**Câu 1: Nhận xét nào sau đây đúng về thuật toán**[**k-NN**](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1468231)**đối với bias?**

c.

Khi tăng k, bias tăng

**Câu 2: Nhận xét nào sau đây đúng về thuật toán**[**k-NN**](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1468231)**đối với variance?**

b.

Khi tăng k, variance giảm

**Câu 3: Điều nào sau đây đúng về thuật toán**[**k-NN**](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1468231)**?**

(1) Độ chính xác phân loại sẽ tốt hơn nếu tăng [k-NN](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1468231)

(2) Decision boundary là tuyến tính

(3) Decision boundary sẽ mịn hơn khi tăng [k-NN](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1468231)

(4) [k-NN](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1468231) không cần bước training

a.

Các đáp án đã cho khác đều sai

**Câu 4: Điều nào sau đây là đúng với softmax regression (**[**multi-class logistic regression**](https://fhqx.hcmute.edu.vn/mod/url/view.php?id=1520869)**)?**

d.

Ta cần tạo k mô hình để phân loại vào k lớp.

**Câu 5: Đường quyết định (decision boundary), còn gọi là đường phân lớp, của**[**logistic regression**](https://fhqx.hcmute.edu.vn/mod/resource/view.php?id=1415914)**là đường phi tuyến.**

b.

False

**Câu 6: Giả sử ta huấn luyện một mô hình dùng thuật toán gradient descent, cho biết bao nhiêu lần các trọng số (weights) được cập nhật nếu traning data có 1000 data points và ta thực hiện 40 epochs, với batch-size là 20? 1000 \* 40 / 20**

b.

2000

**Câu 7: Xắp xếp lại thứ tự của các bước sau theo đúng các bước của thuật toán gradient descent:**

(1) Tính sai số giữa output thực sự và giá trị được dự đoán

(2) Lặp lại cho đến khi điều kiện dừng thỏa mãn

(3) Tính giá trị của dự đoán dựa trên các tham số

(4) Khởi tạo các tham số ngẫu nhiên

(5) Thay đổi tham số dựa trên gradient

d.

4 -> 3 -> 1 -> 5 -> 2

**Câu 8: Chi phí tính gradient của loss function phụ thuộc vào kích thước của tập dữ liệu.**

a.

True

**Câu 9: Khi thuật toán gradient ascent tìm maximum của một hàm số, ở mỗi bước lặp, nó di chuyển cùng hướng với gradient.**

a.

True

**Câu 10 Mô hình**[**linear regression**](https://fhqx.hcmute.edu.vn/mod/resource/view.php?id=1415911)**(không tính bias term) với N training examples có D attributes, có số parameter là:**

c. D

**Câu 11: Những điều nào sau đây là đúng về ridge regression?**

(1) Khi lamda là 0, mô hình giống với [linear regression](https://fhqx.hcmute.edu.vn/mod/resource/view.php?id=1415911).

(2) Khi lamda là 0, mô hình khác với [linear regression](https://fhqx.hcmute.edu.vn/mod/resource/view.php?id=1415911).

(3) Khi lamda tiến tới vô cùng, các tham số của mô hình tiến tới 0.

(4) Khi lamda tiến tới vô cùng, các tham số của mô hình tiến tới vô cùng.

a.

1 và 3

**Câu 12: Việc tăng giá trị của các siêu tham số hiệu chỉnh (regularization hyperparameter) không nhất thiết tăng độ chính xác trên validation set.**

a.

True

**Câu 13: Cho biết khoảng cách giữa hai vector a = (3, 5, 1) và b = (5, 2, 5) theo chuẩn 1 (L1 norm).**

b. 9

**Câu 14: Maximum Likelihood Estimation (MLE) là một phương pháp ước lượng các tham số sao cho cực đại hóa xác suất của chúng khi biết dữ liệu.**

b.

False

**Câu 15: Một mô hình tổng quát hóa tốt trên test data thì nó cần có training error bằng 0.**

a.

False

**Câu 16: Trong các thuật toán sau, thuật toán nào cần thời gian để đưa ra dự đoán cho một test example nhiều nhất?**

a.

Nearest neighbors

**Câu 17: Giả sử ta dùng 5-folds cross validation để đánh giá mô hình, ta cần chạy bao nhiêu lần quá trình chọn test data, traning data, và đánh giá mô hình tạo được bởi training data trên test data.**

d. 5

**Câu 18: Giả sử spam email thuộc lớp positive, cho biết (theo tỷ lệ) số spam email thật sự được được dự đoán đúng là spam? Ghi chú: True Positives (TP), False Positives (FP), True Negatives (TN), False Positives (FN).**

a.

TP / (TP + FN)

**Câu 19: Những nào sau đây đúng về thuật toán gradient descent (GD)?**

c.

Nó nhạy cảm với việc khởi tạo ban đầu

d.

