

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

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

Nếu Radius được chọn quá nhỏ, một phần lớn dữ liệu sẽ không được phân cụm và được xem là *noise*; trong khi đối với giá trị Radius quá cao, các cụm sẽ hợp nhất và phần lớn các điểm sẽ nằm trong cùng một cụm.

Min sample: Là một ngưỡng số điểm dữ liệu tối thiểu được nhóm lại với nhau nhằm xác định một *vùng lân cận radius* có mật độ cao. Số lượng Min sample không bao gồm điểm ở tâm.

Nếu Min sample thấp có nghĩa là nó sẽ xây dựng nhiều cụm hơn từ là tiếng ồn(Min sample ít nhất là 3), trong khi Min sample cao có nghĩa nó có thể bao gồm tất cả các điểm trong tập dữ liệu(nên chọn Min sample = $2 * \text{dim}$, dim là chiều của dữ liệu, còn đối với dữ liệu 2 chiều nên chọn Min sample = 4).

3.