

Mean Shift Clustering

Hoài An, Lương Thắng, Quỳnh Anh, Nhật Nam

PiMA 2021



Trình bày: Nhóm 5, Mean Shift Clustering

August 8, 2021

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

- Tổng quan
- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm

Contents

1 Bài toán phân cụm dữ liệu

■ Giới thiệu

■ Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

■ Động lực

■ Kernel

■ Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

■ Các bước Mean Shift

■ Ý nghĩa toán học

5 Bandwidth

■ Giới thiệu bandwidth

■ Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

■ Ưu điểm

■ Nhược điểm



Bài toán phân cụm dữ liệu là gì?

- Là bài toán rất quan trọng trong khai phá dữ liệu.



Bài toán phân cụm dữ liệu là gì?

- Là bài toán rất quan trọng trong khai phá dữ liệu.
- Thuộc lớp các phương pháp học không giám sát.



Bài toán phân cụm dữ liệu là gì?

- Là bài toán rất quan trọng trong khai phá dữ liệu.
- Thuộc lớp các phương pháp học không giám sát.
- Về bản chất, có thể hiểu phân cụm là đưa đối tượng về các cụm của nó.



Một số thuật toán phân cụm dữ liệu

Example

K-Means, Mean Shift.



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu

- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực

- Kernel

- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

- Tổng quan

- Các bước Mean Shift

- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth

- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm

- Nhược điểm



Ứng dụng của bài toán phân cụm

- Phân mảng hình ảnh.



Ứng dụng của bài toán phân cụm

- Phân mảng hình ảnh.
- Sinh học: Phân nhóm động, thực vật.



Ứng dụng của bài toán phân cụm

- Phân mảng hình ảnh.
- Sinh học: Phân nhóm động, thực vật.
- Marketing: Xác định các nhóm khách hàng tiềm năng, phân loại và dự đoán hành vi khách hàng.

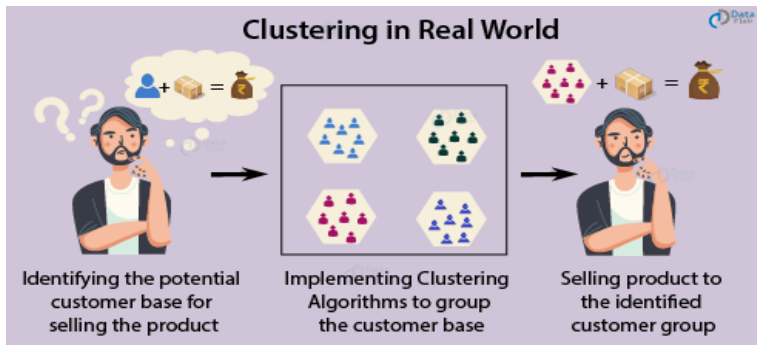


Ứng dụng của bài toán phân cụm

- Phân mảng hình ảnh.
- Sinh học: Phân nhóm động, thực vật.
- Marketing: Xác định các nhóm khách hàng tiềm năng, phân loại và dự đoán hành vi khách hàng.
- Bảo hiểm, tài chính: Phân nhóm các đối tượng, dự đoán xu hướng, phát hiện gian lận tài chính.



Ứng dụng của bài toán phân cụm



Hình: Ứng dụng của bài toán phân cụm trong Marketing

Một số quy ước

- Với x là một vector thuộc không gian d chiều ($x \in \mathbb{R}^d$) thì:

$$\|x\| = \sqrt{\sum_{i=1}^d x_i^2}.$$



Một số quy ước

- Với x là một vector thuộc không gian d chiều ($x \in \mathbb{R}^d$) thì:
$$\|x\| = \sqrt{\sum_{i=1}^d x_i^2}.$$
- S là tập hợp giá trị của miền dữ liệu ($S \subset \mathbb{R}^d$ với d là số chiều của dữ liệu).



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

- Tổng quan
- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Động lực

Giả định rằng có một tập dữ liệu rời rạc hữu hạn được sinh ra theo một phân bố nào đó. KDE là một phương pháp không có tham số (non-parametric) được sử dụng để ước lượng hàm phân bố xác suất (probability density function) đã được dùng để sinh ra tập dữ liệu này.



