

Topic Lecture: Clustering

Trung Can, Long Nguyen, Ngoc Pham, Tuan Pham

PiMA 2021



Trình bày: Trung Can, Caltech

July 17, 2021

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng

Nhắc lại

Định nghĩa vắn tắt (data science)

Data science thu thập và xử lý thông tin từ môi trường \mathcal{M} để được dữ liệu \mathcal{D} và tìm mô hình A để giải thích \mathcal{D}



Nhắc lại

Định nghĩa vắn tắt (data science)

Data science thu thập và xử lý thông tin từ môi trường \mathcal{M} để được dữ liệu \mathcal{D} và tìm mô hình A để giải thích \mathcal{D}

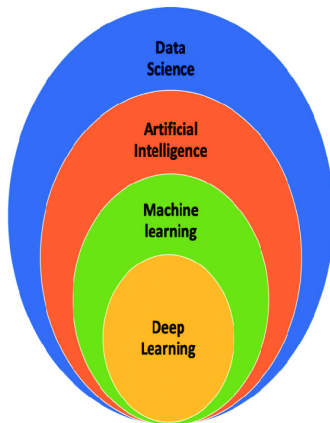
Định nghĩa vắn tắt (machine learning)

Machine learning (máy học) có thể hiểu là các thuật toán giúp máy tính tìm hàm số f trong tập hợp hàm ứng viên \mathcal{H} để giải thích dữ liệu \mathcal{D} hợp lý nhất (tối ưu hóa $L(\mathcal{D}, f)$).



DS and ML

Data science \supset machine learning



Contents

- 1 Machine learning trong data science
 - Học có giám sát
 - Học không giám sát
- 2 Bài toán clustering
 - Một số ví dụ mở đầu
 - Định nghĩa
 - k-means và tối ưu hóa
- 3 Phân loại các thuật toán clustering
 - Dựa trên output
 - Dựa trên giả sử của mô hình
- 4 Kết luận
 - Thực hành và áp dụng



Các loại bài toán học

Bài toán học có giám sát (supervised learning): từ \mathcal{D} là dữ liệu đã được "dán nhãn" (labelled), muốn học cách dán nhãn dữ liệu mới



Các loại bài toán học

Bài toán học có giám sát (supervised learning): từ \mathcal{D} là dữ liệu đã được "dán nhãn" (labelled), muốn học cách dán nhãn dữ liệu mới

Ví dụ: muốn máy tính học cách phân loại email mới từ dữ liệu các emails cũ đã phân loại sẵn thành spam hoặc không phải spam (nhãn dán ở đây là 0 hoặc 1)



Các loại bài toán học

Bài toán học có giám sát (supervised learning): từ \mathcal{D} là dữ liệu đã được "dán nhãn" (labelled), muốn học cách dán nhãn dữ liệu mới

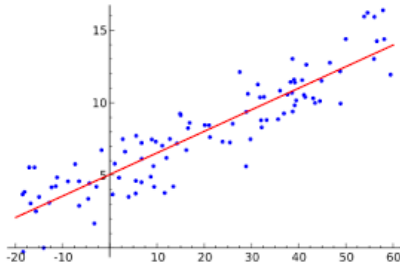
Ví dụ: muốn máy tính học cách phân loại email mới từ dữ liệu các emails cũ đã phân loại sẵn thành spam hoặc không phải spam (nhãn dán ở đây là 0 hoặc 1)

Các bài toán tiêu biểu: hồi quy (regression), phân loại (classification)



Câu hỏi trên lớp

Bài toán tìm đường thẳng least square của một tập hợp điểm $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\} \subset \mathbb{R}^2$ là một ví dụ của hồi quy tuyến tính (linear regression). Nhấn dấn trong bài toán này là gì?



