# 📘 第七章：非監督式學習（Unsupervised Learning）

## 📖 章節簡介與補充重點

非監督式學習著重於資料中潛在結構的發掘，常見任務包括「分群（Clustering）」與「降維（Dimensionality Reduction）」。本章涵蓋下列關鍵主題：  
  
1. 分群演算法：K-Means、K-Medoids、DBSCAN。  
2. 分群效能評估指標：WCSS、Silhouette Score、Davies–Bouldin Index。  
3. 降維技術：主成分分析（PCA）。  
4. 各方法的數學直覺與應用場景差異。  
  
此外，也強調不同演算法對於離群值（outliers）、群集形狀（spherical vs arbitrary）、預設參數（如 k 值、epsilon）的敏感性差異。

## 🧠 名詞解釋（Term Explanation）

非監督式學習（Unsupervised Learning）  
中文：在無標記資料上學習資料的潛在結構。  
英文：Learning hidden patterns from unlabeled data.

K-Means 分群  
中文：根據距離將資料分為 k 個群集，並反覆更新群中心直到收斂。  
英文：Partitions data into k clusters by minimizing the sum of squared distances to the cluster centers.

K-Medoids 分群  
中文：使用實際資料點作為群中心，對離群值較具魯棒性。  
英文：Uses real data points as centers; more robust to outliers.

DBSCAN  
中文：根據密度進行分群，能找出任意形狀群集與噪聲點。  
英文：Density-based clustering algorithm that can discover clusters of arbitrary shapes and detect noise.

WCSS（總群內平方和）  
中文：衡量群內點與中心的距離平方總和。  
英文：Sum of squared distances between each point and its cluster center.

Silhouette Score  
中文：衡量每個點與同群內點與最近其他群點的差異，介於 [-1, 1]。  
英文：Measures similarity within a cluster and dissimilarity to other clusters.

Davies–Bouldin Index  
中文：每個群的內部距離與其最近群的中心距離的比例，愈小愈好。  
英文：Lower DB index indicates better cluster separation and compactness.

主成分分析（PCA）  
中文：透過正交轉換將資料投影到變異量最大的方向，達到降維目的。  
英文：Projects data onto principal components to reduce dimensionality while preserving variance.

## 📚 名詞比較（Term Comparison）

📌 K-Means vs. K-Medoids  
中文：K-Means 使用「群內平均」作為中心，可能受離群值干擾；K-Medoids 使用「資料點」作為中心，更具穩健性。  
英文：K-Means uses the centroid (mean) as cluster center, sensitive to outliers; K-Medoids uses medoid (data point), more robust.  
  
📌 K-Means vs. DBSCAN  
中文：K-Means 假設群為球狀，需事先給定 k 值；DBSCAN 無需指定群數，可辨識任意形狀與離群點。  
英文：K-Means assumes spherical clusters and requires predefined k; DBSCAN does not need k and detects arbitrary shapes and outliers.  
  
📌 WCSS vs. Silhouette Score vs. DB Index  
中文：三者皆為群集品質評估指標：WCSS 越小越集中，Silhouette 越高越佳，DB Index 越小越理想。  
英文：WCSS (lower = better compactness), Silhouette (higher = better clustering), DB Index (lower = better separation).

## 📝 模擬出題與中英文詳解（包含解釋性問題）

Q1: Q1. 為什麼 DBSCAN 不需要事先指定群數 k？這樣有什麼優勢與挑戰？  
答：DBSCAN 根據密度自動決定群數，適合任意形狀資料，但需設定適當的 epsilon 與 minPts，否則可能過分分群或無法分群。  
DBSCAN determines the number of clusters based on density, suitable for arbitrary-shaped data. However, choosing appropriate epsilon and minPts is crucial.

Q2: Q2. 為何 PCA 可以當作降維方法？其在非監督學習中的應用是什麼？  
答：PCA 找出資料變異最大方向進行轉換，保留最多資訊，並可視覺化高維資料結構或用於後續的聚類。  
PCA finds the directions of maximum variance and projects data accordingly, useful for visualization and preprocessing before clustering.

Q3: Q3. 為什麼使用 Silhouette Score 評估 K 值比單用 WCSS 更具判斷力？  
答：因為 Silhouette 同時考量群內一致性與群間分離性，而 WCSS 只看群內誤差，因此前者能更全面反映分群品質。  
Silhouette Score considers both intra-cluster cohesion and inter-cluster separation, offering a more balanced assessment than WCSS.

Q4: Q4. 若資料存在非球狀分佈與離群點，為何 K-Means 效果可能不佳？應選擇何種替代方法？  
答：K-Means 假設球狀群且對離群值敏感，這會導致錯誤分群；可考慮使用 DBSCAN 或 K-Medoids。  
K-Means assumes spherical clusters and is sensitive to outliers, possibly leading to poor clustering. DBSCAN or K-Medoids may be better alternatives.