

Machine Learning com Python

Prof. Luciano Galdino

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Agrupamento Espacial de Aplicações com Ruído Baseado em Densidade.

Mais rápido que o Kmeans e o hierárquico.

Trabalha muito bem com outliers.

Bom para dados que contêm clusters de densidade semelhante.

Encontra amostras centrais de alta densidade e expande clusters a partir delas.

Requer dois parâmetros:

ϵ (eps): raio de alcance.

min_samples: quantidade mínima de pontos necessários para formar um cluster.

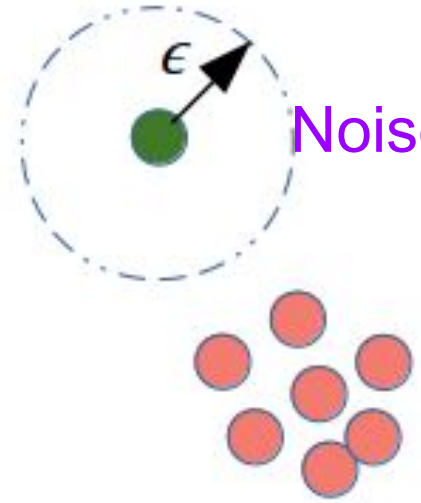
Começa com um ponto de partida aleatório.

Se a vizinhança ϵ desse ponto contiver pontos suficientes, um cluster é iniciado. Caso contrário, o ponto é rotulado como ruído.

Border Point

Core Point

Noise



Vantagens

Não exige especificação do número de clusters.

Pode encontrar clusters de formato arbitrário.

É robusto a outliers.

Requer apenas dois parâmetros.

Mais rápido.

Desvantagens

Pode não agrupar bem conjuntos de dados com grandes diferenças de densidades.

Pode ser difícil escolher um limite de distância (ϵ) adequado.