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Các loại bài toán học

Bài toán học không giám sát (unsupervised learning): \mathcal{D} là dữ liệu chưa dán nhãn (unlabelled)



Các loại bài toán học

Bài toán học không giám sát (unsupervised learning): \mathcal{D} là dữ liệu chưa dán nhãn (unlabelled)

Ví dụ: muốn máy tính khoanh vùng dịch covid từ vị trí của các ca bệnh



Các loại bài toán học

Bài toán học không giám sát (unsupervised learning): \mathcal{D} là dữ liệu chưa dán nhãn (unlabelled)

Ví dụ: muốn máy tính khoanh vùng dịch covid từ vị trí của các ca bệnh

Các bài toán tiêu biểu: giảm chiều dữ liệu (dimensionality reduction), phân cụm (clustering)



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Tại sao cần clustering?

Ví dụ 1: Bài toán đặt nhà kho chứa hàng hóa (warehouse location problem)



Tại sao cần clustering?

Ví dụ 1: Bài toán đặt nhà kho chứa hàng hóa (warehouse location problem)



Tại sao cần clustering?

Ví dụ 1: Bài toán đặt nhà kho chứa hàng hóa (warehouse location problem)



Tổng quát: Bài toán đặt cơ sở bất kì (facility location problem)



Warehouse location problem

Bài toán không đơn giản do nhiều dữ liệu thông tin:

- rất nhiều đại lý bán lẻ,
- số lượng, vị trí đặt nhà kho,
- chi phí giao hàng đến các đại lý bán, v.v.

Một phương án khả thi là chia bài toán làm 2 bước:

- 1 phân cụm các đại lý một cách hợp lý, mỗi cụm sẽ được phân phối bởi một nhà kho (**clustering**)



Warehouse location problem

Bài toán không đơn giản do nhiều dữ liệu thông tin:

- rất nhiều đại lý bán lẻ,
- số lượng, vị trí đặt nhà kho,
- chi phí giao hàng đến các đại lý bán, v.v.

Một phương án khả thi là chia bài toán làm 2 bước:

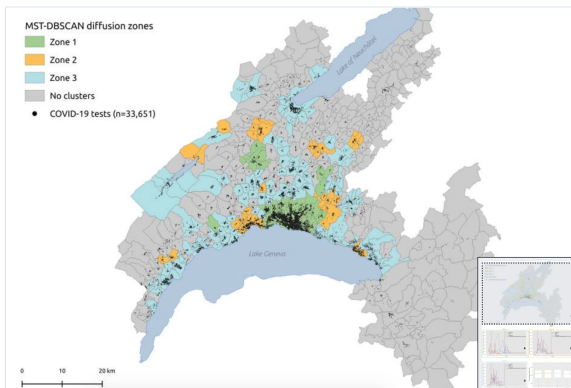
- 1 phân cụm các đại lý một cách hợp lý, mỗi cụm sẽ được phân phối bởi một nhà kho (**clustering**)
- 2 trong từng cụm, sắp xếp lịch phân phối hàng một cách hợp lý nhất (bài toán travelling salesman, bài toán vehicle routing)

[Link tham khảo](#)



Tại sao cần clustering?

Ví dụ 2: Phân vùng dịch bệnh để phòng tránh và đối phó



Một phương án đề xuất xuất gồm 2 bước:

- 1 Phân cụm các điểm bệnh (**clustering**) để xác định các vùng chứa các cụm bệnh
- 2 Phân loại các vùng bệnh để xử lý cho phù hợp: Zone 1, Zone 2, Zone 3 (**classification**)

[Link tham khảo](#)



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Định nghĩa vắn tắt

Bài toán data clustering là bài toán phân cụm cho tập dữ liệu đầu vào một cách hợp lý nhất.

Tập dữ liệu đầu vào: $D = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$

Tập các cụm dữ liệu đầu ra: C_1, C_2, \dots, C_k

Câu hỏi: Như thế nào là hợp lý nhất?

Trả lời:

- Không biết :(



Định nghĩa vắn tắt

Bài toán data clustering là bài toán phân cụm cho tập dữ liệu đầu vào một cách hợp lý nhất.

