ML2 WEEK 5

1) Ý nghĩa tham số radius, min sample trong thuật toán dbscan?

- radius: là bán kính của đường tròn có tâm là các điểm dữ liệu

- minPoints: số điểm dữ liệu tối tối thiểu để tạo thành 1 dense region. Ví dụ: nếu đặt minPoints là 3, thì cần ít nhất 3 điểm để tạo thành 1 dense region

- eps: Nếu radius quá nhỏ thì lượng lớn data sẽ ko được phân cụm. Mô hình sẽ coi chúng là outliers bởi vì có thể sẽ không thỏa mãn điều kiện về minPoints. Mặt khác nếu radius quá lướn thì các clusters sẽ merge với nhau và các cụm dữ liệu lúc này sẽ bị phân về cùng 1 cụm. Radius nên được chọn dựa vào khoảng cách của các điểm dữ liệu.

- minPoints: MinPts thấp giúp model phân nhiều cụm với nhiều noise hoặc outlier hơn. Nếu MinPts cao hơn sẽ đảm bảo việc phân cụm hiệu quả hơn, nhưng nếu Minpoints quá lớn thì các cụm nhỏ hơn sẽ được kết hợp thành các cụm lớn hơn dẫn đến model không phân đúng cụm.

2) So sánh ba thuật toán: kmean, GMM, dbscan.

1) K-Means Clustering:

K-mean là một thuật toán phân cụm dựa trên centroid hoặc dựa trên phân vùng. Thuật toán này phân chia tất cả các điểm trong không gian mẫu thành K nhóm tương tự. Độ tương tự thường được đo bằng Khoảng cách Euclidian.

Thuật toán như sau:

1 K centroid được đặt ngẫu nhiên, một cho mỗi cụm.
2 Khoảng cách của mỗi điểm từ mỗi centroid được tính toán
3 Mỗi điểm dữ liệu được gán cho tâm gần nhất của nó, tạo thành một cụm.
4 Vị trí của K centroid được tính toán lại.

2) GMM

Mô hình hỗn hợp Gauss là một mô hình xác suất dựa trên khoảng cách giả định tất cả các điểm dữ liệu được tạo ra từ sự kết hợp tuyến tính của các phân phối Gaussian đa biến với các tham số chưa biết. Giống như K-mean, nó tính đến các trung tâm của các phân bố Gaussian tiềm ẩn nhưng không giống như K-mean, cấu trúc hiệp phương sai của các phân phối cũng được tính đến. Thuật toán thực hiện thuật toán tối đa hóa kỳ vọng (EM) để tìm lặp đi lặp lại các tham số phân phối tối đa hóa thước đo chất lượng mô hình được gọi là khả năng xảy ra nhật ký.

Thuật toán như sau:

1 Khởi tạo bản phân phối k gaussian

2 Tính xác suất của sự kết hợp của mỗi điểm với mỗi phân phối

3 Tính toán lại các thông số phân phối dựa trên xác suất của từng điểm được liên kết với các phân bố

4 Lặp lại quy trình cho đến khi khả năng ghi được tối đa hóa

3) DBSCAN

DBScan là một thuật toán phân cụm dựa trên mật độ. Thực tế chính của thuật toán này là vùng lân cận của mỗi điểm trong một cụm nằm trong bán kính nhất định (R) phải có số điểm tối thiểu (M). Thuật toán này đã tỏ ra cực kỳ hiệu quả trong việc phát hiện các ngoại lệ và xử lý nhiễu.

Thuật toán như sau:

1 Loại của mỗi điểm được xác định. Mỗi điểm dữ liệu trong tập dữ liệu của chúng tôi có thể là một trong những điểm sau:

- Điểm cốt lõi: Điểm dữ liệu là điểm cốt lõi nếu có ít nhất M điểm trong vùng lân cận của nó, tức là nằm trong bán kính được chỉ định (R).
- Điểm biên giới: Một điểm dữ liệu được phân loại là điểm BIÊN GIỚI nếu: Vùng lân cận của nó chứa ít hơn M điểm dữ liệu, hoặc Nó có thể đạt được từ mộ số điểm cốt lõi là nó nằm trong khoảng cách R từ điểm cốt lõi.
- Điểm ngoại lệ: Điểm ngoại lệ là một điểm không phải là điểm cốt lõi và cũng không đủ gần để có
  thể tiếp cận được từ điểm cốt lõi.

2 Các điểm ngoại lệ bị loại bỏ.

3 Các điểm cốt lõi là láng giềng được kết nối và đặt trong cùng một cụm.

5 Các điểm biên giới được giao cho từng cụm.

3) Khi nào nên sử dụng thuật toán nào? cho ví dụ?

SD GMM :khi đối mặt với phương sai thay đổi có dạng không xác định

SD KMEANS : khi có ý tưởng về việc muốn phân bao nhiêu cụm

SD DBSCAN : thông thường các vùng không gian có mật độ cao sẽ xen kẽ bởi các vùng không gian có mật độ thấp. Nếu như phải dựa vào mật độ để phân chia thì khả năng rất cao những tâm cụm sẽ tập trung vào những vùng không gian có mật độ cao trong khi biên sẽ rơi vào những vùng không gian có mật độ thấp => DBSCAN được sd để phân cụm data sao cho một cụm trong không gian dữ liệu là một vùng có mật độ điểm cao được ngăn cách với các cụm khác bằng các vùng liền kề có mật độ điểm thấp