Notes: Unsupervised Learning Algorithms

## 1. Tổng quan

Unsupervised Learning Algorithms được sử dụng khi dữ liệu không có nhãn (unlabeled data) và ta không có biến mục tiêu (outcome variable) để dự đoán. Các thuật toán này giúp phát hiện cấu trúc ẩn trong dữ liệu hoặc chia nhỏ tập dữ liệu thành các nhóm (clusters) nhỏ hơn.

## 2. Các loại Unsupervised Learning

Hai nhóm chính bao gồm:

- Clustering: dùng dữ liệu chưa gán nhãn để tìm các nhóm có đặc điểm tương đồng. Ví dụ: phân nhóm khách hàng (customer segmentation). Các thuật toán phổ biến gồm K-means, Hierarchical Clustering, DBSCAN.

- Dimensionality Reduction: giảm số chiều (features) của dữ liệu mà vẫn giữ lại thông tin quan trọng. Ví dụ: PCA (Principal Component Analysis), Non-negative Matrix Factorization (NMF).

## 3. Dimensionality Reduction và Curse of Dimensionality

Dimensionality reduction đặc biệt quan trọng khi xử lý dữ liệu có số lượng đặc trưng lớn. Curse of Dimensionality xảy ra khi số lượng features quá nhiều khiến mô hình khó học được mối quan hệ thực sự.

Các vấn đề thường gặp:

- Quá nhiều features dẫn đến overfitting.  
- Tăng yêu cầu tính toán (computational cost).  
- Giảm hiệu suất mô hình (poorer performance).  
- Tăng khả năng xuất hiện outlier.

Giải pháp: giảm số chiều dữ liệu bằng PCA hoặc chọn lọc feature để giảm nhiễu và cải thiện khả năng tổng quát hoá.

## 4. Ứng dụng trong thực tế

Các ứng dụng của Clustering bao gồm:

1. Anomaly Detection: phát hiện giao dịch gian lận hoặc hành vi bất thường trong dữ liệu.  
2. Customer Segmentation: chia khách hàng theo hành vi, nhân khẩu học hoặc tần suất tương tác.  
3. Improve Supervised Learning: sử dụng kết quả clustering làm input cho mô hình phân loại (classification).

Các ứng dụng của Dimensionality Reduction:

1. Image Compression: nén hình ảnh có độ phân giải cao thành dạng dữ liệu nhỏ gọn hơn.  
2. Image Tracking: giảm số lượng đặc trưng chính, giúp tăng tốc độ và giảm chi phí tính toán.

## 5. K-means Clustering

K-means là một thuật toán phân cụm lặp (iterative clustering algorithm) chia dữ liệu thành K nhóm. Thuật toán khởi tạo ngẫu nhiên K điểm làm tâm cụm (centroids), sau đó gán từng điểm dữ liệu vào cụm có tâm gần nhất. Sau mỗi vòng lặp, tâm cụm được cập nhật dựa trên trung bình các điểm trong cụm, cho đến khi không thay đổi nữa.

Ưu điểm: tính toán đơn giản, hiệu quả với dữ liệu lớn.

Nhược điểm: nhạy cảm với điểm khởi tạo (initialization) và không xử lý tốt dữ liệu không cầu phương (non-spherical).

## 6. Chọn số lượng cụm (Model Selection)

Để chọn số cụm K phù hợp, ta có thể dùng các tiêu chí như Inertia hoặc Distortion.  
Inertia đo tổng khoảng cách bình phương từ mỗi điểm đến tâm cụm, giá trị nhỏ hơn nghĩa là cụm chặt hơn.  
Distortion đo trung bình khoảng cách giữa các điểm trong cùng cụm.

Khi số cụm tăng, inertia giảm, nhưng nếu giảm quá ít khi tăng thêm cụm, ta đạt điểm 'elbow' – điểm dừng tối ưu.

## 7. Tổng kết

Unsupervised Learning cho phép khám phá cấu trúc ẩn trong dữ liệu mà không cần nhãn. K-means và PCA là hai kỹ thuật nền tảng được sử dụng rộng rãi trong phân tích dữ liệu, giúp giảm nhiễu, rút trích đặc trưng và hỗ trợ mô hình học có giám sát (Supervised Learning) hiệu quả hơn.