BÀI 5 - PHÂN CỤM

Hạn nộp bài: hết ngày 06/05/2024.

Bài 1. Dùng `sklearn.datasets.make_blobs` để tạo ra 1000 điểm dữ liệu thuộc 4 cụm khác nhau.

a) Viết hàm `kmean_clustering` để thực hiện phân cụm:

- Đầu vào: Ma trận dữ liệu và số cụm;
- Đầu ra: Kết quả phân cụm và toạ độ các tâm cụm.

b) Dùng hàm `kmean_clustering` để phân cụm cho 1000 điểm dữ liệu đã tạo. Trực quan hoá kết quả và nhận xét.

c) Ở đầu vào của hàm `kmean_clustering`, hãy thêm tuỳ chọn `init` để chọn cách khởi tạo tâm cụm. Hai tuỳ chọn là `'random'` (khởi tạo ngẫu nhiên) và `'kmeans++'` (khởi tạo dựa trên phân phối thực nghiệm của dữ liệu).

Gợi ý: dùng hàm `sklearn.cluster.kmeans_plusplus`.

d) Với mỗi tuỳ chọn khởi tạo, thực hiện phân cụm 10 lần. Hãy trực quan hoá bước khởi tạo và đo số vòng lặp, thời gian thực hiện. Nhận xét.

Bài 2. Dùng `sklearn.datasets.make_circles` để tạo ra 1000 điểm dữ liệu có dạng hình tròn với `factor=0.1`, `noise=0.1`.

a) Với $K = 2$, hãy thực hiện phân cụm bằng K-means và GMM trong `sklearn`. Trực quan hoá và nhận xét.

b) Thực hiện lại với $K \in \{3; 4; 5; 6; 7\}$. Trực quan hoá và nhận xét.

Bài 3. Tập dữ liệu Iris (trong `sklearn.datasets.load_iris`) chứa 150 quan trắc về các bông hoa diên vĩ. Giả sử chưa có thông tin về phân loại của từng bông hoa.

a) Trực quan hoá dữ liệu trong hai chiều và nhận xét về số cụm thích hợp để phân cụm.

b) Thực hiện phân cụm bằng K-means với số cụm vừa chọn. Trực quan hoá và nhận xét.

c) Thực hiện phân cụm bằng K-means với số cụm $K \in \{2; \ldots; 10\}$. Trực quan hoá và nhật xét. Thực hiện cross validation với scoring thích hợp để chọn số cụm phù hợp nhất.

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

a) Thực hiện phân cụm bằng K-means và GMM cho X. Trực quan hoá và so sánh.

b) 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 K-means và GMM cho X_1. Trực quan hoá và so sánh.

c) 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 K-means và GMM cho X_2. Trực quan hoá và so sánh.

Bài 5. Dùng K-means để nén ảnh `china` trong `sklearn.datasets.load_sample_image` theo hướng dẫn sau.

a) Trực quan hoá và xem kích thước ảnh. Thay đổi kích thước ảnh thành $(rows * cols, 3)$.

b) Dùng K-means để phân cụm các màu trong ảnh gốc thành 3 cụm. Mỗi điểm ảnh sẽ được thay bằng tâm cụm của cụm nó thuộc về.

c) Làm tròn các toạ độ tâm cụm và tái tạo lại kích thước gốc. Trực quan hoá và nhận xét.

d) Chọn số cụm nhỏ nhất có thể để phân tách rõ các vùng trong ảnh gốc.