What is clustering in Machine Learning?

Clustering is an unsupervised learning technique that groups similar data points into clusters, works with unlabeled data

How is it different from supervised learning (regression/classification)?

Supervised learning requires more effort to label data but can achieve high accuracy for specific prediction tasks, whereas clustering is exploratory and helps find structure or groupings within unlabeled data

Give one real-life example of clustering and one of supervised learning?

clustering

1. Customer Segmentation
2. Streaming Service Recommendations
3. Disease Management in Healthcare

supervised learning

1. Image Classification
2. weather
3. Credit Scoring

2. Clustering Algorithms

Describe the following algorithms:

1. K-Means
2. Hierarchical Clustering
3. DBSCAN

For each algorithm, explain:

1. How it works (basic idea)
2. One real-world use case
3. Main advantages and limitations

K-Means

How it works: Partitions data into K clusters by assigning points to the nearest centroid and updating centroids iteratively until stable

Use case: Recommendations services, Customer segmentation in marketing

Advantages: Simple, fast, scalable to large datasets

Limitations: Requires number of clusters, sensitive to outliers and initial centroid selection, assumes spherical clusters

Hierarchical Clustering

How it works: Creates a tree of clusters by either merging closest points, clusters agglomerative or splitting one cluster recursively divisive

Use case: Gene expression analysis in biology

Advantages: No need to specify clusters number upfront, visual dendrogram

Limitations: Computationally expensive for large datasets, sensitive to noise and outliers

DBSCAN

How it works: Groups densely regions, mark sparse point as noise

Use case: Anomaly detection in network traffic

Advantages: Finds arbitrarily shaped clusters, robust to outliers, no need to specify clusters number

Limitations: Choosing the right neighborhood radius and minimum points can be challenging, struggles with varying cluster densities

3. Clustering Metrics

Define and explain what each of these metrics measures:

Elbow Method: Measures the percentage of variance explained as a function of the number of clusters

Silhouette Score: Measures how similar an object is to its cluster compared to other clusters cohesion vs separation Ranges from -1 to 1

Davies-Bouldin Index: Measures average similarity between each cluster and its most similar cluster, balancing compactness within-cluster and separation between-clusters Lower values indicate better clustering

METRIC	WHAT IT MEASURES	WHAT A "GOOD" VALUE MEANS	WHEN THE METRIC IS MOST USEFUL
ELBOW METHOD	Percentage of variance explained vs number of clusters	The point where adding more clusters no longer improves variance explained significantly ("elbow" point)	Choosing the optimal number of clusters in methods like K-means
SILHOUETTE SCORE	How similar a point is to its own cluster compared to others	Close to +1: good, 0:overlapping clusters; negative: wrong assignment	Validating cluster quality and selecting cluster count
DAVIES-BOULDIN INDEX	compactness and separation	Lower values indicate better	Evaluating cluster clustering quality without assumptions on cluster shape

4. Challenges in Clustering

Why is clustering considered harder than supervised learning?

It harder than supervised learning because it lacks labeled data to guide the learning process

Choosing the right number of clusters k

Choosing too few clusters can oversimplify the data structure, while too many can overfit noise, is often depends heavily on the data distribution and knowledge

Handling Noise and Outliers

It can be sensitive to noise and outliers, which can degrade cluster quality or lead to misclassification

Dealing with high-dimensional data

high-dimensional data where the notion of distance becomes less meaningful and clusters are harder to distinguish, which also complicates tasks

5. Real-World Case Study

Real-world research project or study that used clustering (e.g., consumer segmentation, image segmentation, genetics, social networks)

Summary:

1. Project Objective
2. Data Used
3. Clustering Method Applied
4. Key Findings or Insights

1. Customer Segment _Retail

1. Objective: To group customers for targeted marketing
2. Data: Purchase history, age, income, shopping behavior
3. Method: K-Means clustering
4. Result: Customers in the identified groups prefer high-spending and low-cost discounting

2. Imaging Segment _Healthcare

1. Objective: To diagnose diseases
2. Data: Scanned images
3. Method: K-Means and hierarchical clustering
4. Result: They helped doctors diagnosis and treatment planning

3. Genetics - DNA

1. Goal: To understand genetic relationships
2. Data: Gene database
3. Method: Hierarchical clustering, DBSCAN
4. Result: The person's ancestry is found