Tập dữ liệu đầu vào: $D = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$

Tập các cụm dữ liệu đầu ra: C_1, C_2, \dots, C_k

Câu hỏi: Như thế nào là hợp lý nhất?

Trả lời:

- Không biết :(nếu không có thêm những cấu trúc toán học :D



Định nghĩa vắn tắt

Bài toán data clustering là bài toán phân cụm cho tập dữ liệu đầu vào một cách hợp lý nhất.

Tập dữ liệu đầu vào: $D = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$

Tập các cụm dữ liệu đầu ra: C_1, C_2, \dots, C_k

Câu hỏi: Như thế nào là hợp lý nhất?

Trả lời:

- Không biết :(nếu không có thêm những cấu trúc toán học :D
- Giống như bài toán ML tổng quát, cần thêm các giả sử (assumptions) \Rightarrow cấu trúc toán học \Rightarrow bài toán tối ưu hóa



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Bài toán k-means: Cho trước $\mathcal{D} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$ và số tự nhiên k . Tìm một phân hoạch S_1, S_2, \dots, S_k của \mathcal{D} sao cho hàm số $L = \sum_j \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$ đạt min với $\mu_j = (\sum_{x_i \in S_j} x_i) / |S_j|$.



Bài toán k-means: Cho trước $\mathcal{D} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$ và số tự nhiên k . Tìm một phân hoạch S_1, S_2, \dots, S_k của \mathcal{D} sao cho hàm số $L = \sum_j \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$ đạt min với $\mu_j = (\sum_{x_i \in S_j} x_i) / |S_j|$.



Bài toán k-means: Cho trước $\mathcal{D} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$ và số tự nhiên k . Tìm một phân hoạch S_1, S_2, \dots, S_k của \mathcal{D} sao cho hàm số $L = \sum_j \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$ đạt min với $\mu_j = (\sum_{x_i \in S_j} x_i) / |S_j|$.

Nhận xét: Bài toán tối ưu tổ hợp khó (không có công thức nghiệm chuẩn tắc)



Bài toán k-means: Cho trước $\mathcal{D} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$ và số tự nhiên k . Tìm một phân hoạch S_1, S_2, \dots, S_k của \mathcal{D} sao cho hàm số $L = \sum_j \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$ đạt min với $\mu_j = (\sum_{x_i \in S_j} x_i) / |S_j|$.

Nhận xét: Bài toán tối ưu tổ hợp² khó (không có công thức nghiệm chuẩn tắc)

- Không gian hàm $\mathcal{H} = \{f : \mathcal{D} \rightarrow \{1, \dots, k\}\}$ rời rạc, có độ lớn lũy thừa k^n



Bài toán k-means: Cho trước $\mathcal{D} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^n$ và số tự nhiên k . Tìm một phân hoạch S_1, S_2, \dots, S_k của \mathcal{D} sao cho hàm số $L = \sum_j \sum_{x_i \in S_j} \|x_i - \mu_j\|^2$ đạt min với $\mu_j = (\sum_{x_i \in S_j} x_i) / |S_j|$.

Nhận xét: Bài toán tối ưu tổ hợp² khó (không có công thức nghiệm chuẩn tắc)

- Không gian hàm $\mathcal{H} = \{f : \mathcal{D} \rightarrow \{1, \dots, k\}\}$ rời rạc, có độ lớn lũy thừa k^n
- Hàm mất mát L không có tham số tường minh



Giải pháp: Mẹo để tham số hóa các cấu trúc rời rạc



Giải pháp: Mẹo để tham số hóa các cấu trúc rời rạc

Biến chỉ thị (indicator variable):

Đặt $w_{ij} = 1$ nếu $x_i \in C_j$ (ký hiệu chuẩn là $\mathbb{1}(x_i \in C_j)$)



Giải pháp: Mẹo để tham số hóa các cấu trúc rời rạc

Biến chỉ thị (indicator variable):

Đặt $w_{ij} = 1$ nếu $x_i \in C_j$ (ký hiệu chuẩn là $\mathbb{1}(x_i \in C_j)$)

