

Phân cụm với k-means

Unsupervised learning (học không giám sát)

Trong thuật toán K-means clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho *dữ liệu trong cùng một cụm có tính chất giống nhau*.

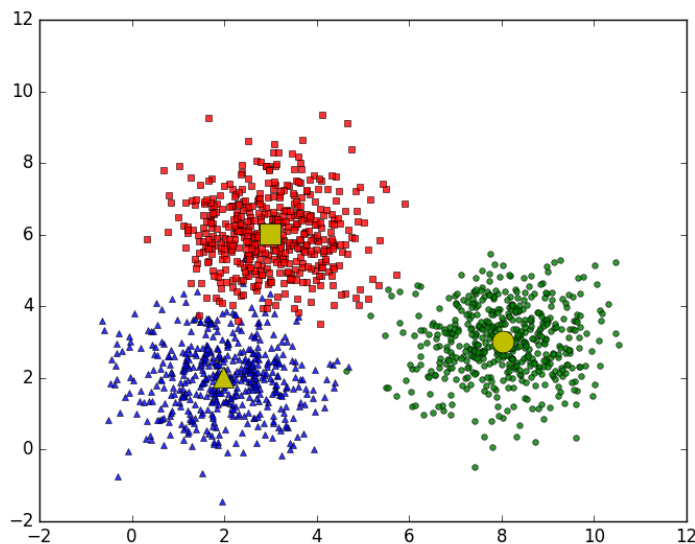
Ví dụ: Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau dựa trên sự tương tác giữa mỗi khách hàng với công ty đó

+ Giả sử công ty đó có rất nhiều dữ liệu của rất nhiều khách hàng nhưng chưa có cách nào chia toàn bộ khách hàng đó thành một số nhóm/cụm khác nhau.

+ Nếu một người biết Machine Learning được đặt câu hỏi này, phương pháp đầu tiên anh (chị) ta nghĩ đến sẽ là K-means Clustering

+ Sau khi đã phân ra được từng nhóm, nhân viên công ty đó có thể lựa chọn ra một vài khách hàng trong mỗi nhóm để quyết định xem mỗi nhóm tương ứng với nhóm khách hàng nào. Phần việc cuối cùng này cần sự can thiệp của con người, nhưng lượng công việc đã được rút gọn đi rất nhiều.

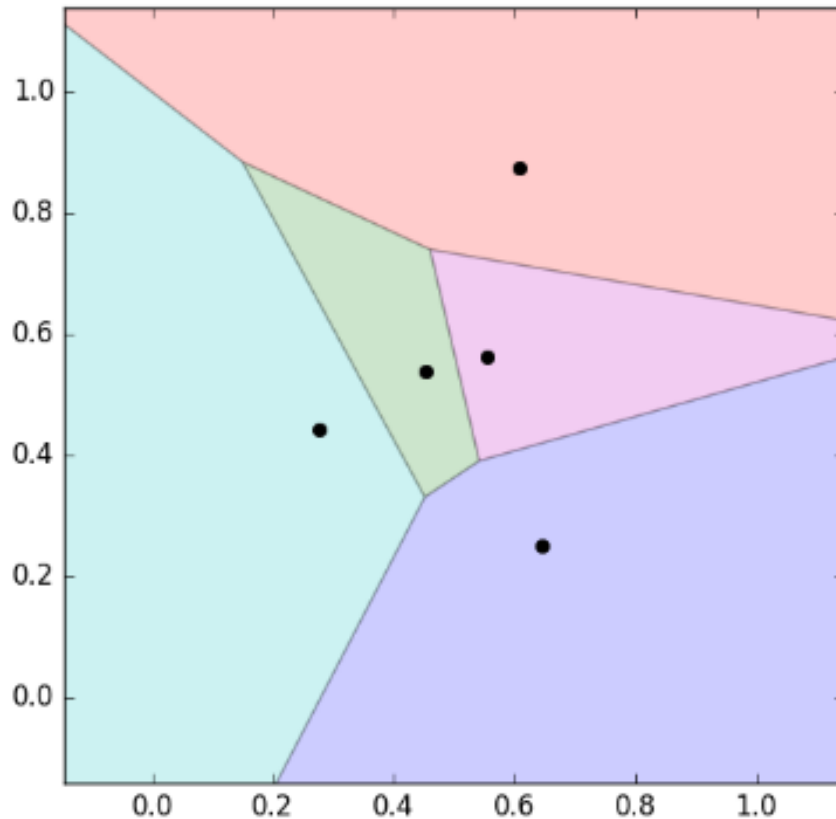
Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở *gần nhau trong một không gian nào đó* (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu (từ giờ tôi sẽ viết gọn là *cluster*).