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Kernel là gì?

Hàm $K(\vec{x}) : S \rightarrow \mathbb{R}$ được gọi là một kernel khi và chỉ khi tồn tại một hàm $k : [0, +\infty) \rightarrow \mathbb{R}$ thoả mãn:



Kernel là gì?

Hàm $K(\vec{x}) : S \rightarrow \mathbb{R}$ được gọi là một kernel khi và chỉ khi tồn tại một hàm $k : [0, +\infty) \rightarrow \mathbb{R}$ thoả mãn:

- $K(\vec{x}) = k(\|\vec{x}\|^2)$



Kernel là gì?

Hàm $K(\vec{x}) : S \rightarrow \mathbb{R}$ được gọi là một kernel khi và chỉ khi tồn tại một hàm $k : [0, +\infty) \rightarrow \mathbb{R}$ thoả mãn:

- $K(\vec{x}) = k(\|\vec{x}\|^2)$
- $k(x)$ không âm



Kernel là gì?

Hàm $K(\vec{x}) : S \rightarrow \mathbb{R}$ được gọi là một kernel khi và chỉ khi tồn tại một hàm $k : [0, +\infty) \rightarrow \mathbb{R}$ thoả mãn:

- $K(\vec{x}) = k(\|\vec{x}\|^2)$
- $k(x)$ không âm
- $k(x)$ không tăng, tức là $\forall a < b : k(a) \geq k(b)$



Kernel là gì?

Hàm $K(\vec{x}) : S \rightarrow \mathbb{R}$ được gọi là một kernel khi và chỉ khi tồn tại một hàm $k : [0, +\infty) \rightarrow \mathbb{R}$ thoả mãn:

- $K(\vec{x}) = k(\|\vec{x}\|^2)$
- $k(x)$ không âm
- $k(x)$ không tăng, tức là $\forall a < b : k(a) \geq k(b)$
- $k(x)$ khả vi với mọi $x \in [0, +\infty)$



Kernel là gì?

Hàm $K(\vec{x}) : S \rightarrow \mathbb{R}$ được gọi là một kernel khi và chỉ khi tồn tại một hàm $k : [0, +\infty) \rightarrow \mathbb{R}$ thoả mãn:

- $K(\vec{x}) = k(\|\vec{x}\|^2)$
- $k(x)$ không âm
- $k(x)$ không tăng, tức là $\forall a < b : k(a) \geq k(b)$
- $k(x)$ khả vi với mọi $x \in [0, +\infty)$
- $\int_0^{+\infty} k(x)dx < +\infty$ (thông thường để chuẩn hoá, tích phân này có giá trị bằng 1)



Một số kernel thường gặp

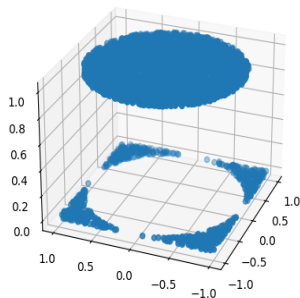
Flat, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, quadratic.



Flat kernel

Flat kernel

$$k(x) = \begin{cases} 1 & \text{if } x \leq \lambda \\ 0 & \text{if } x > \lambda \end{cases}$$



Hình: Unit Flat Kernel K cho dữ liệu 2 chiều

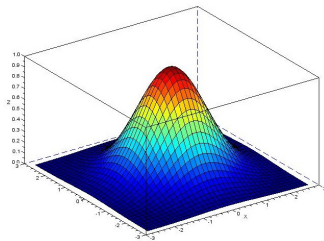


Gaussian Kernel

Gaussian kernel

$$k(x) = e^{-\frac{x^2}{2\sigma^2}},$$

Trong đó tham số độ lệch chuẩn σ được coi như là tham số bandwidth của thuật toán Mean Shift (sẽ được giải thích rõ hơn sau).



Hình: Unit Gaussian Kernel cho dữ liệu 2 chiều



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



KDE là gì?

