

DBSCAN

1. Cara kerja algoritma

Algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah metode clustering yang digunakan untuk mengelompokkan data berdasarkan kepadatan titik data dalam ruang. DBSCAN sangat efektif dalam mengidentifikasi cluster dengan bentuk yang tidak teratur dan dapat menangani noise dalam dataset.

1. Menentukan parameter

- Epsilon

Jarak maksimum antara dua titik untuk dianggap berada dalam cluster yang sama.

- MinPts

Jumlah minimum titik yang diperlukan untuk membentuk sebuah cluster. Parameter ini menentukan seberapa padat suatu area untuk dianggap sebagai cluster.

2. Menentukan titik inti, titik border, dan noise

Setelah parameter ditentukan, langkah berikutnya adalah mengidentifikasi jenis titik:

- Titik inti

Titik yang memiliki setidaknya MinPts titik lain dalam radius epsilon.

- Titik border

Titik yang berada dalam radius epsilon dari titik inti tetapi tidak memiliki cukup titik di sekitarnya untuk menjadi titik inti.

- Noise

Titik yang tidak termasuk dalam cluster mana pun, yaitu tidak memenuhi syarat untuk menjadi titik inti atau border.

3. Membangun cluster

Proses clustering dimulai dengan memilih titik acak yang belum dikunjungi. Jika titik tersebut adalah titik inti, maka cluster baru akan dibentuk. Semua titik yang dapat dijangkau dari titik inti (dalam radius epsilon) akan ditambahkan ke cluster. Proses ini akan berlanjut dengan memeriksa titik-titik baru yang ditambahkan ke cluster untuk melihat apakah mereka juga merupakan titik inti.

4. Ulangi proses

Proses ini diulang untuk setiap titik yang belum dikunjungi dalam dataset. Jika titik yang dipilih adalah titik border atau noise, kita akan melanjutkan ke titik berikutnya. Proses ini berlanjut hingga semua titik dalam dataset telah dikunjungi..

5. Evaluasi

Setelah proses clustering selesai, kita dapat mengevaluasi hasil DBSCAN dengan memvisualisasikan cluster yang terbentuk dan menganalisis jumlah titik yang teridentifikasi sebagai noise serta menggunakan metrik seperti Silhouette score.

2. Perbandingan model buatan sendiri dan library

Berdasarkan hasil evaluasi, kedua model DBSCAN, baik yang menggunakan scikit-learn, maupun implementasi sendiri, menunjukkan performa yang sama dalam hal Silhouette Score, yaitu sekitar 0.13. Ini menunjukkan bahwa implementasi sendiri berhasil menangkap struktur data dengan baik dan menghasilkan clustering yang sebanding dengan model yang sudah teruji.