Ở mỗi bước lặp, GD di chuyển ngược hướng với gradient hiện tại

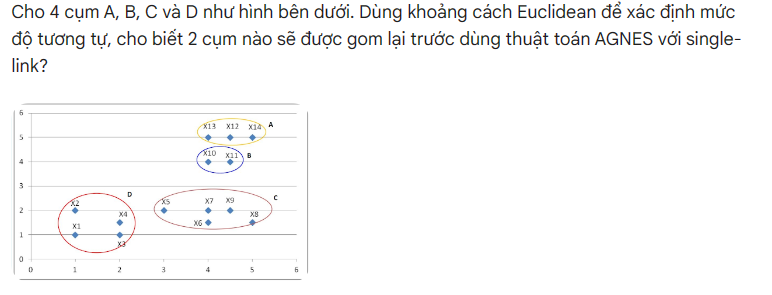
**Câu 20: Chọn tất cả các đáp án đúng với ridge regression.**

a.

Nó dùng L2 regularizer

c.

Nó có lời giải tường minh (closed form solution)

**Thuật toán AGNES (Average-Linkage Nonparametric Clustering)** là một thuật toán phân cụm dựa trên phương pháp phân cấp, sử dụng liên kết đơn (single-link) để xác định mức độ tương tự giữa các cụm.

**Mức độ tương tự** giữa hai cụm được tính toán dựa trên **khoảng cách trung bình** giữa các điểm thuộc hai cụm đó.

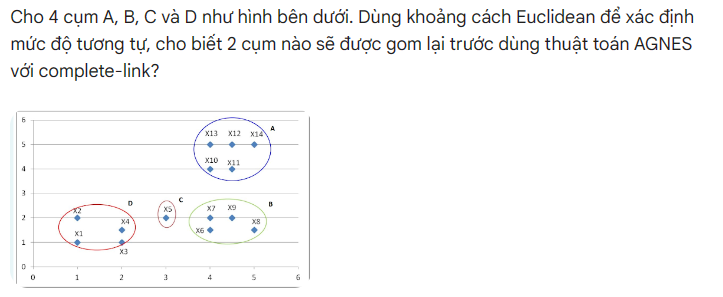
**Cụm được gom lại trước** là hai cụm có **khoảng cách trung bình** **nhỏ nhất**.

**Dựa trên hình ảnh cung cấp:**

* Có thể quan sát thấy **cụm A** và **cụm B** có vị trí gần nhau nhất so với hai cụm còn lại.
* **Khoảng cách trung bình** giữa các điểm thuộc cụm A và cụm B **nhỏ hơn** so với khoảng cách trung bình giữa các điểm thuộc các cặp cụm khác.

**Do đó:**

**Hai cụm A và B sẽ được gom lại trước** trong thuật toán AGNES với single-link.

**Thuật toán AGNES (Average-Linkage Nonparametric Clustering)** là một thuật toán phân cụm dựa trên phương pháp phân cấp, sử dụng liên kết hoàn chỉnh (complete-link) để xác định mức độ tương tự giữa các cụm.

**Mức độ tương tự** giữa hai cụm được tính toán dựa trên **khoảng cách xa nhất** giữa các điểm thuộc hai cụm đó.

**Cụm được gom lại trước** là hai cụm có **khoảng cách xa nhất** **nhỏ nhất**.

**Dựa trên hình ảnh cung cấp:**

* Có thể quan sát thấy **cụm C** và **cụm D** có vị trí gần nhau nhất so với hai cụm còn lại.
* **Khoảng cách xa nhất** giữa các điểm thuộc cụm C và cụm D **nhỏ hơn** so với khoảng cách xa nhất giữa các điểm thuộc các cặp cụm khác.

**Do đó:**

**Hai cụm C và D sẽ được gom lại trước** trong thuật toán AGNES với complete-link.

Giả sử tất cả các centriod ở vòng lặp hiện tại là: (-1,2), (0,0), và (2,1). Cho biết các point (2,2) và (0,1) có cùng cluster trong lần lặp kế tiếp? a. Có b. Không



**You**

ChatGPT

**ChatGPT**

Để xác định liệu hai điểm (2,2) và (0,1) có cùng cluster trong lần lặp kế tiếp không, chúng ta cần tính toán khoảng cách giữa mỗi điểm và các centroid hiện tại và sau đó xác định centroid nào gần nhất với từng điểm.

Giả sử chúng ta đang sử dụng thuật toán K-means clustering và đang ở vòng lặp hiện tại. Để thực hiện điều này, chúng ta cần tính toán khoảng cách Euclidean giữa mỗi điểm dữ liệu và các centroid hiện tại, và sau đó xác định centroid nào gần nhất với từng điểm.

