**Báo Cáo**

Thuật toán tự viết:

* Hàm khởi tạo các tham số trong class Mahalanobis gồm số cụm, các tâm cụm và nhãn:

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, số

Nội dung do AI tạo ra có thể không chính xác.

* Hàm chọn ngẫu nhiên các điểm dữ liệu trong bộ dữ liệu làm các tâm cụm đầu tiên:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

* Hàm tính khoảng cách mahalanobis

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, Phông chữ

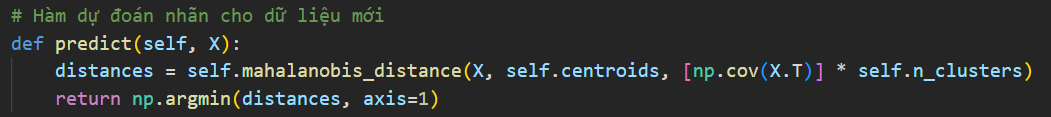
Nội dung do AI tạo ra có thể không chính xác.

* Hàm fit mô hình:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

* Hàm dự đoán nhãn cho dữ liệu:



* Triển khai mô hình và vẽ biểu đồ scatter cho dữ liệu:

Ảnh có chứa văn bản, ảnh chụp màn hình

Nội dung do AI tạo ra có thể không chính xác.

Ảnh có chứa biểu đồ, văn bản, ảnh chụp màn hình, hàng

Nội dung do AI tạo ra có thể không chính xác.

* Kiểm tra mô hình và vẽ dữ liệu với nhiều tâm cụm:

Ảnh có chứa văn bản, biểu đồ, mẫu

Nội dung do AI tạo ra có thể không chính xác.

Nhận xét 3 thuật toán GMM, K-means và Mahalanobis k-means:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **GMM (Gaussian Mixture Model)** | **K-means** | **Mahalanobis k-means** |
| **Ý tưởng** | Phân cụm dựa trên xác suất: mỗi cụm là một **phân phối Gaussian** | Phân cụm dựa trên trung tâm (centroid) và khoảng cách **Euclidean** | Giống K-means nhưng dùng **khoảng cách Mahalanobis** để đo đạc |
| **Hình dạng** | Elip | Cầu | Elip |
| **Tốc độ** | Chậm (do thuật toán EM) | Nhanh nhất | Chậm hơn K-means và nhanh hơn GMM |
| **Chi phí tính toán** | Cao | Thấp | Cao hơn K-means và thấp hơn GMM |
| **Dữ liệu đầu vào** | Do sử dụng mô hình xác suất nên có thể khắc phục tương đối dữ liệu nhiễu và outliers. Hoạt động tốt với dữ liệu nhiều chiều, nên chuẩn hóa dữ liệu trước để tăng độ ổn định. | Hoạt động tốt nhất khi chiều dữ liệu <= 10, và nên chuẩn hóa dữ liệu trước. Nhạy cảm với nhiễu và outliers | Ít nhạy cảm với dữ liệu nhiễu và outliers hơn K-means. Cần chuẩn hóa dữ liệu để ma trận hiệp phương sai chính xác. |
| **Ưu điểm** | Mô hình hóa linh hoạt, xác suất, ước lượng mật độ | Đơn giản, nhanh, hiệu quả | Xử lý cụm phi cầu, ít nhạy cảm với tỷ lệ chiều |
| **Nhược điểm** | Phức tạp, tính toán nặng | Nhạy cảm khởi tạo, K cố định, chỉ cụm cầu | Phức tạp hơn K-Means, nhạy cảm cụm nhỏ với nhiễu và outliers, K cố định |
| **Ứng dụng** | Bài toán nhận dạng giọng nói, nhận dạng khuôn mặt, phân tích ảnh y tế. | Bài toán phân khúc tài liệu, phân khúc khách hàng, nén ảnh, nhận dạng ảnh, phân tích dữ liệu địa lý. | Bài toán phân tích dữ liệu cảm biến, phân cụm tế bào trong sinh học. |