# Notes: Hierarchical Clustering

## 1. Tổng quan

Hierarchical Clustering là thuật toán phân cụm hoạt động dựa trên ý tưởng chia nhỏ hoặc gộp dần dữ liệu cho đến khi đạt trạng thái hội tụ. Không giống như K-means, thuật toán này không yêu cầu xác định trước số cụm (k).

Có hai hướng tiếp cận chính:  
- Agglomerative (Bottom-up): bắt đầu từ từng điểm riêng lẻ, sau đó gộp dần thành cụm lớn hơn.  
- Divisive (Top-down): bắt đầu từ một cụm tổng thể, sau đó chia nhỏ dần thành các cụm con.

## 2. Các phương pháp Linkage (Liên kết cụm)

Các phương pháp Linkage được sử dụng để xác định khoảng cách giữa các cụm trong quá trình hợp nhất.

### a. Single Linkage (Minimum Pairwise Distance)

Đo khoảng cách nhỏ nhất giữa hai điểm thuộc hai cụm khác nhau.  
• Ưu điểm: giúp duy trì khoảng cách nhỏ giữa các cụm.  
• Nhược điểm: nhạy cảm với nhiễu (noise) và outlier.

### b. Complete Linkage (Maximum Pairwise Distance)

Đo khoảng cách lớn nhất giữa hai điểm thuộc hai cụm khác nhau.  
• Ưu điểm: giúp tách biệt cụm tốt hơn khi dữ liệu có nhiễu.  
• Nhược điểm: dễ chia nhỏ cụm lớn, dẫn đến nhiều cụm nhỏ hơn cần thiết.

### c. Average Linkage (Average Pairwise Distance)

Tính trung bình khoảng cách giữa tất cả các điểm thuộc hai cụm khác nhau.  
• Ưu điểm: cân bằng giữa Single và Complete Linkage.  
• Nhược điểm: có thể phá vỡ các cụm lớn hiện có nếu dữ liệu phân bố không đều.

### d. Ward Linkage (Inertia-Based Merge)

Ward Linkage hợp nhất hai cụm sao cho tổng bình phương khoảng cách (inertia) giữa các điểm và tâm cụm tăng ít nhất.  
• Ưu điểm: tạo ra các cụm có kích thước đồng đều và giảm tổng sai số nội cụm.  
• Nhược điểm: chi phí tính toán cao, đặc biệt khi dữ liệu lớn.

## 3. Ưu và Nhược điểm của Hierarchical Clustering

Ưu điểm:  
- Không cần xác định trước số cụm (k).  
- Dễ trực quan hóa qua biểu đồ dendrogram.  
- Hiểu rõ cấu trúc phân cấp tự nhiên của dữ liệu.

Nhược điểm:  
- Chi phí tính toán cao (O(n²) hoặc hơn).  
- Không thích hợp với dữ liệu lớn.  
- Nhạy cảm với lựa chọn metric và linkage method.

## 4. Cách triển khai trong Python

Triển khai bằng thư viện scikit-learn:  
  
from sklearn.cluster import AgglomerativeClustering  
  
agg = AgglomerativeClustering(n\_clusters=3, affinity='euclidean', linkage='ward')  
agg.fit(X)  
y\_pred = agg.labels\_

Trong đó:  
- n\_clusters: số cụm cần chia.  
- affinity: phương pháp đo khoảng cách (euclidean, manhattan,...).  
- linkage: phương pháp liên kết (‘ward’, ‘complete’, ‘average’, ‘single’).

## 5. Tổng kết

Hierarchical Clustering là công cụ mạnh để khám phá cấu trúc dữ liệu và trực quan hóa các mối quan hệ giữa các điểm dữ liệu. Tuy nhiên, cần cân nhắc khi áp dụng cho dữ liệu lớn vì độ phức tạp tính toán cao.