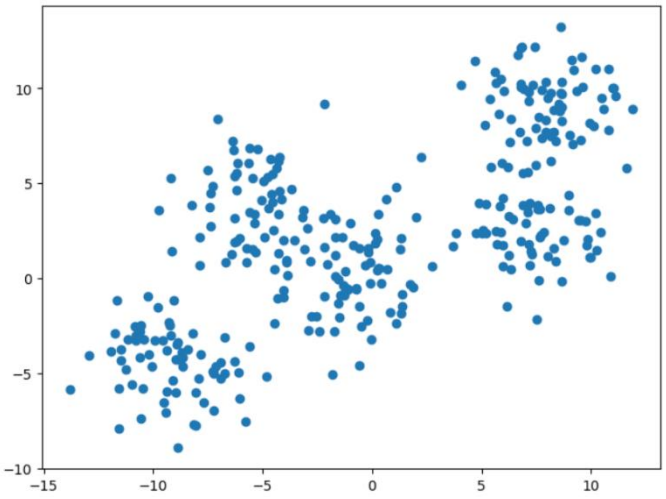
BÁO CÁO THỰC HÀNH KHAI THÁC DỮ LIỆU – TUẦN 8

Họ tên : Nguyễn Tiến Phong
MSSV : 20280071

Thuật toán Mahalanobis k-means :

- Tạo dữ liệu với hàm make_blobs và vẽ scatterplot

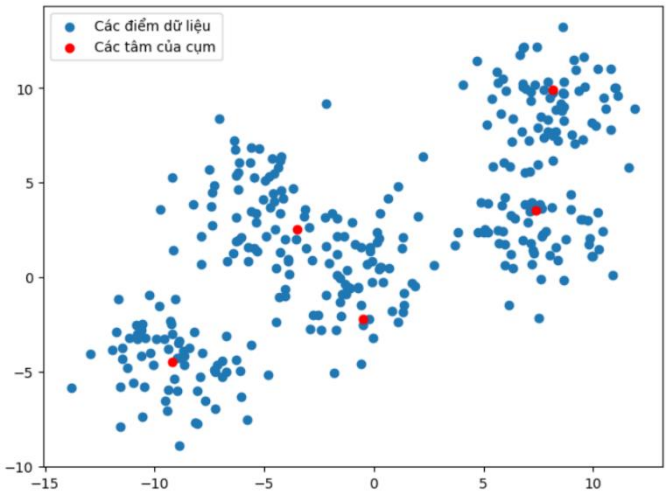


- Hàm tính khoảng cách Mahalanobis theo công thức :

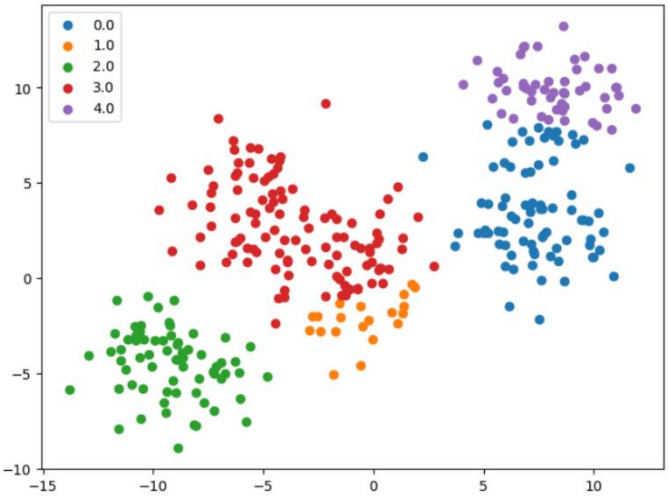
$$Maha(\overline{X}, \overline{\mu_r}, \overline{\Sigma_r}) = \sqrt{(\overline{X} - \overline{\mu_r})\sigma_r^{-1}(\overline{X} - (\overline{\mu_r})^T}$$

