Gaussian Mixture Models

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

Mô tả Dự án

Gaussian mixture model (**GMM**) là mô hình xác suất miêu tả các quần thể con của một tập dữ liệu không gắn nhãn. **GMM** được phát biểu lần đầu bởi **Karl Pearsons** vào năm **1894**. Về sau, với sự phát triển và phổ biến của maximum likelihood estimation, mô hình này được nghiên cứu và ứng dụng trong các bài toán như phân cụm dữ liệu (clustering), nhận dạng chữ viết tay (handwriting recognition), phân đoạn hình ảnh (image segmentation). Trong dự án này, chúng ta sẽ tìm hiểu về ứng dụng của GMM trong clustering và cách cập nhật các tham số của phân phối Gaussian và các cụm dữ liệu với thuật toán expectation maximization (**EM** algorithm).

Yêu cầu cơ bản

Lý thuyết:

- Phát biểu và trình bày ý tưởng của mô hình GMM trong bài toán phân cụm.
- Trình bày cụ thể mô hình GMM theo ngôn ngữ toán học.
- Xây dựng hàm mục tiêu cần tối ưu và điều kiện ràng buộc.
- Trong phần lớn các trường hợp, người ta sử dụng thuật toán EM để tìm các tham số của những phân phối Gaussian thay vì trực tiếp giải bài toán tối ưu hàm mục tiêu. Hãy trình bày thuật toán EM và giải thích, chứng minh ý nghĩa toán học của các công thức cập nhật.
- Trong quá trình trình bày, chỉ rõ ra được ý nghĩa hình học của các tham số, hàm số và nêu một số ví dụ cụ thể trong bài làm để giải thích các bước thực hiện.

Thực hành:

- Tìm hiểu cách sử dụng GMM và thuật toán EM với thư viện scikit-learn.
- Áp dụng mô hình GMM và thuật toán EM vào một hoặc nhiều dữ liệu cụ thể.
- Nhận xét về kết quả và đánh giá mô hình. Nếu có thể, hãy trình bày dữ liệu một cách trực quan.

Câu hỏi nâng cao

- Lập trình thuật toán EM mà không sử dụng thư viện sklearn.
- Tìm hiểu những phương pháp tìm tham số của những phân phối Gaussian bên cạnh thuật toán EM (V.d: moment matching, spectral method) và nhận xét ưu,nhược điểm của những phương pháp đó.

Kiến thức

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

- **Kiến thức Toán:** Xác suất và thống kê, các phép toán cơ bản trên ma trận, tối ưu hóa
- **Một số từ khóa:** Gaussian mixture models, EM algorithm, unsupervised learning, MLE, clustering, Gaussian distribution.

Tham Khảo

[1] Các bài giảng PiMA 2021.
[2] `https://en.wikipedia.org/wiki/Mixture$_$model`