Presentation on unsupervised learning: K-means and DBSCAN clustering

AMNIL TECHNOLOGIES

Presented by: Sejal Panta
13th October, 2025

What is Unsupervised Learning?

Unsupervised learning is a type of machine learning where the model learns patterns from unlabeled data. Unlike supervised learning, there is no right answer provided here and the algorithm itself tries to find structure or groupings in the data.

Importance over Supervised Learning:

1. No labeled data required: Useful when labeling is expensive or time-consuming.
2. Discover hidden patterns: Can reveal natural groupings in data that we didn’t know exist.
3. Preprocessing for supervised models: Can reduce dimensions or extract important features.
4. Detect anomalies: Finds unusual data points automatically.

Example:

Unsupervised Learning with Penguins

- Dataset: penguins.csv contains flipper length, culmen length, culmen depth, body mass, etc.
- Goal: Group penguins into clusters.
- Use Case:
  - Identify natural clusters of penguins based on physical features (bill length, flipper length, body mass)
  - Can help explore the dataset before building a supervised model to predict species

K-Means Clustering

- Definition:
  K-Means groups data into K clusters based on similarity. Each cluster has a centroid, and data points are assigned to the nearest centroid.

- How it works?
  1. Pick K random points as centroids
  2. Assign each point to the nearest centroid
  3. Recalculate centroid of each cluster
  4. Repeat until clusters stabilize

Pros:

- Simple, easy to understand
- Works well for well-separated clusters

Cons:

- Must choose K in advance
- Sensitive to outliers
- Assumes clusters are roughly circular

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

- Definition:
  DBSCAN groups dense regions of data and treats low-density points as outliers.

- How it works?
  1. Define eps (distance threshold) and minPts (minimum points for a cluster)
  2. Identify core points with ≥ minPts neighbors within eps
  3. Expand clusters from core points
  4. Points not in any cluster = noise/outliers

Pros:

- Can find clusters of any shape
- Automatically detects outliers
- No need to predefine number of clusters

Cons:

- Sensitive to eps and minPts values
- Struggles if clusters have very different densities

Comparison :

K-Means: Divides data into a fixed number of clusters (k) by finding cluster centers. Works best for roughly circular, equally sized clusters. Need to choose k beforehand, and it’s sensitive to outliers.

DBSCAN: Groups points based on density, automatically finding clusters of any shape and marking sparse points as noise. We don’t need to specify the number of clusters, it handles outliers well, but doesn’t give explicit cluster centers.

Summary: **K-Means** = fixed, center-based clustering
DBSCAN = flexible, density-based clustering.

Interpretation of different silhouette scores:

k(clusters)	Silhouette score	Meaning
2	0.3969	Weak moderate structure
3	0.3960	Similar to k=2, not well separated
4	0.3781	Slightly worse
5	0.5214	Good cluster structure starts forming
6	**0.5460**	Best structure, most distinct
7	0.4745	Declining, too many clusters
8	0.4977	Slight improvement but almost same as k=7

THANK YOU!