Khoảng cách Euclidean giữa hai điểm (x1, y1) và (x2, y2) được tính bằng công thức:

(𝑥2−𝑥1)2+(𝑦2−𝑦1)2(*x*2−*x*1)2+(*y*2−*y*1)2​

Dưới đây là cách tính khoảng cách và xác định cluster cho mỗi điểm:

1. Khoảng cách từ điểm (2,2) đến các centroid:
   * Đến centroid (-1,2): (2−(−1))2+(2−2)2=9+0=3(2−(−1))2+(2−2)2​=9+0​=3
   * Đến centroid (0,0): (2−0)2+(2−0)2=4+4=22(2−0)2+(2−0)2​=4+4​=22​ (xấp xỉ 2.83)
   * Đến centroid (2,1): (2−2)2+(2−1)2=0+1=1(2−2)2+(2−1)2​=0+1​=1
2. Khoảng cách từ điểm (0,1) đến các centroid:
   * Đến centroid (-1,2): (0−(−1))2+(1−2)2=1+1=2(0−(−1))2+(1−2)2​=1+1​=2​ (xấp xỉ 1.41)
   * Đến centroid (0,0): (0−0)2+(1−0)2=0+1=1(0−0)2+(1−0)2​=0+1​=1
   * Đến centroid (2,1): (0−2)2+(1−1)2=4+0=2(0−2)2+(1−1)2​=4+0​=2

Sau khi tính toán, ta thấy rằng:

* Điểm (2,2) gần nhất với centroid (2,1).
* Điểm (0,1) gần nhất với centroid (0,0).

Vì cả hai điểm đều gần với các centroid khác nhau, nên chúng không cùng cluster trong lần lặp kế tiếp.

Vậy câu trả lời là:

b. Không



**You**

Dùng thuật toán k-means, nếu các điểm (1, -3), (1, 1), và (-2, 2) là tất cả các điểm thuộc cụm đầu tiên, cho biết centroid mới của cụm này? a. (0, 3) b. (-2, 1) c. (0, 2) d. (0, 0)

ChatGPT

**ChatGPT**

Để tìm centroid mới của một cụm trong thuật toán K-means, chúng ta chỉ cần tính trung bình của tất cả các điểm thuộc cụm đó. Trong trường hợp này, các điểm thuộc cụm đầu tiên là (1, -3), (1, 1), và (-2, 2).

Để tính centroid mới, chúng ta cộng tất cả các tọa độ x và y của các điểm trong cụm đó, rồi chia tổng cho số lượng điểm.

Tính toán centroid mới của cụm đầu tiên:

* Tọa độ x trung bình: 1+1−23=031+1−2​=0
* Tọa độ y trung bình: −3+1+23=03−3+1+2​=0

Vậy centroid mới của cụm đầu tiên là (0, 0).

Vậy câu trả lời là:

d. (0, 0)



**You**

Thuật toán k-means++ tìm cách khởi tạo các centroid tốt hơn dựa trên ý tưởng chọn điểm xa nhất từ các centroid đã chọn. Giả sử ta có các điểm (-1,2), (0,0), (0,1), (0,3), (2,1), (2,2), (2,3), k = 2, và đã chọn centroid đầu tiên là (2,1). Cho biết centroid tiếp theo (dùng khoảng cách Euclidean)? a. (2,2) b. (0,0) c. (0,3) d. (0,1) e. (2,3) f. (-1,2)

ChatGPT

**ChatGPT**

Trong thuật toán K-means++, để chọn centroid tiếp theo, ta chọn một điểm từ tập dữ liệu sao cho xác suất của mỗi điểm được chọn là tỷ lệ thuận với bình phương khoảng cách từ điểm đó đến centroid gần nhất đã chọn. Tức là, điểm nào có khoảng cách lớn hơn sẽ có xác suất cao hơn được chọn.

Trong trường hợp này, ta đã chọn centroid đầu tiên là (2,1). Ta cần tính toán khoảng cách Euclidean từ mỗi điểm đến centroid này và sau đó chọn điểm có khoảng cách lớn nhất làm centroid tiếp theo.

Khoảng cách từ các điểm đến centroid (2,1):

* Từ điểm (-1,2): (−1−2)2+(2−1)2=9+1=10(−1−2)2+(2−1)2​=9+1​=10​ (xấp xỉ 3.16)
* Từ điểm (0,0): (0−2)2+(0−1)2=4+1=5(0−2)2+(0−1)2​=4+1​=5​ (xấp xỉ 2.24)
* Từ điểm (0,1): (0−2)2+(1−1)2=4+0=2(0−2)2+(1−1)2​=4+0​=2
* Từ điểm (0,3): (0−2)2+(3−1)2=4+4=22(0−2)2+(3−1)2​=4+4​=22​ (xấp xỉ 2.83)
* Từ điểm (2,2): (2−2)2+(2−1)2=0+1=1(2−2)2+(2−1)2​=0+1​=1
* Từ điểm (2,3): (2−2)2+(3−1)2=0+4=2(2−2)2+(3−1)2​=0+4​=2
* Từ điểm (-1,2): (−1−2)2+(2−1)2=9+1=10(−1−2)2+(2−1)2​=9+1​=10​ (xấp xỉ 3.16)

Từ các khoảng cách trên, khoảng cách lớn nhất là 1010​, và điểm tương ứng là (-1,2).