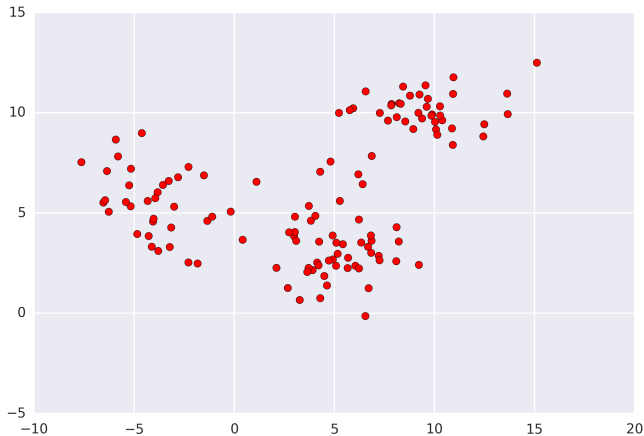
Gọi (x_1, x_2, \dots, x_n) là n điểm dữ liệu được lấy mẫu độc lập với nhau từ một phân bố nào đó có hàm mật độ f . Chúng ta cần ước lượng hình dáng hàm f này. Hàm f có thể được ước lượng bởi hàm số \hat{f} , gọi là KDE và tính bởi công thức :

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right). \quad (1)$$

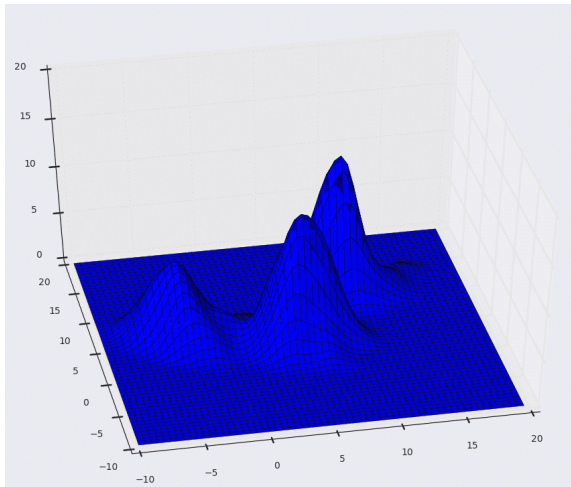
Trong đó K là hàm Kernel còn h là tham số bán kính (bandwidth), ảnh hưởng đến độ "trơn" trong ước lượng của phân phối.



KDE là gì?



KDE là gì?



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Lịch sử

Thuật toán Mean Shift Clustering ra đời vào năm 1975 và thường được ứng dụng vào các bài toán: phân cụm dữ liệu, phân mảng hình ảnh, dò theo đối tượng hình ảnh.



Tổng quan thuật toán Mean Shift

- Sử dụng phương pháp Multiple Restart Gradient Ascent để đưa các điểm dữ liệu về các cực đại địa phương của hàm KDE. Khi thuật toán dừng, mỗi điểm được gán cho một cụm.



Tổng quan thuật toán Mean Shift

- Sử dụng phương pháp Multiple Restart Gradient Ascent để đưa các điểm dữ liệu về các cực đại địa phương của hàm KDE. Khi thuật toán dừng, mỗi điểm được gán cho một cụm.
- Mean Shift không yêu cầu chỉ định trước số lượng cụm.



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

■ Các bước Mean Shift

■ Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Giá trị mean

Cho $Q \subset S$ là một tập hữu hạn dữ liệu xung quanh điểm x và hàm Kernel K . Giá trị trung bình có trọng số với Kernel K tại điểm x được định nghĩa như sau:

$$m(x) = \frac{\sum_{x_i \in Q} K(x_i - x)x_i}{\sum_{x_i \in Q} K(x_i - x)}. \quad (2)$$

Khi đó, $m(x) - x$, được gọi là mean shift vector v , sẽ chỉ mỗi điểm về hướng của cụm có mật độ cao. Thuật toán Mean Shift sẽ gán $x \leftarrow m(x)$ và lặp đi lặp lại cho đến khi hội tụ.



