# Notes: Distance Metrics in Unsupervised Learning

## 1. Tổng quan

Trong các thuật toán phân cụm (Clustering Methods), việc xác định 'khoảng cách' (Distance Metric) giữa các điểm dữ liệu là yếu tố cốt lõi. Lựa chọn Distance Metric phù hợp sẽ ảnh hưởng trực tiếp đến hiệu quả và độ chính xác của các thuật toán như K-means, Hierarchical Clustering.

## 2. Các loại Distance Metrics phổ biến

### a. Euclidean Distance (L2 Distance)

Đây là loại khoảng cách quen thuộc nhất, tính bằng căn bậc hai của tổng bình phương hiệu giữa các tọa độ:  
d(p, q) = √Σ (pi - qi)²  
Metric này đo khoảng cách thực tế giữa hai điểm trong không gian và được dùng phổ biến trong K-means.

• Ưu điểm: Trực quan, dễ hiểu, hiệu quả với dữ liệu có ít chiều.  
• Nhược điểm: Dễ bị ảnh hưởng bởi Curse of Dimensionality khi số chiều tăng cao.

### b. Manhattan Distance (L1 or City Block Distance)

Khoảng cách Manhattan đo bằng tổng giá trị tuyệt đối của chênh lệch giữa các tọa độ:  
d(p, q) = Σ |pi - qi|  
Giống như việc di chuyển theo đường phố trong thành phố hình lưới (City Block).

• Ưu điểm: Ổn định hơn khi có outlier.  
• Nhược điểm: Có thể kém chính xác trong không gian nhiều chiều.

### c. Cosine Distance

Cosine Distance dựa trên góc giữa hai vector, thay vì độ lớn của chúng:  
cos(θ) = (A · B) / (||A|| ||B||)  
Nó đo mức độ tương đồng hướng giữa hai vector, tức là chúng có cùng hướng đến mức nào.

• Ưu điểm: Phù hợp với dữ liệu văn bản (text data) hoặc vector TF-IDF, nơi độ lớn không quan trọng.  
• Nhược điểm: Không phản ánh tốt khoảng cách vật lý giữa các điểm.

### d. Jaccard Distance

Jaccard Distance được dùng cho dữ liệu dạng tập hợp (sets), đặc biệt hữu ích trong phân tích văn bản. Công thức:  
J(A, B) = 1 - (|A ∩ B| / |A ∪ B|)  
Giá trị càng nhỏ thì hai tập càng giống nhau.

Ví dụ:  
A = {I, like, chocolate, ice, cream}  
B = {Do, I, want, chocolate, cream, or, vanilla}  
→ |A ∩ B| = 3, |A ∪ B| = 9 ⇒ J(A, B) = 1 - 3/9 = 0.67

## 3. So sánh và Ứng dụng

• Euclidean Distance: dùng cho dữ liệu tọa độ liên tục, nhưng nhạy cảm với curse of dimensionality.  
• Manhattan Distance: phù hợp khi dữ liệu có nhiều chiều và tồn tại outlier.  
• Cosine Distance: dùng cho vector hướng, đặc biệt trong xử lý văn bản (text mining).  
• Jaccard Distance: dùng cho dữ liệu rời rạc hoặc phân tích văn bản theo từ (word occurrence).

## 4. Tổng kết

Việc lựa chọn Distance Metric phụ thuộc vào loại dữ liệu và mục tiêu bài toán. Không có metric nào là tối ưu cho mọi trường hợp. Trong thực tế, Euclidean và Cosine Distance là hai loại phổ biến nhất trong các bài toán Unsupervised Learning.