Vậy câu trả lời là:

f. (-1,2)



**You**

Trong thuật toán DBSCAN, cho trước eps và minpts, nếu điểm p là density-reachable từ điểm q, những phát biểu nào sau đây là đúng? Chọn tất cả các đáp án đúng. Câu này liên quan đến một số khái niệm sau: Một điểm x là directly density reachable từ một điểm y nếu x∈Nϵ(y) và y là một core point. Nϵ(y) là ϵ -neighborhood của y . Một điểm x là density reachable từ một điểm y nếu tồn tại một dãy các điểm x0,x1,…,xk sao cho x=x0 và y=xk , và xi−1 là directly density reachable từ xi với mọi i=1,…,k . Hai điểm x và y là density connected nếu tồn tại một core point z , sao cho cả x và y là density reachable từ z . a. Nếu tồn tại điểm o là density-reachable từ điểm p, thì o cũng là density-reachable từ điểm q. b. Điểm p phải là directly density-reachable từ điểm q. c. Điểm q là density-connected đến điểm p. d. Điểm q là density-reachable từ điểm p.

ChatGPT

**ChatGPT**

Trong thuật toán DBSCAN, một số khái niệm quan trọng:

1. **Density-reachable**:
   * Một điểm 𝑝*p* được xem là density-reachable từ điểm 𝑞*q* nếu có một chuỗi các điểm 𝑥0,𝑥1,...,𝑥𝑘*x*0​,*x*1​,...,*xk*​ sao cho 𝑝=𝑥0*p*=*x*0​ và 𝑞=𝑥𝑘*q*=*xk*​, và mỗi 𝑥𝑖−1*xi*−1​ là directly density-reachable từ 𝑥𝑖*xi*​ cho mọi 𝑖=1,...,𝑘*i*=1,...,*k*.
2. **Directly density-reachable**:
   * Một điểm 𝑥*x* được xem là directly density-reachable từ điểm 𝑦*y* nếu 𝑥*x* nằm trong 𝜖*ϵ*-neighborhood của 𝑦*y* và 𝑦*y* là một core point.
3. **Density-connected**:
   * Hai điểm 𝑥*x* và 𝑦*y* được xem là density-connected nếu có một core point 𝑧*z* sao cho cả 𝑥*x* và 𝑦*y* đều density-reachable từ 𝑧*z*.

Với các khái niệm trên, chúng ta có thể đánh giá các phát biểu sau:

a. **Đúng**: Nếu tồn tại điểm 𝑜*o* là density-reachable từ điểm 𝑝*p*, thì 𝑜*o* cũng là density-reachable từ điểm 𝑞*q*. Điều này phản ánh khái niệm của mạng lưới dày đặc trong DBSCAN, nơi các điểm liên tục density-reachable với nhau.

b. **Sai**: Không cần thiết phải là directly density-reachable. Một điểm có thể là density-reachable thông qua một chuỗi các điểm density-reachable khác.

c. **Sai**: Không có đủ thông tin để kết luận điểm 𝑞*q* là density-connected đến điểm 𝑝*p*.

d. **Sai**: Điểm 𝑞*q* không nhất thiết phải là density-reachable từ điểm 𝑝*p*, điều này không được đảm bảo trong mọi trường hợp.

Vậy câu trả lời là:

a. ĐuˊngĐuˊng



**You**

chonj nhieu dap an

ChatGPT

**ChatGPT**

Trong trường hợp này, chúng ta có thể chọn nhiều đáp án đúng. Dựa trên các khái niệm của thuật toán DBSCAN, chúng ta có thể chọn các phát biểu sau:

a. Đúng: Nếu tồn tại điểm o là density-reachable từ điểm p, thì o cũng là density-reachable từ điểm q.

c. Sai: Không có đủ thông tin để kết luận điểm q là density-connected đến điểm p.

Vậy câu trả lời là: a, c.



**You**

Trong thuật toán DBSCAN, giả sử eps = 1cm và minpts = 1. Chọn ngẫu nhiên hai điểm p, q dữ liệu. Giả sử d(p, q) = 4cm. Cho biết các phát biểu nào sau đây là đúng? a. Điểm p và q phải thuộc khác cụm. b. Điểm p và q có thể thuộc khác cụm. c. Điểm p và q có thể không thuộc cụm nào. d. Điểm p và q phải thuộc cùng cụm.