Thuật toán mean shift clustering gồm 2 quá trình



Thuật toán mean shift clustering gồm 2 quá trình

1 Mean shift

1 Tính toán Mean Shift vector $v(x_i^t)$



Thuật toán mean shift clustering gồm 2 quá trình

1 Mean shift

- 1 Tính toán Mean Shift vector $v(x_i^t)$
- 2 Dịch chuyển điểm x_i^t tới vị trí mới $x_i^{t+1} = x_i^t + v(x_i^t)$



Thuật toán mean shift clustering gồm 2 quá trình

1 Mean shift

- 1 Tính toán Mean Shift vector $v(x_i^t)$
- 2 Dịch chuyển điểm x_i^t tới vị trí mới $x_i^{t+1} = x_i^t + v(x_i^t)$
- 3 Lặp lại 2 bước trên cho tới khi x_i^{t+1} gần như hội tụ với tâm của một trong những cụm có mật độ dày đặc hoặc đạt đến điều kiện dừng.



Thuật toán mean shift clustering gồm 2 quá trình

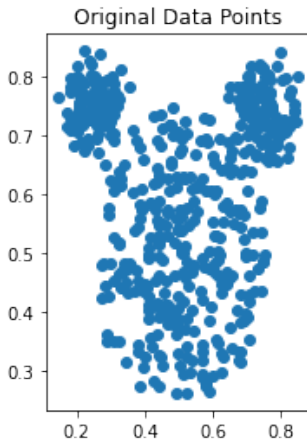
1 Mean shift

- 1 Tính toán Mean Shift vector $v(x_i^t)$
- 2 Dịch chuyển điểm x_i^t tới vị trí mới $x_i^{t+1} = x_i^t + v(x_i^t)$
- 3 Lặp lại 2 bước trên cho tới khi x_i^{t+1} gần như hội tụ với tâm của một trong những cụm có mật độ dày đặc hoặc đạt đến điều kiện dừng.