Xem L là hàm số $L(w_{ij}, \mu_j)$ và áp dụng một mô hình tối ưu kỳ vọng (expectation maximization) gồm 2 bước:

- 1 (E) $w_{ij} := 1$ nếu $j = \operatorname{argmin}_k \|x_i - \mu_k\|$
- 2 (M) $\mu_j := (\sum_i w_{ij} x_i) / (\sum_i w_{ij})$



Giải pháp: Mẹo để tham số hóa các cấu trúc rời rạc

Biến chỉ thị (indicator variable):

Đặt $w_{ij} = 1$ nếu $x_i \in C_j$ (ký hiệu chuẩn là $\mathbb{1}(x_i \in C_j)$)

Xem L là hàm số $L(w_{ij}, \mu_j)$ và áp dụng một mô hình tối ưu kỳ vọng (expectation maximization) gồm 2 bước:

1 (E) $w_{ij} := 1$ nếu $j = \operatorname{argmin}_k \|x_i - \mu_k\|$

2 (M) $\mu_j := (\sum_i w_{ij} x_i) / (\sum_i w_{ij})$

Nhận xét: đây chính là thuật toán k-means cơ bản



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Hard clustering

Output: Mỗi điểm dữ x_i liệu thuộc đúng một cụm duy nhất C_j hoặc không thuộc cụm nào, gọi là các điểm ngoài lề (outliers).

Ví dụ: k-means cơ bản, mean-shift clustering, spectral clustering



Soft clustering

Output: Mỗi điểm dữ liệu x_i thuộc cụm C_j với xác suất p_{ij} .

Ví dụ: k-means có trọng số (weighted), mô hình hỗn hợp Gaussian (Gaussian mixture model)

Câu hỏi trên lớp

Các giá trị $\{p_{ij}\}$ phải thỏa mãn điều kiện gì?



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Các giả sử về dữ liệu và cụm hợp lý sẽ giới hạn không gian tìm kiếm \mathcal{H} . Một số nhóm phổ biến:

- 1 phân cụm dựa trên trọng tâm (centroid-based)
- 2 phân cụm dựa trên tính liên thông (connectivity-based)
- 3 phân cụm dựa trên mật độ dữ liệu (density-based)
- 4 phân cụm dựa trên phân phối của dữ liệu (distribution-based)



Cụm trọng tâm

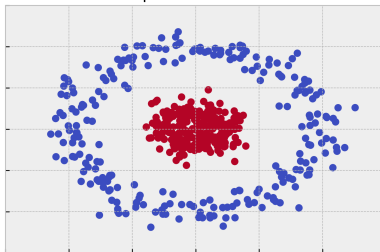
Giả sử: Mỗi cụm được đại diện bởi một trọng tâm và các điểm dữ liệu được phân cụm dựa trên các trọng tâm này. Ví dụ: k-means



Cụm liên thông

Giả sử: Các điểm giống nhau sẽ ở chung cụm và sự chung cụm có tính bắc cầu (transitive). Ví dụ: spectral clustering (spectral ở đây là từ spectral graph theory)

Spectral Circles



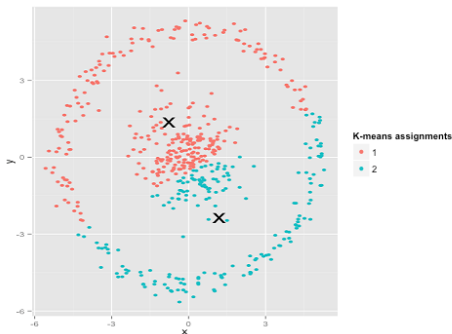
Visualization



Cụm liên thông

Câu hỏi trên lớp

Tại sao k -means với $k = 2$ không phân cụm được vòng tròn xanh bên ngoài trong dữ liệu ở dưới?



