BÀI 6 - PHÂN CỤM DỰA TRÊN MẬT ĐỘ DBSCAN

Hạn nộp bài: hết ngày 16/05/2024.

Bài 1. Dùng `sklearn.datasets.make_circles` để tạo ra 1000 điểm dữ liệu có dạng hình tròn với `factor=0.1`, `noise=0.1`.

a) Thực hiện phân cụm bằng DBSCAN trong `sklearn`. Trực quan hoá và nhận xét.

b) Điều chỉnh các tham số `eps` và `min_samples` khi phân cụm. Trực quan hoá và nhận xét. So sánh với K-means và GMM.

c) Viết hàm `DBSCAN_clustering(data, eps, minPts)` để thực hiện phân cụm dựa trên mật độ cho tập dữ liệu `data` theo các bước sau:

Bước 1. Viết hàm `classify_points(data)` để phân loại điểm nhân, điểm biên và điểm nhiễu trong tập dữ liệu;

Bước 2. Viết hàm `density_reach(data, i)` để tìm danh sách các điểm density-reachable từ điểm thứ `i` trong `data`;

Bước 3. Sử dụng 2 hàm trên để viết hàm phân cụm.

d) Sử dụng hàm `DBSCAN_clustering` vừa viết để phân cụm cho dữ liệu đã tạo từ đầu bài và so sánh với DBSCAN trong `sklearn`.

Bài 2. Dùng `sklearn.datasets.make_moons` để tạo ra 1000 điểm dữ liệu có dạng hình mặt trăng với `noise=0.1`.

a) Thực hiện phân cụm bằng DBSCAN. Trực quan hoá và nhận xét.

b) Điều chỉnh các tham số `eps` và `min_samples` để được kết quả thích hợp. Trực quan hoá và nhận xét.

Bài 3. Bộ dữ liệu shopping-data bao gồm các quan sát về giới tính, độ tuổi, thu nhập và điểm chi tiêu của 200 khách hàng. Ta cần phân cụm tập khách hàng này vào những nhóm có chung đặc tính và hành vi mua sắm để chăm sóc và phục vụ họ tốt hơn.

a) Đọc dữ liệu và tiền xử lý dữ liệu.

b) Với $minPts = 11$, hãy chọn ε thích hợp dựa vào biểu đồ k-distance.

c) Với tham số đã chọn, thực hiện phân cụm bằng DBSCAN. Nhận xét.

d) Phân cụm dữ liệu trên bằng K-means và GMM. So sánh kết quả với câu c.

Bài 4. Dùng `sklearn.datasets.make_blobs` để tạo ra ma trận X có 1500 điểm dữ liệu thuộc 3 cụm khác nhau.

a) Cho ma trận

$$A = \begin{bmatrix} 0.6 & -0.6 \\ -0.4 & 0.8 \end{bmatrix}.$$

Khi đó ma trận $X_1 = X \cdot A$ sẽ có phân bố dị hướng. Thực hiện phân cụm bằng DBSCAN cho X_1. Trực quan hoá và so sánh với K-means và GMM.

b) Tạo ma trận X_2 từ X, trong đó lấy 500 điểm thuộc cụm 0, 100 điểm thuộc cụm 1 và 10 điểm thuộc cụm 2. Thực hiện phân cụm bằng DBSCAN cho X_2. Trực quan hoá và so sánh với K-means và GMM.

c) Tạo ra ma trận X_3 có 1500 điểm dữ liệu thuộc 3 cụm khác nhau với độ lệch chuẩn khác nhau: $[1.0, 2.5, 0.5]$. Thực hiện phân cụm bằng DBSCAN cho X_3. Trực quan hoá và so sánh với K-means và GMM.