2 Clustering: Xếp các điểm dữ liệu vào các cụm thích hợp.



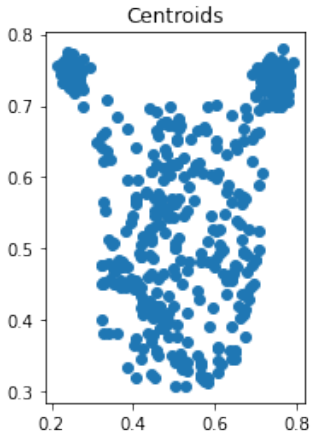
Demo Mean Shift



Hình: Bộ dữ liệu mouse.csv



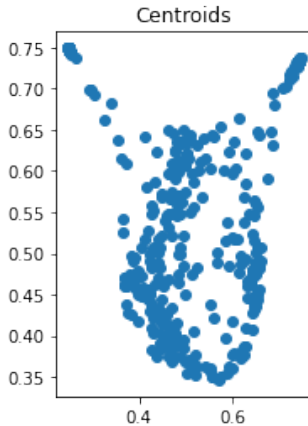
Demo Mean Shift



Hình: Iteration = 1



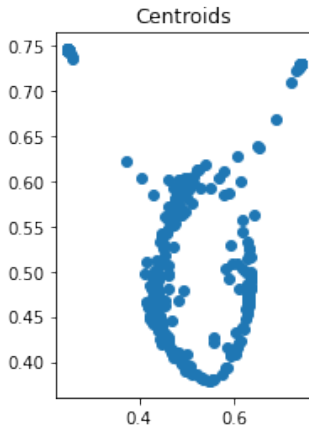
Demo Mean Shift



Hình: Iteration = 2



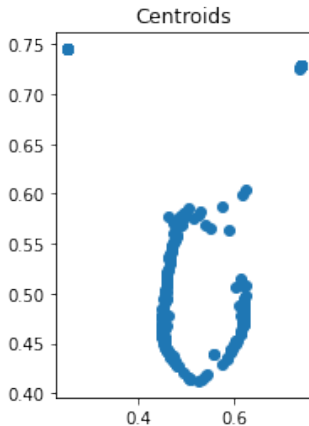
Demo Mean Shift



Hình: Iteration = 3



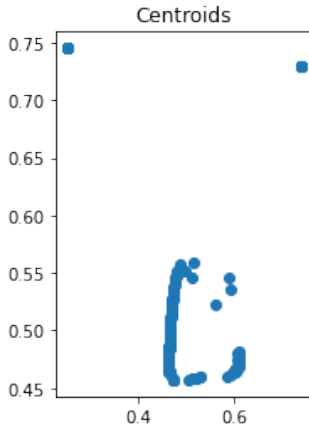
Demo Mean Shift



Hình: Iteration = 4



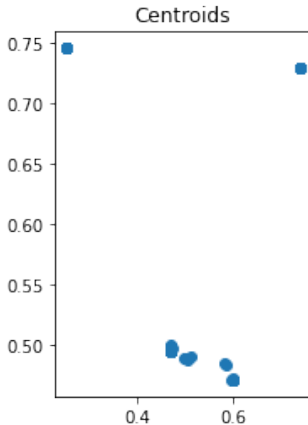
Demo Mean Shift



Hình: Iteration = 5



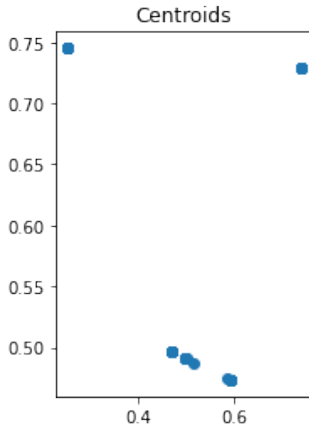
Demo Mean Shift



Hình: Iteration = 6



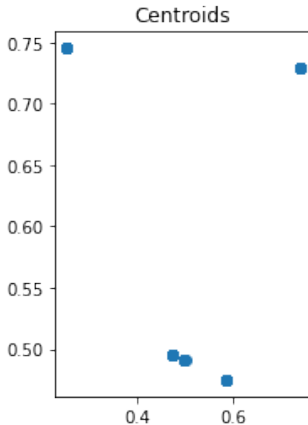
Demo Mean Shift



Hình: Iteration = 7



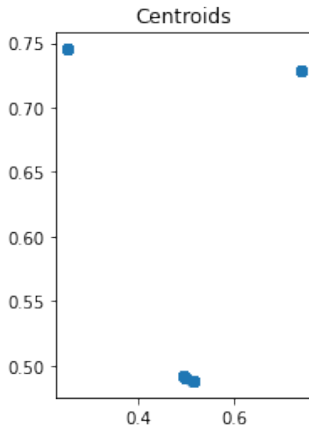
Demo Mean Shift



Hình: Iteration = 8



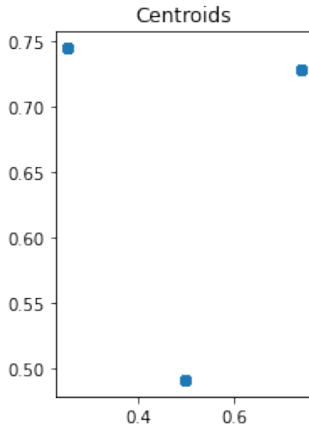
Demo Mean Shift



Hình: Iteration = 9



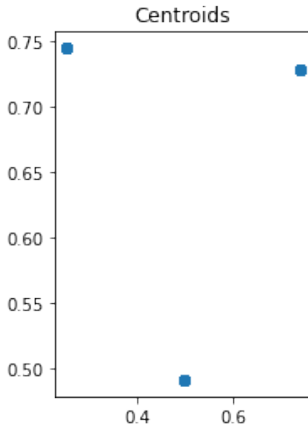
Demo Mean Shift



Hình: Iteration = 10



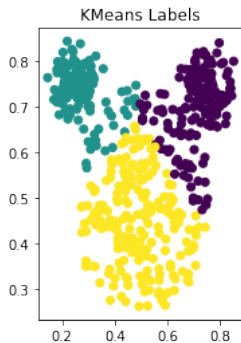
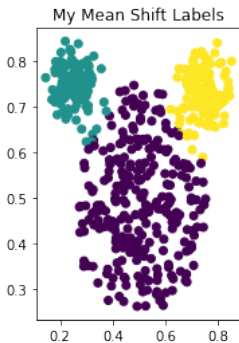
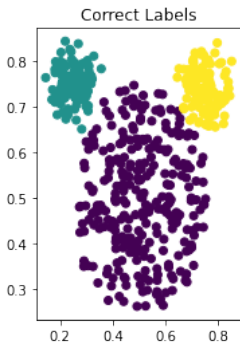
Demo Mean Shift



Hình: Iteration = 11



Demo Mean Shift



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

■ Các bước Mean Shift

■ Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Về bản chất

Quá trình Mean Shift là quá trình tìm ra các cực trị địa phương của hàm KDE. Hàm KDE được định nghĩa như sau:

$$f_K(x, Q) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right). \quad (3)$$

Trong đó, h là tham số bandwidth chỉ bán kính của Kernel và hàm $K(x)$ được định nghĩa như sau:

$$K(x) = ck(||x||^2) \quad (4)$$

Với c là hằng số chuẩn hóa và k là hàm kernel.