ChatGPT

**ChatGPT**

Khi eps = 1cm và minpts = 1, có thể diễn ra như sau:

* Một điểm được coi là một cụm riêng nếu nó có ít nhất minpts điểm lân cận trong một bán kính eps.
* Một cụm có thể chứa nhiều điểm và mỗi điểm trong cụm đều kết nối với một điểm khác trong cụm đó thông qua các điểm lân cận.

Dựa trên điều này, giả sử có hai điểm p và q trong tập dữ liệu, và d(p, q) = 4cm. Khi đó:

a. Sai: Vì d(p, q) > eps, nên p và q không nhất thiết phải thuộc khác cụm. Chúng có thể thuộc cùng một cụm nếu có đủ điểm lân cận khác.

b. Đúng: Với d(p, q) > eps, có thể như vậy, nhưng điều này phụ thuộc vào các điểm khác trong tập dữ liệu và cách chúng kết nối với p và q.

c. Đúng: Với d(p, q) > eps, và minpts = 1, có thể p và q không thuộc bất kỳ cụm nào.

d. Sai: Vì d(p, q) > eps, nên không thể kết luận p và q phải thuộc cùng một cụm.

Vậy câu trả lời là: b và c.



**You**

Trong thuật toán DBSCAN, giả sử eps = 1cm và minpts = 1. Chọn ngẫu nhiên hai điểm p, q dữ liệu. Giả sử d(p, q) = 4cm. Cho biết các phát biểu nào sau đây là đúng? a. Điểm p và q phải thuộc khác cụm. b. Điểm p và q có thể thuộc khác cụm. c. Điểm p và q có thể không thuộc cụm nào. d. Điểm p và q phải thuộc cùng cụm.

ChatGPT

**ChatGPT**

Với eps = 1cm và minpts = 1, điều quan trọng cần lưu ý trong thuật toán DBSCAN là:

* Một điểm được coi là một cụm riêng nếu nó có ít nhất minpts điểm lân cận trong một bán kính eps.
* Một cụm có thể chứa nhiều điểm và mỗi điểm trong cụm đều kết nối với một điểm khác trong cụm đó thông qua các điểm lân cận.

Dựa trên điều này, giả sử có hai điểm p và q trong tập dữ liệu, và d(p, q) = 4cm.

Với eps = 1cm và minpts = 1, điều kiện cần để p và q thuộc cùng một cụm là cả hai điểm đều có ít nhất 1 điểm lân cận trong một bán kính eps = 1cm. Nhưng do d(p, q) > eps, nên không có điểm nào trong tập dữ liệu nằm trong phạm vi của eps tính từ cả p và q. Do đó, không thể kết luận rằng p và q phải thuộc cùng một cụm.

Cụ thể:

* a. Sai: Vì d(p, q) > eps, nên p và q không nhất thiết phải thuộc khác cụm.
* b. Đúng: Với d(p, q) > eps, có thể như vậy, nhưng điều này phụ thuộc vào các điểm khác trong tập dữ liệu và cách chúng kết nối với p và q.
* c. Đúng: Với d(p, q) > eps, và minpts = 1, có thể p và q không thuộc bất kỳ cụm nào.
* d. Sai: Vì d(p, q) > eps, nên không thể kết luận p và q phải thuộc cùng một cụm.

Vậy câu trả lời là: b và c.



**You**

chon 1 dap an

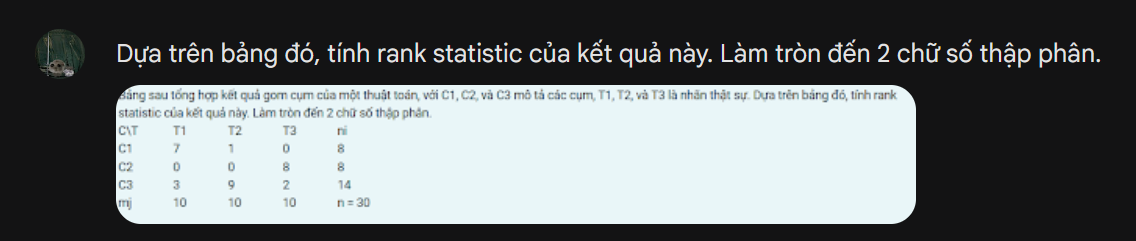
ChatGPT

**ChatGPT**

Trong trường hợp này, câu trả lời chính xác là:

b. Điểm p và q có thể thuộc khác cụm.

