matrix.
 2. Treat each data point as a cluster.
 3. Merge the two closest clusters.
 4. Update the similarity or distance matrix.
 5. Repeat until a single cluster remains.

Distance Matrix Updates:

- **Methods:**
 - Min (single linkage).
 - Max (complete linkage).
 - Group average.
 - Centroid-based distance.
-

6. Running Example

Gene Expression Data:

A dataset of genes and their expression levels under different conditions:

Gene	WT	Mutant 1	Mutant 2	Mutant 3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5

Gene WT Mutant 1 Mutant 2 Mutant 3

At4g27450 1.5 1 1 1.5

At2g34930 10 25 23 15

At2g05540 1 1 2 1

Steps:

1. Compute the similarity matrix (e.g., correlation).
2. Treat each gene as a cluster.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770	1				
At1g30720	0.9733	1			
At4g27450	-1	-0.9733	1		
At2g34930	0.9493	0.9909	-0.9493	1	
At2g05540	0.5774	0.562	-0.5774	0.4528	1

3. Merge clusters based on the largest correlation or smallest distance.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733 ->-0.9493			
At2g34930	0.9493 ->0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528 ->0.562	

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9493			
At2g34930	0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.562	

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-1	-0.9493			
At2g34930	->-0.9493	-0.9493	-0.9493		
At2g05540	0.5774	0.562 ->0.5774	-0.5774	0.562 ->0.5774	

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.5774	-0.5774	0.5774	

4. Repeat until only one cluster remains.

		At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
Node2	At4g35770					
	At1g30720					
Node1	At4g27450	-0.5774	-0.5774			
Node3	At2g34930			-0.5774		
	At2g05540			-0.5774		

7. Applications of Clustering

1. Gene Clustering:

- Identify co-expressed or differentially expressed genes.

2. Grouping Patients:

- Discover disease sub-types or new patient groups.

3. Dimensionality Reduction:

- Simplify high-dimensional datasets while preserving key patterns.