

Lesson 6: Unsupervised Learning – Part 2

◇ Introduction

Welcome back to the second part of our journey into Unsupervised Learning! In Lesson 5, we discovered what unsupervised learning is and explored powerful applications of clustering.

Now, we'll explore the most popular clustering algorithms, learn how they work, and understand when and why to use each one.

◇ Popular Clustering Algorithms

1. K-Means Clustering

What is it?

- A simple and widely-used clustering algorithm.
- Divides data into K distinct, non-overlapping clusters based on distance.

How it works:

1. Choose the number of clusters K.
2. Randomly assign K cluster centers (called centroids).
3. Assign each point to the nearest centroid.
4. Recalculate centroids as the average of assigned points.
5. Repeat steps 3 and 4 until convergence.

When to use it:

- When you have an idea of how many clusters (K) you want.
- Works well with spherical, evenly sized clusters.

Example Use Case:

- Customer segmentation: Grouping customers into categories based on spending behavior.

2. Hierarchical Clustering

What is it?

- A method that builds a tree-like structure (dendrogram) showing how data points are grouped at various levels.

Two types:

- Agglomerative (bottom-up): Each point starts as its own cluster, and clusters merge as similarity increases.
- Divisive (top-down): Starts with one cluster, then splits apart.

When to use it:

- When you want to visualize relationships or don't know how many clusters you need.
- Great for genomics, text analysis, and cases with natural hierarchy.

Example Use Case:

- Document similarity: Grouping articles or research papers by topic hierarchy.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

What is it?

- A powerful clustering method that doesn't require specifying the number of clusters and can identify outliers.

How it works:

- Groups together points that are close in dense regions.
- Points in sparse regions are labeled as noise or outliers.




When to use it:

- When clusters are of arbitrary shape and size.
- When you expect noise in the data (e.g., GPS data or spatial patterns).

Example Use Case:

- Anomaly detection in geospatial data or network traffic.

◇ Clustering Comparison Table

Algorithm	Requires K?	Detects Noise	Shape of Clusters	Suitable For
K-Means	<input checked="" type="checkbox"/> Yes	<input checked="" type="checkbox"/> No	 Spherical	Simple, structured data
Hierarchical	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> No	 Any shape	Visual exploration, hierarchies
DBSCAN	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Yes	 Arbitrary	Noisy, spatial, complex data

◇ How to Evaluate Clustering?

Since clustering is unsupervised, evaluating its performance isn't as straightforward as accuracy in supervised learning. Here are some common techniques:

- Silhouette Score: Measures how close a point is to its cluster vs other clusters (ranges from -1 to 1).
- Elbow Method (for K-Means): Helps determine the optimal value of K.
- Dendrogram (for Hierarchical): Visualizes the best place to "cut" and form clusters.

◇ Outro

In this lesson, you've learned:

- The top three clustering algorithms: K-Means, Hierarchical Clustering, and DBSCAN.
- How they work, where they shine, and how to pick the right one.
- Ways to evaluate the quality of your clustering results.

Clustering is just one part of unsupervised learning. In the next lessons, we'll explore dimensionality reduction techniques like PCA and anomaly detection — two more critical tools in the unsupervised learning toolbox.