KMeans

Nguyễn Văn Huy & Lê Duy An

Ngày 27 tháng 7 năm 2020

Mục lục

1 Giới thiệu

2 Phân tích toán học

Với dữ liệu đầu vào của thuật toán là tập hợp các điểm dữ liệu $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ với $\mathbf{x}_i$ (có d phần tử) là một vector mang giá trị của mỗi điểm, N là số lượng các vector và số lượng K các nhóm cần phân loại từ các điểm dữ liệu đó với $K < N$ (vì số lượng nhóm cần phân loại không được lớn hơn số lượng các phần tử). Điều mà chúng ta cần phải làm là làm thế nào để xác định các điểm thuộc về nhóm nào một cách gắn kết nhất, ở đây để cho dễ gọi và tính toán thì chúng ta cho rằng K nhóm cần phần loại được gọi là nhóm $1, 2, 3, ..K$. Trong phần này chúng ta chỉ đề cập đến bài toán chỉ có một điểm dữ liệu thuộc vào một nhóm duy nhất. Ban đầu chúng ta phải có được các điểm gốc ban đầu của các nhóm có thể chọn k điểm bất kì hoặc có thể lấy các điểm dữ liệu có sẵn trong tập dữ liệu ban đầu. Gọi các điểm gốc ban đầu là $\mathbf{m} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, ..., \mathbf{m}_K]$ với mỗi điểm $\mathbf{m}_k$ cũng có có d các giá trị tương tự như các điểm dữ liệu $\mathbf{x}_i$. Dựa vào tập các điểm gốc $\mathbf{m}_k$ chúng ta phải xác định xem điểm $\mathbf{x}_i$ thuộc vào nhóm nào và gán nhãn cho các điểm đó bằng vector $\mathbf{y}$ trong đó $\mathbf{y}_{ij} = 0$ và $\mathbf{y}_{ik} = 1$, nghĩa là vector $\mathbf{y}$ có K giá trị và vị trí ở vị trí k có giá trị bằng 1 thì đồng nghĩa là vector $\mathbf{x}_i$ được gán vào nhóm k.

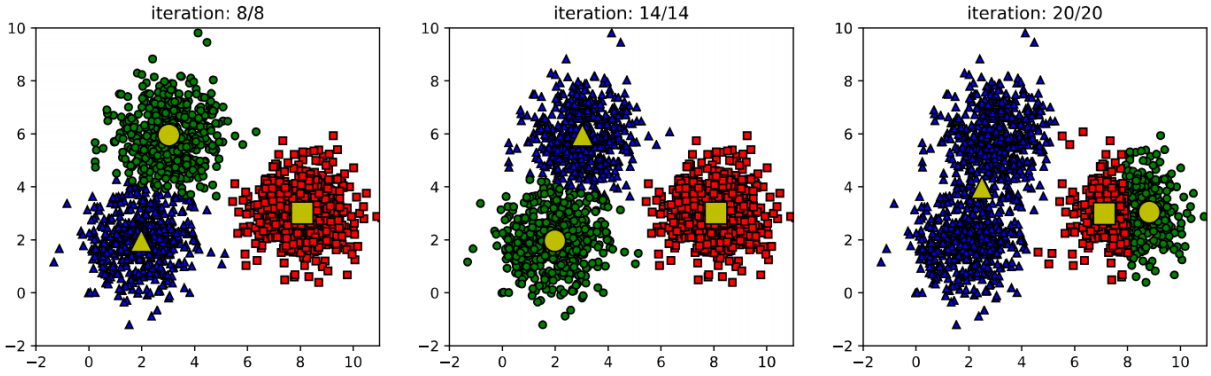
3 Ưu và nhược điểm.

