DBSCAN

Procedure 方法

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) adalah sebuah algoritma machine learning unsupervised yang berfungsi untuk membuat cluster dari data yang diberikan. Berbeda dari Kmeans, DBSCAN tidak akan mencoba untuk memasukan semua titik ke cluster, titik yang kurang dekat untuk masuk cluster akan ditandai sebagai noise.

1. **Inisialisasi model dengan parameter epsilon dan min_sample**
   Epsilon disini adalah jarak maksimum antara dua titik data untuk dianggap berada di dalam cluster yang sama. Min_sample disini berarti jumlah minimum titik untuk diperbolehkan membuat sebuah cluster.
2. **Inisialisasi Centroid**
   Centroid adalah titik pusat dari kluster. Centroid dapat dinisialisasi secara random, atau dengan kmeans++ random sesuai distribusi jarak antar titik agar centroid lebih tersebar dengan baik.
3. **Tentukan jenis titik**
   Terdapat tiga jenis titik di DBSCAN: Core, Border, dan Noise. Titik Core adalah titik yang memiliki setidaknya min_sample titik yang berada dalam radius epsilonnya. Titik Border adalah titik yang masuk ke dalam radius epsilon titik core, namun tidak memiliki jumlah min_sample titik di dalam radius epsilonnya sendiri. Terakhir, titik noise adalah titik yang tidak masuk kedalam cluster.
4. **Buat cluster**
   Iterasi data yang diberikan. Cek apakah titik sudah dikunjungi. Jika belum, cari jumlah tetangga yang masuk ke dalam radius epsilonnya. Jika lebih dari min_sample, coba buat cluster dari situ. Lalu propagasikan cluster melalui pengecekan radius tetangganya dalam sebuah queue. Selama propagasi terus berlanjut, maka semua yang terkena akan menjadi cluster yang sama. Jika propagasi berhenti, ganti id_cluster dan lanjut ke titik yang belum dikunjungi lainnya untuk mencoba membuat cluster baru.
5. **Evaluasi**
   Hasil clustering bisa dievaluasi dengan silhouette score.

VS Sklearn

Hasil di notebook DoE menghasilkan clustering yang sama persis.