Principal Component Analysis (PCA)

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

Mô tả dự án

Thuật toán Principal Component Analysis ("Phân tích chiều chính") là một trong những thuật toán giảm chiều dữ liệu không giám sát được phát minh sớm nhất mà vẫn được dùng rộng rãi tới ngày nay. Mục tiêu của PCA là tìm một phép chiếu trên không gian ít chiều sao cho hình chiếu của dữ liệu mới qua phép chiếu này có các tính chất tối ưu nhất định.

Yêu cầu cơ bản

Lý thuyết:

- Phát biểu ngắn gọn bài toán mà thuật toán PCA muốn giải và ý tưởng đằng sau nó.
- Mô tả từng bước cụ thể của PCA 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 ĐSTT.
- Tìm hàm nhúng (embedding) và hàm xấp xỉ ngược của PCA.
- Mô tả thuật toán Kernel PCA. Giải thích sự giống và khác nhau giữa Kernel PCA và PCA thông thường, và ý nghĩa của Kernel trick.

Thực hành:

- Áp dụng PCA 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 nêu nhận xét, giải thích kết quả.
- Áp dụng PCA lên một dataset tự chọn và nhận xét, giải thích kết quả.

Yêu cầu nâng cao

Lý thuyết:

- Có một cách định nghĩa tương đương của PCA bằng bài toán xấp xỉ ngược tuyến tính từ không gian ít chiều. Hãy phát biểu bài toán này và chứng minh sự tương đương giữa nghiệm 2 bài toán.
- Nêu ra các trường hợp các bài toán tối ưu trong thuật toán không giải được hoặc không có nghiệm duy nhất (nếu có) và nêu một số cách khắc phục.

- Nêu một số hàm kernel thông dụng và ý nghĩa của chúng.
- So sánh PCA với thuật toán TruncatedSVD và rút ra điểm mạnh/yếu tương quan giữa 2 thuật toán.

Thực hành:

- Sử dụng PCA để nén một ảnh tự chọn, thể hiện các chiều chính dưới dạng heatmap và nhận xét. Tái tạo lại ảnh bằng hàm xấp xỉ ngược và nhận xét kết quả.

Kiến thức

Kiến thức toán: Đại số tuyến tính, Giải tích nhiều biến.

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

- Thông tin chung: Orthogonal Projection, Covariance Matrix, Linear Approximation.
- Tối ưu hóa: Min-max theorem, Eigendecomposition.

Tham Khảo

[1] Các bài giảng PiMA 2021.
[2] https://en.wikipedia.org/wiki/Principal_component_analysis
[3] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html