Về bản chất

Thay (4) vào (3), hàm dự đoán mật độ (3) sẽ trở thành:

$$f_K(x, Q) = \frac{c}{n} \sum_{i=1}^n k \left(\left\| \frac{x - x_i}{h} \right\|^2 \right). \quad (5)$$



Về bản chất

Thay (4) vào (3), hàm dự đoán mật độ (3) sẽ trở thành:

$$f_K(x, Q) = \frac{c}{n} \sum_{i=1}^n k \left(\left\| \frac{x - x_i}{h} \right\|^2 \right). \quad (5)$$

Hàm dự đoán mật độ gradient có được bằng cách lấy gradient của hàm (5):

$$\nabla f_K(x, Q) = \frac{2c}{nh^2} \sum_{i=1}^n (x - x_i) k' \left(\left\| \frac{x - x_i}{h} \right\|^2 \right). \quad (6)$$



Về bản chất

Đặt

$$g(x) = -k'(x) \quad (7)$$



Về bản chất

Đặt

$$g(x) = -k'(x) \quad (7)$$

Thay (7) vào (6), ta được:

$$\begin{aligned} \nabla f_K(x, Q) &= \frac{2c}{nh^2} \sum_{i=1}^n (x_i - x) g\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \\ &= \frac{2c}{nh^2} \left[\sum_{i=1}^n g\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \right] \left[\frac{\sum_{i=1}^n g\left(\left\|\frac{x - x_i}{h}\right\|^2\right) x_i}{\sum_{i=1}^n g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \right]. \end{aligned}$$



Về bản chất

Suy ra,

$$\frac{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)x_i}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} = x + \frac{nh^2}{2c \sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} \nabla f_K(x, S). \quad (8)$$



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu

- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực

- Kernel

- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

- Tổng quan

- Các bước Mean Shift

- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth

- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm

- Nhược điểm



Bandwidth là gì

- Một tham số tự do có ảnh hưởng mạnh mẽ đến kết quả ước tính.



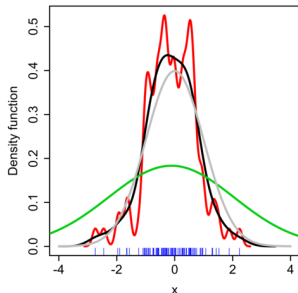
Bandwidth là gì

- Một tham số tự do có ảnh hưởng mạnh mẽ đến kết quả ước tính.
- Bandwidth khác nhau sẽ cho kết quả phân cụm khác nhau.



Minh họa

Đường cong màu xám là hàm mật độ thực (có giá trị trung bình là 0 và phương sai 1). Với $h = 0.05$, $h = 2$, $h = 0.337$.



Hình: KDE với các bandwidth khác nhau của một phép thử ngẫu nhiên 100 điểm từ phân phối chuẩn.

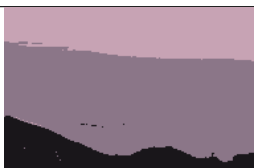
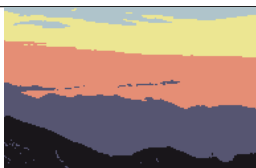


Demo lựa chọn bandwidth



Hình: Ảnh gốc

Demo lựa chọn bandwidth



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

- Tổng quan
- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Đánh giá độ tốt của bandwidth

Cách phổ biến nhất là sử dụng hàm mất mát L_2 (MISE - mean integrated squared error).

$$MISE(h) = E \left[\int (\hat{f}_h(x) - f(x))^2 dx \right]. \quad (9)$$

Trong đó \hat{f} là hàm KDE mà ta dự đoán còn f là hàm mật độ xác suất thật sự. Bằng chứng minh toán học, giá trị h làm cho hàm MISE đạt giá trị nhỏ nhất là:

$$h = \frac{\int K(x)^2 dx}{\left[\int x^2 K(x) dx \right] \left[\int f''(x)^2 dx \right] n^{1/5}}. \quad (10)$$



Đánh giá độ tốt của bandwidth

Lưu ý: công thức ở trên không thể áp dụng được trong thực tế do chưa biết được f nên cũng không thể tính được f'' .



