# KHAI THÁC DỮ LIỆU

Nguyễn Xuân Việt Đức - 22280012

Bài tập lý thuyết - Lần 2

## 1 Kỹ thuật ISOMAP (Isometric Mapping)

ISOMAP là một kỹ thuật giảm chiều dữ liệu phi tuyến tính, được phát triển như một phương pháp cải tiến so với các phương pháp tuyến tính truyền thống như PCA. ISOMAP cố gắng bảo toàn khoảng cách trắc địa (geodesic distance) giữa các điểm dữ liệu, thay vì khoảng cách Euclidean thông thường.

### 1.1 Nguyên lý hoạt động

ISOMAP hoạt động qua 3 bước chính:

1. Xây dựng đồ thị láng giềng (neighborhood graph) từ dữ liệu đầu vào
2. Tính toán khoảng cách trắc địa giữa tất cả các cặp điểm
3. Áp dụng MDS (Multidimensional Scaling) để giảm chiều dữ liệu

Thuật toán ISOMAP có thể mô tả chi tiết như sau:

1. **Xây dựng đồ thị láng giềng:** Xác định $k$ láng giềng gần nhất của mỗi điểm dựa trên khoảng cách Euclidean, hoặc kết nối các điểm có khoảng cách nhỏ hơn ngưỡng $\epsilon$.
2. **Tính toán khoảng cách trắc địa:** Sử dụng thuật toán đường đi ngắn nhất (như Floyd-Warshall hoặc Dijkstra) để ước tính khoảng cách trắc địa giữa mọi cặp điểm trên đa tạp.
3. **Thực hiện embedding:** Áp dụng MDS cổ điển lên ma trận khoảng cách trắc địa để tìm tọa độ trong không gian thấp chiều sao cho bảo toàn các khoảng cách.

### 1.2 Ưu điểm của ISOMAP

- **Khả năng xử lý dữ liệu phi tuyến tính:** ISOMAP có thể phát hiện và bảo toàn cấu trúc phi tuyến tính trong dữ liệu nhiều chiều, khắc phục hạn chế của các phương pháp tuyến tính như PCA.

- **Bảo toàn tốt hơn khoảng cách thực sự:** Bằng cách sử dụng khoảng cách trắc địa, ISOMAP nắm bắt được khoảng cách thực sự giữa các điểm dữ liệu dọc theo đa tạp (manifold).
- **Hiệu quả với dữ liệu nằm trên đa tạp:** Đặc biệt hiệu quả khi dữ liệu nằm trên một đa tạp có tính chất "gần như đẳng cự" với không gian Euclidean thấp chiều.
- **Cung cấp không gian nhúng có ý nghĩa:** Không gian nhúng được tạo ra thường có thể biểu diễn được các đặc trưng có ý nghĩa của dữ liệu.
- **Khả năng trực quan hóa dữ liệu:** Rất hữu ích cho việc trực quan hóa dữ liệu nhiều chiều bằng cách giảm xuống 2 hoặc 3 chiều.

## 1.3 Nhược điểm của ISOMAP

- **Độ phức tạp tính toán cao:** Yêu cầu tính toán khoảng cách giữa mọi cặp điểm, độ phức tạp là $O(n^3)$ với $n$ là số lượng điểm dữ liệu, khiến nó không hiệu quả với tập dữ liệu lớn.
- **Nhạy cảm với nhiễu:** Kết quả của ISOMAP có thể bị ảnh hưởng đáng kể bởi nhiễu trong dữ liệu và lựa chọn tham số láng giềng $k$.
- **Vấn đề với "lỗ thủng" trong dữ liệu:** ISOMAP gặp khó khăn khi đa tạp có các "lỗ thủng" (holes) hoặc không liên thông, có thể dẫn đến kết quả không chính xác.
- **Không có ánh xạ ngược:** ISOMAP không cung cấp ánh xạ ngược từ không gian thấp chiều về không gian ban đầu.
- **Vấn đề với việc chọn tham số $k$:** Hiệu suất phụ thuộc nhiều vào việc lựa chọn số lượng láng giềng $k$ phù hợp, nhưng không có phương pháp tối ưu để chọn giá trị này.
- **Tính toàn vẹn của đồ thị:** Dòi hỏi đồ thị láng giềng phải liên thông, nếu không sẽ gặp vấn đề với việc tính khoảng cách trắc địa.

## 1.4 So sánh với các phương pháp khác

ISOMAP thường được so sánh với các kỹ thuật giảm chiều khác như PCA, LLE (Locally Linear Embedding), và t-SNE:

- **So với PCA:** ISOMAP hiệu quả hơn với dữ liệu phi tuyến tính, nhưng tốn kém tính toán hơn. PCA tìm kiếm các hướng có phương sai lớn nhất, trong khi ISOMAP tìm kiếm cấu trúc đa tạp cơ bản.
- **So với LLE:** Cả hai đều xử lý dữ liệu phi tuyến tốt, nhưng ISOMAP có xu hướng bảo toàn cấu trúc toàn cục tốt hơn. LLE tập trung vào việc bảo toàn cấu trúc cục bộ.

- **So với t-SNE:** ISOMAP tập trung vào cấu trúc toàn cục, trong khi t-SNE chú trọng hơn đến cấu trúc cục bộ và thường tốt hơn cho việc trực quan hóa các cụm dữ liệu.

## 1.5 Kết luận

ISOMAP là một lựa chọn tốt khi cần giảm chiều dữ liệu phi tuyến có cấu trúc đa tạp rõ ràng, đặc biệt khi cần bảo toàn khoảng cách trên toàn bộ đa tạp. Tuy nhiên, phương pháp này không phù hợp với tập dữ liệu rất lớn do chi phí tính toán cao và có những hạn chế nhất định khi đa tạp có cấu trúc phức tạp hoặc dữ liệu có nhiều nhiễu.

Định nghĩa toán học của ISOMAP có thể được tóm tắt như sau:

Với một tập dữ liệu $X = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^D$, ISOMAP tìm một ánh xạ $f : X \to \mathbb{R}^d$ (với $d < D$) sao cho:

$$||\mathbf{f}(x_i) - \mathbf{f}(x_j)|| \approx d_G(x_i, x_j) \tag{1}$$

trong đó $d_G(x_i, x_j)$ là khoảng cách trắc địa giữa $x_i$ và $x_j$ trên đa tạp dữ liệu.