Isomap

*E-mail: pima.vn@gmail.com

Mô tả dự án

Manifold Learning ("Học đa tạp") là tên một nhóm các thuật toán giảm chiều dữ liệu không giám sát với điểm chung là bắt đầu với giả thiết các điểm dữ liệu trong không gian D chiều đều nằm trên một manifold (đa tạp) d chiều với $d < D$, và sử dụng các tính chất cục bộ (local properties) của manifold để tìm ra biểu diễn d chiều của các điểm dữ liệu ban đầu.

Thuật toán **Isomap** (Isometric mapping - "Ánh xạ đẳng cự") là một thuật toán Manifold Learning với mục tiêu bảo toàn khoảng cách địa lý (geodesic distance) giữa mọi cặp điểm trong input data, tận dụng hình dạng hình học của manifold.

Yêu cầu cơ bản

Lý thuyết:

- Phát biểu ngắn gọn bài toán mà thuật toán Isomap muốn giải và ý tưởng đằng sau nó. Gợi ý: khoảng cách trên manifold khác gì khoảng cách euclid? Việc xây dựng neighbor graph giúp ích gì cho việc phát hiện cấu trúc manifold?
- Mô tả một số thuật toán giúp xây dựng đồ thị dựa trên khoảng cách (neighbor graph) từ input data.
- Mô tả từng bước cụ thể của Isomap và các bài toán tối ưu mà thuật toán này giải.
- Chứng minh công thức nghiệm hoặc giải thích phương pháp tìm nghiệm cho từng bài toán tối ưu trên. Gợi ý: có 1 bài toán tối ưu trên graph và một bài toán tối ưu ĐSTT.
- Nêu ra các trường hợp các bài toán tối ưu trong thuật toán không giải được hoặc không có nghiệm duy nhất và nêu một số cách khắc phục.

Thực hành:

- Áp dụng Isomap vào dataset Swiss Roll của sklearn để giảm dữ liệu xuống 2 chiều, thể hiện các điểm output data lên một scatter plot. Thử lại nhiều lần với các số nearest neighbors khác nhau và nhận xét về ảnh hưởng số nearest neighbors lên kết quả.
- Áp dụng Isomap lên một dataset tự chọn, theo phương pháp tự chọn và nhận xét, giải thích kết quả.

Yêu cầu nâng cao

Lý thuyết:

- Hãy nêu một hoặc một số phương pháp giúp chọn số nearest neighbors phù hợp cho Isomap trên từng dataset.
- Mô tả một số cải biến giúp cải thiện thời gian chạy của Isomap.

Thực hành:

- Hãy thử cải thiện phương pháp xây dựng neighbor graphs so với phiên bản Isomap của scikit-learn. So sánh kết quả của phiên bản cải biến Isomap của các bạn và phiên bản scikit-learn trên cùng một dataset. Lưu ý: các bạn có thể sẽ phải viết lại thuật toán mà không dùng thư viện.
- Sử dụng Isomap để nén một ảnh tự chọn, thể hiện các chiều mới dưới dạng heatmap và nhận xét. Tái tạo lại ảnh bằng phương pháp Leave-one-out prediction với kernel tự chọn và nhận xét kết quả.

Kiến thức

Kiến thức toán: Đại số tuyến tính, Giải tích nhiều biến, Lý thuyết đồ thị.

Một số từ khóa giúp các bạn tìm kiếm thông tin hiệu quả hơn:

- Thông tin chung: Isomap, Manifold Learning, k-nearest neighbors, Shortest path on graph.
- Tối ưu hóa: Multi-dimensional scaling, Min-max theorem, Eigendecomposition.

Tham Khảo

[1] Các bài giảng PiMA 2021.
[2] https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction
[3] https://scikit-learn.org/stable/modules/manifold.html