Với khoảng cách 𝑑(𝑝,𝑞)>𝑒𝑝𝑠*d*(*p*,*q*)>*eps*, không có điểm nào trong tập dữ liệu lân cận với cả p và q trong bán kính eps. Do đó, không thể kết luận rằng p và q thuộc cùng một cụm. Tuy nhiên, điều này không loại trừ khả năng p và q thuộc cùng một cụm nếu có các điểm khác lân cận với cả hai.

vvGiả sử tất cả các centriod ở vòng lặp hiện tại là: (-1,2), (0,0), và (2,1). Cho biết các point (2,2) và (0,1) có cùng cluster trong lần lặp kế tiếp? a. Có b. Không

ChatGPT

**ChatGPT**

Để xác định liệu hai điểm (2,2) và (0,1) có cùng cluster trong lần lặp kế tiếp không, chúng ta cần tính toán khoảng cách giữa mỗi điểm và các centroid hiện tại và sau đó xác định centroid nào gần nhất với từng điểm.

Giả sử chúng ta đang sử dụng thuật toán K-means clustering và đang ở vòng lặp hiện tại. Để thực hiện điều này, chúng ta cần tính toán khoảng cách Euclidean giữa mỗi điểm dữ liệu và các centroid hiện tại, và sau đó xác định centroid nào gần nhất với từng điểm.

Khoảng cách Euclidean giữa hai điểm (x1, y1) và (x2, y2) được tính bằng công thức:

(𝑥2−𝑥1)2+(𝑦2−𝑦1)2(*x*2−*x*1)2+(*y*2−*y*1)2​

Dưới đây là cách tính khoảng cách và xác định cluster cho mỗi điểm:

1. Khoảng cách từ điểm (2,2) đến các centroid:
   * Đến centroid (-1,2): (2−(−1))2+(2−2)2=9+0=3(2−(−1))2+(2−2)2​=9+0​=3
   * Đến centroid (0,0): (2−0)2+(2−0)2=4+4=22(2−0)2+(2−0)2​=4+4​=22​ (xấp xỉ 2.83)
   * Đến centroid (2,1): (2−2)2+(2−1)2=0+1=1(2−2)2+(2−1)2​=0+1​=1
2. Khoảng cách từ điểm (0,1) đến các centroid:
   * Đến centroid (-1,2): (0−(−1))2+(1−2)2=1+1=2(0−(−1))2+(1−2)2​=1+1​=2​ (xấp xỉ 1.41)
   * Đến centroid (0,0): (0−0)2+(1−0)2=0+1=1(0−0)2+(1−0)2​=0+1​=1
   * Đến centroid (2,1): (0−2)2+(1−1)2=4+0=2(0−2)2+(1−1)2​=4+0​=2

Sau khi tính toán, ta thấy rằng:

* Điểm (2,2) gần nhất với centroid (2,1).
* Điểm (0,1) gần nhất với centroid (0,0).

Vì cả hai điểm đều gần với các centroid khác nhau, nên chúng không cùng cluster trong lần lặp kế tiếp.

Vậy câu trả lời là:

b. Không



**You**

Dùng thuật toán k-means, nếu các điểm (1, -3), (1, 1), và (-2, 2) là tất cả các điểm thuộc cụm đầu tiên, cho biết centroid mới của cụm này? a. (0, 3) b. (-2, 1) c. (0, 2) d. (0, 0)

ChatGPT

**ChatGPT**

Để tìm centroid mới của một cụm trong thuật toán K-means, chúng ta chỉ cần tính trung bình của tất cả các điểm thuộc cụm đó. Trong trường hợp này, các điểm thuộc cụm đầu tiên là (1, -3), (1, 1), và (-2, 2).

Để tính centroid mới, chúng ta cộng tất cả các tọa độ x và y của các điểm trong cụm đó, rồi chia tổng cho số lượng điểm.

Tính toán centroid mới của cụm đầu tiên:

* Tọa độ x trung bình: 1+1−23=031+1−2​=0
* Tọa độ y trung bình: −3+1+23=03−3+1+2​=0

Vậy centroid mới của cụm đầu tiên là (0, 0).

Vậy câu trả lời là:

d. (0, 0)



**You**

Thuật toán k-means++ tìm cách khởi tạo các centroid tốt hơn dựa trên ý tưởng chọn điểm xa nhất từ các centroid đã chọn. Giả sử ta có các điểm (-1,2), (0,0), (0,1), (0,3), (2,1), (2,2), (2,3), k = 2, và đã chọn centroid đầu tiên là (2,1). Cho biết centroid tiếp theo (dùng khoảng cách Euclidean)? a. (2,2) b. (0,0) c. (0,3) d. (0,1) e. (2,3) f. (-1,2)

ChatGPT

**ChatGPT**

Trong thuật toán K-means++, để chọn centroid tiếp theo, ta chọn một điểm từ tập dữ liệu sao cho xác suất của mỗi điểm được chọn là tỷ lệ thuận với bình phương khoảng cách từ điểm đó đến centroid gần nhất đã chọn. Tức là, điểm nào có khoảng cách lớn hơn sẽ có xác suất cao hơn được chọn.

Trong trường hợp này, ta đã chọn centroid đầu tiên là (2,1). Ta cần tính toán khoảng cách Euclidean từ mỗi điểm đến centroid này và sau đó chọn điểm có khoảng cách lớn nhất làm centroid tiếp theo.

