Principal Component Analysis

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

Mô Tả Dự Án

Principal Component Analysis (PCA) là một trong những kỹ thuật làm giảm số chiều (dimension reduction) của không gian dữ liệu đang xét, rất thông dụng trong thống kê (statistics), học máy (machine learning), hay lý thuyết thông tin (information theory).

Công cụ chính để xây dựng phương pháp này là ma trận và cơ sở tương ứng với vector riêng trong Đại số Tuyến tính. Ý tưởng chính là biểu diễn lại các vector dữ liệu trong một cơ sở con thích hợp (cơ sở ứng với các giá trị riêng lớn nhất) để chỉ cần lưu lại những thông tin trọng yếu của toàn bộ dữ liệu.

Trong dự án này, các bạn sẽ tìm hiểu nền tảng lý thuyết của PCA và áp dụng công cụ này vào một dữ liệu cụ thể.

Yêu cầu

Hãy tìm hiểu và trình bày mô hình PCA theo các yêu cầu sau.

(1) Mô tả cụ thể cách xây dựng một ma trận dữ liệu từ một tập dữ liệu cho trước. Tại sao lại phải chuẩn hóa dữ liệu về quanh gốc tọa độ?
(2) Mô tả ma trận hiệp phương sai (covariance matrix) và ý nghĩa của nó.
(3) Trình bày và chứng minh những kết quả chính của lý thuyết ma trận đối xứng và trị riêng dùng cho PCA.
(4) Nêu ý nghĩa của các kết quả trong việc biểu diễn lại vector thông tin cũng như tính toán.
(5) Ứng dụng mô hình PCA vào một dữ liệu cụ thể. Thay đổi số lượng trị riêng được dùng và đưa ra nhận xét về dữ liệu tương ứng.
(6) Trong thực tế, việc giải đa thức đặc trưng để tìm các trị riêng không phải là một việc dễ (đặc biệt khi bậc của đa thức ≥ 5). Hãy tìm và mô tả một phương pháp thay thế khác để thực hiện bước này (có thể là xấp xỉ các giá trị riêng lớn nhất).

Một số từ khóa: Subspace, Basis, Projection, Dimension Reduction, Covariance Matrix, Symmetric Matrix, Eigenbasis, Eigenvalue Algorithm.

Tham Khảo

[1] Các bài giảng PiMA 2018.
[2] `https://en.wikipedia.org/wiki/Principal_component_analysis`
[3] Chen Yu. *Linear Algebra and Face Recognition*, Lecture Note, Indiana University.
Link: `http://www.indiana.edu/~dll/B657/B657_lec_pca.pdf`