Đánh giá độ tốt của bandwidth

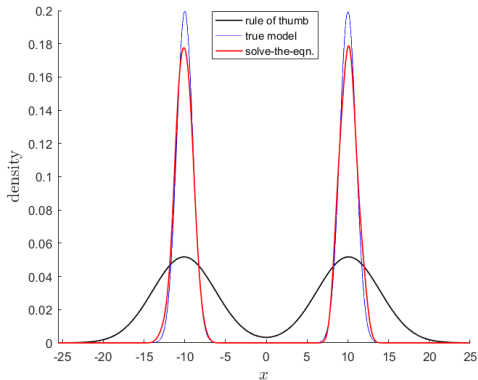
Lưu ý: công thức ở trên không thể áp dụng được trong thực tế do chưa biết được f nên cũng không thể tính được f'' .

Nếu giả định rằng dữ liệu được sinh bằng phân phối chuẩn, ta có thể chứng minh được hàm mất mát L_2 đạt giá trị cực tiểu tại:

$$h = \left(\frac{4\sigma^5}{3n} \right)^{\frac{1}{5}} \quad (11)$$



Đánh giá độ tốt của bandwidth



Hình: Một số sự lựa chọn bandwidth



Sự hội tụ

Thuật toán Mean Shift đã sử dụng Gradient Ascent để đưa mỗi điểm đến cực đại địa phương một cách độc lập \Rightarrow sự hội tụ của thuật toán Mean Shift là hệ quả của việc sử dụng Gradient Ascent đối với các điểm x riêng lẻ.



Tính dừng của thuật toán

- Đối với dữ liệu rời rạc, số bước hội tụ phụ thuộc vào kernel được sử dụng.

Tính dừng của thuật toán

- Đối với dữ liệu rời rạc, số bước hội tụ phụ thuộc vào kernel được sử dụng.
- Khi G là flat kernel, sự hội tụ đạt được trong một số bước hữu hạn.



Tính dừng của thuật toán

- Đối với dữ liệu rời rạc, số bước hội tụ phụ thuộc vào kernel được sử dụng.
- Khi G là flat kernel, sự hội tụ đạt được trong một số bước hữu hạn.
- Khi sử dụng một kernel khác (Gaussian Kernel), thì sự hội tụ sau hữu hạn bước là không chắc chắn. Giải pháp: đặt giới hạn dưới cho v .



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Ưu điểm

- Không dựa trên một số giả định như thuật toán K-means.



Ưu điểm

- Không dựa trên một số giả định như thuật toán K-means.
- Chỉ phụ thuộc một siêu tham số bandwidth.



Ưu điểm

- Không dựa trên một số giả định như thuật toán K-means.
- Chỉ phụ thuộc một siêu tham số bandwidth.
- Có thể xử lí các điểm dữ liệu ngoại lai (outliers) tốt hơn khá nhiều so với thuật toán K-means.



Contents

1 Bài toán phân cụm dữ liệu

- Giới thiệu
- Ứng dụng

2 Quy ước

3 Kernel Density Estimator (KDE)

- Động lực
- Kernel
- Hàm Kernel Density Estimator

4 Thuật toán mean shift clustering

■ Tổng quan

- Các bước Mean Shift
- Ý nghĩa toán học

5 Bandwidth

- Giới thiệu bandwidth
- Độ tốt của bandwidth

6 Sự hội tụ và tính dừng của thuật toán

7 Ưu, nhược điểm của thuật toán Mean Shift

- Ưu điểm
- Nhược điểm



Nhược điểm

- Độ phức tạp cao.



Nhược điểm

- Độ phức tạp cao.
- Cần phải tìm ra một giá trị bandwidth phù hợp.