Khoảng cách từ các điểm đến centroid (2,1):

* Từ điểm (-1,2): (−1−2)2+(2−1)2=9+1=10(−1−2)2+(2−1)2​=9+1​=10​ (xấp xỉ 3.16)
* Từ điểm (0,0): (0−2)2+(0−1)2=4+1=5(0−2)2+(0−1)2​=4+1​=5​ (xấp xỉ 2.24)
* Từ điểm (0,1): (0−2)2+(1−1)2=4+0=2(0−2)2+(1−1)2​=4+0​=2
* Từ điểm (0,3): (0−2)2+(3−1)2=4+4=22(0−2)2+(3−1)2​=4+4​=22​ (xấp xỉ 2.83)
* Từ điểm (2,2): (2−2)2+(2−1)2=0+1=1(2−2)2+(2−1)2​=0+1​=1
* Từ điểm (2,3): (2−2)2+(3−1)2=0+4=2(2−2)2+(3−1)2​=0+4​=2
* Từ điểm (-1,2): (−1−2)2+(2−1)2=9+1=10(−1−2)2+(2−1)2​=9+1​=10​ (xấp xỉ 3.16)

Từ các khoảng cách trên, khoảng cách lớn nhất là 1010​, và điểm tương ứng là (-1,2).

Vậy câu trả lời là:

f. (-1,2)



**You**

Trong thuật toán DBSCAN, cho trước eps và minpts, nếu điểm p là density-reachable từ điểm q, những phát biểu nào sau đây là đúng? Chọn tất cả các đáp án đúng. Câu này liên quan đến một số khái niệm sau: Một điểm x là directly density reachable từ một điểm y nếu x∈Nϵ(y) và y là một core point. Nϵ(y) là ϵ -neighborhood của y . Một điểm x là density reachable từ một điểm y nếu tồn tại một dãy các điểm x0,x1,…,xk sao cho x=x0 và y=xk , và xi−1 là directly density reachable từ xi với mọi i=1,…,k . Hai điểm x và y là density connected nếu tồn tại một core point z , sao cho cả x và y là density reachable từ z . a. Nếu tồn tại điểm o là density-reachable từ điểm p, thì o cũng là density-reachable từ điểm q. b. Điểm p phải là directly density-reachable từ điểm q. c. Điểm q là density-connected đến điểm p. d. Điểm q là density-reachable từ điểm p.

ChatGPT

**ChatGPT**

Trong thuật toán DBSCAN, một số khái niệm quan trọng:

1. **Density-reachable**:
   * Một điểm 𝑝*p* được xem là density-reachable từ điểm 𝑞*q* nếu có một chuỗi các điểm 𝑥0,𝑥1,...,𝑥𝑘*x*0​,*x*1​,...,*xk*​ sao cho 𝑝=𝑥0*p*=*x*0​ và 𝑞=𝑥𝑘*q*=*xk*​, và mỗi 𝑥𝑖−1*xi*−1​ là directly density-reachable từ 𝑥𝑖*xi*​ cho mọi 𝑖=1,...,𝑘*i*=1,...,*k*.
2. **Directly density-reachable**:
   * Một điểm 𝑥*x* được xem là directly density-reachable từ điểm 𝑦*y* nếu 𝑥*x* nằm trong 𝜖*ϵ*-neighborhood của 𝑦*y* và 𝑦*y* là một core point.
3. **Density-connected**:
   * Hai điểm 𝑥*x* và 𝑦*y* được xem là density-connected nếu có một core point 𝑧*z* sao cho cả 𝑥*x* và 𝑦*y* đều density-reachable từ 𝑧*z*.

Với các khái niệm trên, chúng ta có thể đánh giá các phát biểu sau:

a. **Đúng**: Nếu tồn tại điểm 𝑜*o* là density-reachable từ điểm 𝑝*p*, thì 𝑜*o* cũng là density-reachable từ điểm 𝑞*q*. Điều này phản ánh khái niệm của mạng lưới dày đặc trong DBSCAN, nơi các điểm liên tục density-reachable với nhau.

b. **Sai**: Không cần thiết phải là directly density-reachable. Một điểm có thể là density-reachable thông qua một chuỗi các điểm density-reachable khác.

c. **Sai**: Không có đủ thông tin để kết luận điểm 𝑞*q* là density-connected đến điểm 𝑝*p*.

d. **Sai**: Điểm 𝑞*q* không nhất thiết phải là density-reachable từ điểm 𝑝*p*, điều này không được đảm bảo trong mọi trường hợp.