3.1 Ưu điểm

- Thuật toán đơn giản, hiệu quả
- Sử dụng được với bộ số liệu lớn

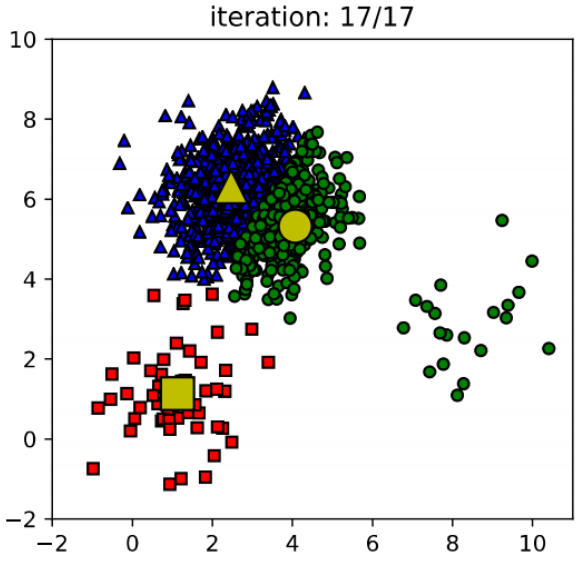
3.2 Khuyết điểm

- Cần phải xác định trước số lượng cluster. Trong thực tế, cần phải sử dụng thêm một số biện pháp giúp xác định giá trị K, chẳng hạn như elbow method.
- Thuật toán KMeans không đảm bảo tìm được nghiệm tối ưu toàn cục nên nghiệm cuối cùng phụ thuộc rất nhiều vào các centroid ban đầu.



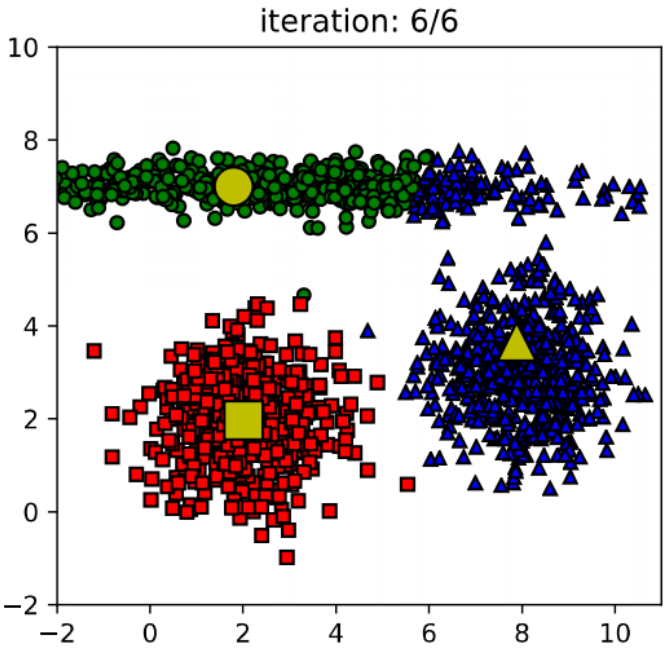
Hình 1: Các nghiệm khác nhau do khởi tạo ban đầu khác nhau

- Các cluster cần phải có số lượng điểm gần bằng nhau



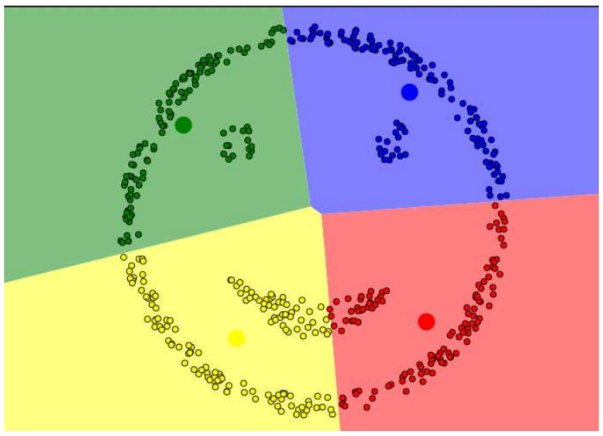
Hình 2: Các nghiệm trong cluster này bị nhầm vào cluster khác

- Các cluster cần có dạng hình tròn (cầu)



Hình 3: Các nghiệm trong cluster này bị nhầm vào cluster khác

- Centroid có thể bị xê dịch bởi các ngoại lệ, hoặc các ngoại lệ có thể có cụm riêng thay vì bị bỏ qua.
- Cho kết quả sai khi một cluster này bị bao bọc bằng một cluster khác



Hình 4: KMeans chia hình mặt làm 4 phần thay vì gom chung các bộ phận trong khuôn mặt thành 1 cụm

4 Cách tìm K cụm tối ưu nhất