Cụm mật độ

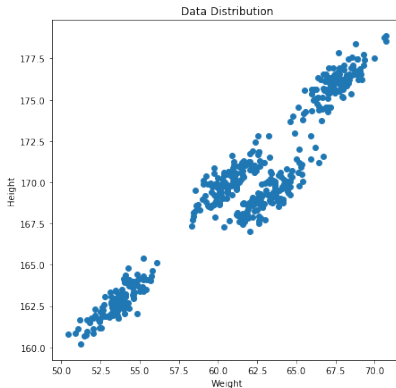
Giả sử: Các cụm được tạo thành từ các vùng có mật độ điểm dữ liệu dày đặc. Ví dụ: mean-shift

Visualization

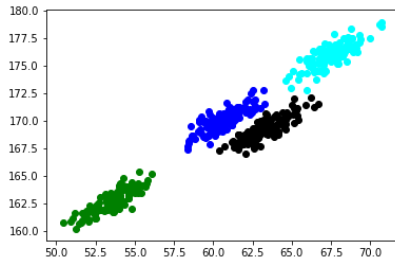
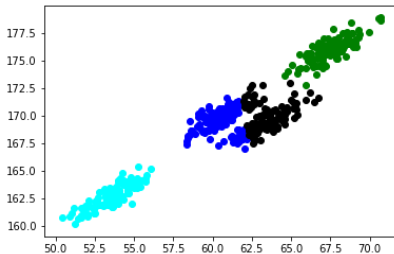


Cụm phân phối

Giả sử: Các điểm dữ liệu được lấy ngẫu nhiên (random sampling) theo một phối xác suất nào đó. Ví dụ: Gaussian mixture model



Cụm phân phối



Một số ý chính

- Clustering thuộc nhóm thuật toán học không giám sát



Một số ý chính

- Clustering thuộc nhóm thuật toán học không giám sát
- Có nhiều mô hình clustering khác nhau dựa trên những giả sử khác nhau



Một số ý chính

- Clustering thuộc nhóm thuật toán học không giám sát
- Có nhiều mô hình clustering khác nhau dựa trên những giả sử khác nhau
- Đưa về một bài toán tối ưu hóa không tầm thường \Rightarrow sử dụng các phương pháp xấp xỉ



Một số ý chính

- Clustering thuộc nhóm thuật toán học không giám sát
- Có nhiều mô hình clustering khác nhau dựa trên những giả sử khác nhau
- Đưa về một bài toán tối ưu hóa không tầm thường \Rightarrow sử dụng các phương pháp xấp xỉ
- Mỗi mô hình có những ưu và nhược điểm riêng \Rightarrow xây dựng các phiên bản cải tiến hơn và các mô hình khác



Một số ý chính

- Clustering thuộc nhóm thuật toán học không giám sát
- Có nhiều mô hình clustering khác nhau dựa trên những giả sử khác nhau
- Đưa về một bài toán tối ưu hóa không tầm thường \Rightarrow sử dụng các phương pháp xấp xỉ
- Mỗi mô hình có những ưu và nhược điểm riêng \Rightarrow xây dựng các phiên bản cải tiến hơn và các mô hình khác
- Áp dụng mô hình nào phụ thuộc vào dữ liệu và mục đích cụ thể trong ứng dụng



Contents

1 Machine learning trong data science

- Học có giám sát
- Học không giám sát

2 Bài toán clustering

- Một số ví dụ mở đầu
- Định nghĩa
- k-means và tối ưu hóa

3 Phân loại các thuật toán clustering

- Dựa trên output
- Dựa trên giả sử của mô hình

4 Kết luận

- Thực hành và áp dụng



Một số vấn đề chưa được cover



Một số vấn đề chưa được cover

- Khởi tạo các siêu tham số một cách hợp lý (ví dụ: chọn k và vị trí k trọng tâm ban đầu cho k -means)



Một số vấn đề chưa được cover

- Khởi tạo các siêu tham số một cách hợp lý (ví dụ: chọn k và vị trí k trọng tâm ban đầu cho k -means)
- Đánh giá độ hiệu quả của thuật toán cho một dữ liệu cụ thể (ví dụ: homogeneity score, completeness score)