Vậy câu trả lời là:

a. ĐuˊngĐuˊng



**You**

chonj nhieu dap an

ChatGPT

**ChatGPT**

Trong trường hợp này, chúng ta có thể chọn nhiều đáp án đúng. Dựa trên các khái niệm của thuật toán DBSCAN, chúng ta có thể chọn các phát biểu sau:

a. Đúng: Nếu tồn tại điểm o là density-reachable từ điểm p, thì o cũng là density-reachable từ điểm q.

c. Sai: Không có đủ thông tin để kết luận điểm q là density-connected đến điểm p.

Vậy câu trả lời là: a, c.



**You**

Trong thuật toán DBSCAN, giả sử eps = 1cm và minpts = 1. Chọn ngẫu nhiên hai điểm p, q dữ liệu. Giả sử d(p, q) = 4cm. Cho biết các phát biểu nào sau đây là đúng? a. Điểm p và q phải thuộc khác cụm. b. Điểm p và q có thể thuộc khác cụm. c. Điểm p và q có thể không thuộc cụm nào. d. Điểm p và q phải thuộc cùng cụm.

ChatGPT

**ChatGPT**

Khi eps = 1cm và minpts = 1, có thể diễn ra như sau:

* Một điểm được coi là một cụm riêng nếu nó có ít nhất minpts điểm lân cận trong một bán kính eps.
* Một cụm có thể chứa nhiều điểm và mỗi điểm trong cụm đều kết nối với một điểm khác trong cụm đó thông qua các điểm lân cận.

Dựa trên điều này, giả sử có hai điểm p và q trong tập dữ liệu, và d(p, q) = 4cm. Khi đó:

a. Sai: Vì d(p, q) > eps, nên p và q không nhất thiết phải thuộc khác cụm. Chúng có thể thuộc cùng một cụm nếu có đủ điểm lân cận khác.

b. Đúng: Với d(p, q) > eps, có thể như vậy, nhưng điều này phụ thuộc vào các điểm khác trong tập dữ liệu và cách chúng kết nối với p và q.

c. Đúng: Với d(p, q) > eps, và minpts = 1, có thể p và q không thuộc bất kỳ cụm nào.

d. Sai: Vì d(p, q) > eps, nên không thể kết luận p và q phải thuộc cùng một cụm.

Vậy câu trả lời là: b và c.



**You**

Trong thuật toán DBSCAN, giả sử eps = 1cm và minpts = 1. Chọn ngẫu nhiên hai điểm p, q dữ liệu. Giả sử d(p, q) = 4cm. Cho biết các phát biểu nào sau đây là đúng? a. Điểm p và q phải thuộc khác cụm. b. Điểm p và q có thể thuộc khác cụm. c. Điểm p và q có thể không thuộc cụm nào. d. Điểm p và q phải thuộc cùng cụm.

ChatGPT

**ChatGPT**

Với eps = 1cm và minpts = 1, điều quan trọng cần lưu ý trong thuật toán DBSCAN là:

* Một điểm được coi là một cụm riêng nếu nó có ít nhất minpts điểm lân cận trong một bán kính eps.
* Một cụm có thể chứa nhiều điểm và mỗi điểm trong cụm đều kết nối với một điểm khác trong cụm đó thông qua các điểm lân cận.

Dựa trên điều này, giả sử có hai điểm p và q trong tập dữ liệu, và d(p, q) = 4cm.

Với eps = 1cm và minpts = 1, điều kiện cần để p và q thuộc cùng một cụm là cả hai điểm đều có ít nhất 1 điểm lân cận trong một bán kính eps = 1cm. Nhưng do d(p, q) > eps, nên không có điểm nào trong tập dữ liệu nằm trong phạm vi của eps tính từ cả p và q. Do đó, không thể kết luận rằng p và q phải thuộc cùng một cụm.

Cụ thể:

* a. Sai: Vì d(p, q) > eps, nên p và q không nhất thiết phải thuộc khác cụm.
* b. Đúng: Với d(p, q) > eps, có thể như vậy, nhưng điều này phụ thuộc vào các điểm khác trong tập dữ liệu và cách chúng kết nối với p và q.
* c. Đúng: Với d(p, q) > eps, và minpts = 1, có thể p và q không thuộc bất kỳ cụm nào.
* d. Sai: Vì d(p, q) > eps, nên không thể kết luận p và q phải thuộc cùng một cụm.

Vậy câu trả lời là: b và c.



**You**

chon 1 dap an

ChatGPT

**ChatGPT**

Trong trường hợp này, câu trả lời chính xác là:

b. Điểm p và q có thể thuộc khác cụm.

Với khoảng cách 𝑑(𝑝,𝑞)>𝑒𝑝𝑠*d*(*p*,*q*)>*eps*, không có điểm nào trong tập dữ liệu lân cận với cả p và q trong bán kính eps. Do đó, không thể kết luận rằng p và q thuộc cùng một cụm. Tuy nhiên, điều này không loại trừ khả năng p và q thuộc cùng một cụm nếu có các điểm khác lân cận với cả hai.