DBSCAN & Hierarchical Clustering – Complete Beginner-to-Advanced Revision Guide

---

1. Introduction

What are DBSCAN and Hierarchical Clustering?

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and **Hierarchical Clustering** are **unsupervised clustering algorithms** that do **not require pre-specifying the number of clusters**.

They are especially useful when: - Clusters have arbitrary shapes - Data contains noise and outliers

---

When to Use Them

- Unknown number of clusters
- Non-spherical cluster shapes
- Presence of noise

Real-World Applications

- Geospatial data analysis
- Anomaly detection
- Image segmentation
- Biological data clustering

---

2. DBSCAN (Density-Based Clustering)

2.1 Core Concepts

DBSCAN groups points based on **local density**.

Key Parameters

- ε (eps): neighbourhood radius
- MinPts: minimum points in eps-neighbourhood

2.2 Point Types

- **Core point**: ≥ MinPts in eps-neighbourhood
- **Border point**: < MinPts but reachable from core
- **Noise point**: neither core nor border

2.3 Mathematical Definition

Eps-neighbourhood:

$$ N_\varepsilon(x) = {y \mid d(x,y) \le \varepsilon} $$

Core point condition:

$$ |N_\varepsilon(x)| \ge \text{MinPts} $$

2.4 Density Reachability

- Directly density-reachable
- Density-reachable
- Density-connected

Defines cluster formation

2.5 DBSCAN Algorithm

1. Pick unvisited point
2. Check eps-neighbourhood
3. Expand cluster if core
4. Mark noise otherwise

2.6 Strengths & Limitations

Strengths

- Finds arbitrary-shaped clusters
- Automatically detects noise
- No need for K

Limitations

- Sensitive to eps
- Struggles with varying densities
- Poor in high dimensions

---

3. Hierarchical Clustering

3.1 Core Idea

Builds a **tree of clusters (dendrogram)** showing nested grouping.

Two approaches: - Agglomerative (bottom-up) - Divisive (top-down)

3.2 Agglomerative Clustering

1. Start with each point as its own cluster
2. Merge closest clusters iteratively

3.3 Linkage Criteria

Single Linkage

$$ d(A, B) = \min_{x \in A, y \in B} d(x,y) $$

Complete Linkage

$$ d(A, B) = \max_{x \in A, y \in B} d(x,y) $$

Average Linkage

$$ d(A, B) = \frac{ 1 }{|A||B|} \sum d(x,y) $$

Ward's Method

Minimises variance increase

3.4 Dendrogram Interpretation

- Height = distance
- Cut height defines clusters

3.5 Strengths & Limitations

Strengths

- No need to choose K initially
- Dendrogram gives insight

Limitations

- Computationally expensive
- Sensitive to noise

---

4. Model Evaluation

Internal Metrics

- Silhouette score
- Davies–Bouldin index

External Metrics

- Adjusted Rand Index
- Normalised Mutual Information

---

5. Interpretation

- DBSCAN clusters = dense regions
- Hierarchical clusters = nested structure
- Dendrogram explains hierarchy

---

6. Assumptions

DBSCAN

- Density defines clusters
- Meaningful distance metric

Hierarchical

- Pairwise distances meaningful

---

7. Common Pitfalls & Misconceptions

- ❌ Using DBSCAN without scaling
- ❌ Expecting DBSCAN to work in high dimensions
- ❌ Misreading dendrogram heights
- ❌ Assuming hierarchical clustering is scalable

---

8. Python Implementation

8.1 DBSCAN (scikit-learn)

```
from sklearn.cluster import DBSCAN
from sklearn.preprocessing import StandardScaler

X_scaled = StandardScaler().fit_transform(X)

model = DBSCAN(eps=0.5, min_samples=5)
labels = model.fit_predict(X_scaled)
```

8.2 Hierarchical Clustering

```
from sklearn.cluster import AgglomerativeClustering

model = AgglomerativeClustering(n_clusters=3, linkage='ward')
labels = model.fit_predict(X_scaled)
```

8.3 Dendrogram Visualization

```
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt

dendrogram = sch.dendrogram(sch.linkage(X_scaled, method='ward'))
plt.show()
```

---

9. DBSCAN vs Hierarchical vs K-Means

Feature	DBSCAN	Hierarchical	K-Means
Requires K	No	No	Yes
Noise Handling	Yes	No	No
Shape	Arbitrary	Any	Spherical
Scalability	Medium	Poor	Excellent

---

10. When NOT to Use

- DBSCAN: varying densities
- Hierarchical: large datasets

---

11. Best Practices

- Scale features
- Tune eps using k-distance plot
- Use dendrogram to choose cut

---

12. Summary

- DBSCAN excels at noise and shape
- Hierarchical reveals structure
- Both complement K-Means

End of Revision Guide