- Hàm thuật toán Mahalanobis k-means :
  1. Khởi tạo tâm cụm ngẫu nhiên
  2. Tính toán khoảng cách Mahalanobis đến các centroid trên, gán cluster cho từng điểm dữ liệu
  3. Cập nhật lại tâm cụm bằng cách lấy trung bình cộng các điểm thuộc cluster đó
  4. Lặp lại bước 2,3 đến khi kết quả hai lần chạy liên tiếp không thay đổi

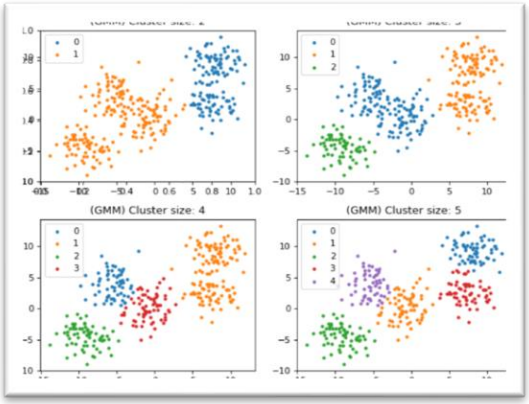
- Vẽ scatterplot với các tâm của từng cụm (số cụm = 5)



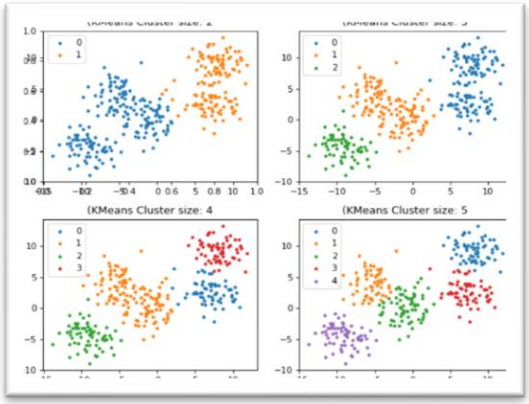
- Vẽ scatterplot phân biệt các cụm



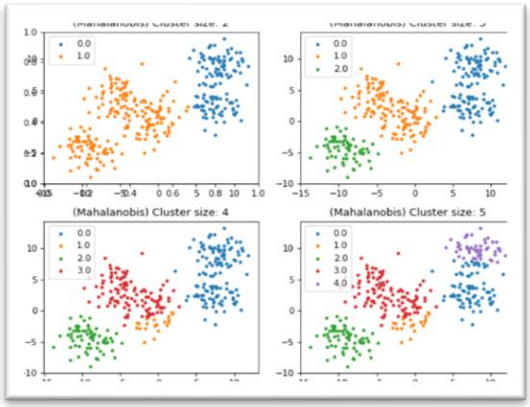
- **Scatterplot của các thuật toán khi số cụm từ 2 → 5 :**



GGM



KMeans



Mahalanobis KMeans

- **Nhận xét :**
  - **GGM :** Mạnh mẽ đối với các dữ liệu phức tạp và không tuân theo phân phối chuẩn, khả năng ước lượng các tham số của phân phối Gaussian. Tuy nhiên, độ phức tạp tính toán cao, phải ước lượng số lượng cụm cần tính toán.

  - **K-Means :** Đơn giản dễ hiểu, phù hợp với dữ liệu lớn. Tuy nhiên, mô hình không xử lý được dữ liệu có độ lệch và các đặc trưng không tuân theo phân phối chuẩn.

  - **Mahalanobis K-Means :** Sử dụng khoảng cách Mahalanobis xử lý các trường hợp không tuân theo phân phối chuẩn, tích hợp sự tương quan và phân bố của dữ liệu vào quá trình gom cụm. Tuy nhiên, mô hình đòi hỏi tính toán ma trận hiệp phương sai cho từng cụm tốn tài nguyên hơn và cần dữ liệu đủ lớn để ước lượng đúng ma trận hiệp phương sai.