Homework DBSCAN

Exercise 2: Ý nghĩa tham số radius, min sample trong thuật toán dbscan? Nếu chỉ số lớn, nhỏ ảnh hưởng thế nào tới thuật toán?

epsilon : Một giá trị khoảng cách được sử dụng để xác định vùng lân cận epsilon của bất kỳ điểm dữ liệu nào

- Nếu epsilon được chọn quá nhỏ, một phần lớn dữ liệu sẽ không được phân cụm và được xem là nhiễu
- Nếu epsilon được chọn quá cao, các cụm sẽ hợp nhất và phần lớn các điểm sẽ nằm trong cùng một cụm.

min samples: Số lượng tối thiểu các điểm láng giềng xung quanh một điểm để xác định một điểm lõi (core point) và số lượng này cũng đã bao gồm điểm lõi. Tương đương với minPts+1

- MinPts = 1: giá trị không có ý nghĩa, vì khi đó mọi điểm bản thân nó đều là một cụm.
- MinPts = 2: kết quả sẽ giống như phân cụm phân cấp (hierarchical clustering) với single linkage với biểu đồ dendrogram được cắt ở độ cao epsilon.

  ⇒ MinPts ít nhất phải bằng 3

Exercise 3: Biến đổi lại và so sánh ba thuật toán: kmean, GMM, dbscan. Khi nào nên sử dụng thuật toán nào? cho ví dụ?

Ưu và nhược điểm của K-means

Ưu điểm:

- Là 1 thuật toán đơn giản và phổ biến nhất giải quyết vấn đề phân cụm

Nhược điểm:

- Thuật toán rất nhạy cảm với outliers và noise: Khi xuất hiện outliers thì thường khiến cho tâm cụm bị chệch và do đó dự báo cụm không còn chuẩn xác.
- Chúng ta cần phải xác định trước số cụm cho thuật toán. Vị trí tâm của cụm sẽ bị phụ thuộc vào điểm khởi tạo ban đầu của chúng: Những vị trí khởi tạo khác nhau có thể dẫn tới cách phân cụm khác nhau, mặc dù thuật toán có cùng thiết lập số cụm.
- Gặp vấn đề khi các nhóm có kích thước, mật độ khác nhau hoặc hình dáng không phải hình cầu.
- K-means là hard assignment
- Với những cụm dữ liệu chồng lên nhau thì sẽ phân loại sai

=> Trường hợp sử dụng: kích thước cụm đồng đều, hình học phẳng, không quá nhiều cụm

Ưu và nhược điểm của GMM

Ưu điểm:

- Xử lý nhiều hình dạng hơn, chủ yếu là các cụm tạo thành hình elip (K-Means chỉ thực sự tốt ở các cụm có dạng gần giống hình cầu)

- Soft assignment: Trong GMM 1 điểm có thể thuộc vào nhiều cluster với mức độ khác nhau. Điều này hữu ích trong một số task như 1 người có thể quan tâm đến nhiều chủ đề(phân loại đối tượng trên Facebook).

Nhược điểm:

- Không xác định chính xác được một số cụm có hình dạng khác

=> Trường hợp sử dụng: ước tính mật độ và hình học phẳng

Ưu và nhược điểm của DBScan

Ưu điểm:

- Không cần khai báo trước số lượng cụm cần phân chia
- Tự động loại bỏ được các điểm dữ liệu nhiễu
- Nó phân thành các cụm có hình dạng bất kỳ và có thể không cùng kích thước
- Tốc độ tính toán nhanh

Nhược điểm:

- Không hiệu quả đối với những dữ liệu có mật độ thưa thớt hoặc có mật độ thay đổi
- Nhạy cảm với các thông số eps và minPts.

=> Trường hợp sử dụng: kích thước cụm không đồng đều và hình học không phẳng