

# Spectral Clustering

\*E-mail: [pima.vn@gmail.com](mailto:pima.vn@gmail.com)

## Mô tả Dự án

Mô hình **Spectral Clustering** trong Machine Learning được phổ biến bởi **Shi, Malik** (2000) và **Ng, Jordan, Weiss** (2002). Mô hình này lấy ý tưởng từ việc biểu diễn dữ liệu dưới dạng đồ thị và sử dụng thông tin về dữ liệu thông qua các ma trận đặc biệt từ đồ thị đó. Trong dự án này, chúng ta sẽ tìm hiểu về thuật toán Spectral Clustering.

## Yêu cầu chung

Lý thuyết: Trình bày mô hình Spectral Clustering thỏa mãn các yêu cầu sau.

- Tìm hiểu các cách chuyển dữ liệu về đồ thị và phân tích ưu, nhược điểm của các cách.
- Phát biểu bài toán Clustering dưới dạng đồ thị. Trình bày và chứng minh các bước chuyển bài toán đồ thị đó về bài toán tối ưu Trace minimization problem.
- Trình bày thuật toán Spectral Clustering và giải thích ý nghĩa toán học của thuật toán.
- Tìm hiểu và nêu ra các đặc điểm và ý nghĩa của các ma trận đặc biệt trong thuật toán.
- Tìm hiểu cách chọn tham số  $k$  (số cụm) bằng cách eigengap heuristic.

Thực hành:

- Tìm hiểu cách sử dụng thư viện trên scikit-learn
- Áp dụng mô hình Spectral Clustering vào một hoặc nhiều dữ liệu cụ thể của thư viện scikit-learn
- Plot kết quả ở các bước và nhận xét về kết quả của thuật toán

## Câu hỏi nâng cao

Lý thuyết: Tìm hiểu và phân tích về ưu, nhược điểm của việc lựa chọn giữa ba ma trận Laplacian matrix, Symmetric normalized Laplacian và Random walk normalized Laplacian (Đây sẽ dẫn đến 3 phiên bản khác nhau của thuật toán)

Thực hành: Lập trình thuật toán Spectral Clustering mà không sử dụng thư viện.

## Kiến thức

- Kiến thức: Đại số tuyến tính, Đồ thị, Tối ưu hóa
- Một số từ khóa: Spectral clustering, Adjacency matrix, Laplacian matrix, Normalized Laplacian matrices, Mincut, Trace minimization problem.

## Tham Khảo

[1] Các bài giảng PiMA 2021.

[2] [https://en.wikipedia.org/wiki/Spectral\\_clustering](https://en.wikipedia.org/wiki/Spectral_clustering)