Giả sử mỗi cluster có một điểm đại diện (*center*) màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

Tới đây, chúng ta có một bài toán thú vị: *Trên một vùng biển hình vuông lớn có ba đảo hình vuông, tam giác, và tròn màu vàng như hình trên. Một điểm trên biển được gọi là thuộc lãnh hải của một đảo nếu nó nằm gần đảo này hơn so với hai đảo kia . Hãy xác định ranh giới lãnh hải của các đảo.*

Hình dưới đây là một hình minh họa cho việc phân chia lãnh hải nếu có 5 đảo khác nhau được biểu diễn bằng các hình tròn màu đen:



Chúng ta thấy rằng đường phân định giữa các lãnh hải là các đường thẳng (chính xác hơn thì chúng là các đường trung trực của các cặp điểm gần nhau). Vì vậy, lãnh hải của một đảo sẽ là một hình đa giác.

2. Phân tích toán học

Mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm chúng ta muốn tìm, hãy chỉ ra center của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng. Giả sử thêm rằng mỗi điểm dữ liệu chỉ thuộc vào đúng một nhóm.

Một số ký hiệu toán học

Giả sử có N điểm dữ liệu là $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ và $K < N$ là số cluster chúng ta muốn phân chia. Chúng ta cần tìm các center $\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_K \in \mathbb{R}^{d \times 1}$ và label của mỗi điểm dữ liệu.

Lưu ý về ký hiệu toán học: trong các bài viết của tôi, các số vô hướng được biểu diễn bởi các chữ cái viết ở dạng không in đậm, có thể viết hoa, ví dụ x_1, N, y, k . Các vector được biểu diễn bằng các chữ cái thường in đậm, ví dụ \mathbf{m}, \mathbf{x}_1 . Các ma trận được biểu diễn bởi các chữ viết hoa in đậm, ví dụ $\mathbf{X}, \mathbf{M}, \mathbf{Y}$. Lưu ý này đã được nêu ở bài [Linear Regression](#). Tôi xin được không nhắc lại trong các bài tiếp theo.

Với mỗi điểm dữ liệu \mathbf{x}_i đặt $\mathbf{y}_i = [y_{i1}, y_{i2}, \dots, y_{iK}]$ là label vector của nó, trong đó nếu \mathbf{x}_i được phân vào cluster k thì $y_{ik} = 1$ và $y_{ij} = 0, \forall j \neq k$. Điều này có nghĩa là có đúng một phần tử của vector \mathbf{y}_i là bằng 1 (tương ứng với cluster của \mathbf{x}_i), các phần tử còn lại bằng 0. Ví dụ: nếu một điểm dữ liệu có label vector là $[1, 0, 0, \dots, 0]$ thì nó thuộc vào cluster 1, là $[0, 1, 0, \dots, 0]$ thì nó thuộc vào cluster 2, \dots . Cách mã hóa label của dữ liệu như thế này được gọi là biểu diễn *one-hot*. Chúng ta sẽ thấy cách biểu diễn one-hot này rất phổ biến trong Machine Learning ở các bài tiếp theo.

Ràng buộc của \mathbf{y}_i có thể viết dưới dạng toán học như sau:

$$y_{ik} \in \{0, 1\}, \quad \sum_{k=1}^K y_{ik} = 1 \quad (1)$$

Hàm mất mát và bài toán tối ưu

Nếu ta coi center \mathbf{m}_k là center (hoặc representative) của mỗi cluster và ước lượng tất cả các điểm được phân vào cluster này bởi \mathbf{m}_k , thì một điểm dữ liệu \mathbf{x}_i được phân vào cluster k sẽ bị sai số là $(\mathbf{x}_i - \mathbf{m}_k)$. Chúng ta mong muốn sai số này có trị tuyệt đối nhỏ nhất nên (giống như trong bài [Linear Regression](#)) ta sẽ tìm cách để đại lượng sau đây đạt giá trị nhỏ nhất:

$$\|\mathbf{x}_i - \mathbf{m}_k\|_2^2$$

Hơn nữa, vì \mathbf{x}_i được phân vào cluster k nên $y_{ik} = 1, y_{ij} = 0, \forall j \neq k$. Khi đó, biểu thức bên trên sẽ được viết lại là:

$$y_{ik} \|\mathbf{x}_i - \mathbf{m}_k\|_2^2 = \sum_{j=1}^K y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2$$

(Hy vọng chỗ này không quá khó hiểu)

Tài liệu tham khảo

Web: <https://machinelearningcoban.com/2017/01/01/kmeans/#-gioi-thieu>