DBSCAN Clustering

1. Giới thiệu

* DBSCAN là một thuật toán phân cụm dựa trên mật độ thích hợp để sử dụng khi kiểm tra dữ liệu không gian
* Nhược điểm của phân cụm truyền thống:

+ Các kĩ thuật phân cụm truyền thống như Kmeans, phân cấp, phân cụm mờ có thể được sử dụng để nhóm dữ liệu theo cách không giám sát

+ Tuy nhiên khi áp dụng với các tác vụ với cụm hoặc hình dạng tùy ý trong cụm, kỹ thuật truyền thống có thể không đạt được kết quả tốt: các phần tử trong cụm có thể không thể có điểm tương tự , hiệu suất có thể kém.

+ Phân vùng dựa trên Kmeans tuy dễ hiểu và dễ thực hiện trong thực tế nhưng thuật toán không có khái niệm về các outlier, tất cả các điểm sẽ được gán cho một cụm nào đó ngay cả khi điểm đó không thuộc về cụm nào cả.

+ Trong lĩnh vực phát hiện bất thường, điều này gây ra vấn đề vì các điểm bất thường sẽ được gán cho cùng một cụm như các điểm bình thường. 🡪 outlier sẽ kéo trọng tâm cụm về phía chúng làm cho việc phân loại chúng thành các outlier trở nên khó khăn hơn

* Ưu điểm của phân cụm dựa trên mật độ :

+ Định vị các vùng của mật độ cao được phân tách với nhau bởi các khu vực có mật độ thấp.

+ Mật độ là số điểm trong bán kính chỉ định

* DBSCAN là một kiểu phân cụm dựa trên mật độ phổ biến
* DBSCAN đặc biệt hiệu quả trong nhận dạng lớp học trên một bối cảnh không gian
* Ưu điểm DBSCAN:

+ có thể tìm ra bất kì cụm hình dạng tùy ý mà không bị ảnh hưởng bởi nhiễu ( outlier)

+ DBSCAN có thểt tìm thất phần dày đặc hơn của các mẫu tập trung vào dữ liệu bằng cách bỏ qua khu vực ít nhiễu hoặc bỏ qua nhiễu .

1. DBSCAN

* Hoạt động dựa trên ý tưởng : nếu một điểm cụ thể thuộc về một cụm thì nó phải ở gần nhiều điểm khác trong cụm đó.
* Hoạt động dựa trên 2 tham số : bán kính ( R) và điểm tối thiểu.(M)
* DBSCAN: đến thăm từng điểm và tìm loại đầu tiên, sau đó nhóm các điểm dưới dạng cụm dựa trên các loại của chúng.

+ Chọn 1 điểm ngẫu nhiên

+ Kiểm tra xem đó có phải là core point hay không

+ Kết nối tất cả các điểm cốt lõi hàng xóm và đặt trong một cụm

( Một cụm được hình thành như ít nhất một điểm cốt lõi cộng với tất cả các điểm cốt lõi có thể tiếp cận và các đường viền của chúng.

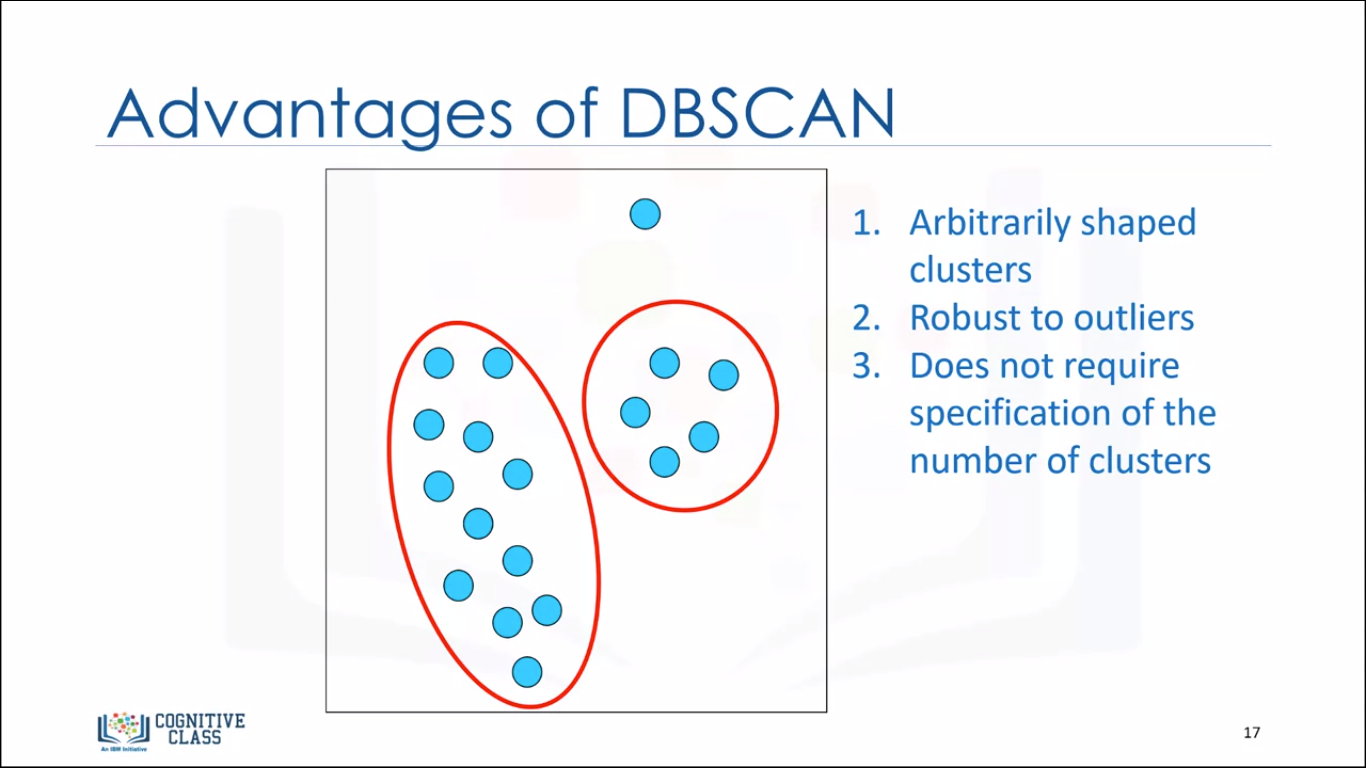
* Mỗi điểm trong tập dữ liệu có thể là core point , border point, outlier point

+ Core point ( Điểm cốt lõi) : nếu trong khu vực lân cận có ít nhất M điểm

+ Border point ( Điểm biên giới) : nếu vùng lân cận của nó chứa ít hơn M điểm dữ liệu , hoặc nó có thể truy cập từ một số điểm cốt lõi.

+ Outlier point ( Điểm ngoại lệ) : là một điểm không phải điểm cốt lõi, cũng không đủ gần để tiếp cận từ điểm cốt lõi

1. Advance DBSCAN



1. Câu hỏi

Which of the following are the characteristics of density-based clustering?

Density-based clustering algorithms have no notion of outliers.

Density-based clustering algorithms are proper for arbitrary shape clusters.

Density-based clustering algorithms locate regions of high density that are separated from one another by regions of low density.