Rizqika Mulia Pratama
13522126

K-Means

1. Cara kerja algoritma

Algoritma K-Means adalah metode clustering yang digunakan untuk mengelompokkan data ke dalam sejumlah kelompok (cluster) berdasarkan kesamaan atau kedekatan antar data. K-Means sangat populer dalam analisis data dan machine learning karena kesederhanaannya dan efisiensinya dalam menangani dataset besar.

1. Menentukan jumlah cluster (K)

   Langkah pertama dalam K-Means adalah menentukan jumlah cluster (K) yang diinginkan. Pemilihan nilai K sangat penting karena akan mempengaruhi hasil clustering. Nilai K yang terlalu kecil dapat mengabaikan struktur data yang lebih kompleks, sedangkan nilai K yang terlalu besar dapat menyebabkan overfitting.

2. Inisialisasi centroid

   Selain menentukan K, langkah selanjutnya adalah menginisialisasi posisi centroid untuk setiap cluster. Centroid adalah titik pusat dari cluster yang akan diperbarui selama proses iterasi. Inisialisasi dapat dilakukan secara acak dengan diperbarui selama proses iterasi. Inisialisasi dapat dilakukan secara acak dengan memilih K titik dari dataset sebagai centroid awal.

3. Menghitung jarak dan mengelompokkan data

   Pada langkah ini, kita menghitung jarak antara setiap titik data dengan centroid dari setiap cluster. Jarak yang umum digunakan adalah jarak Euclidean. Setelah menghitung jarak, setiap titik data akan dikelompokkan ke dalam cluster yang memiliki centroid terdekat.

4. Memperbarui centroid

   Setelah semua titik data dikelompokkan, langkah selanjutnya adalah memperbarui posisi centroid. Centroid baru dihitung sebagai rata-rata dari semua titik data yang termasuk dalam cluster tersebut.

5. Iterasi

   Langkah 3 dan 4 diulang hingga posisi centroid tidak berubah secara signifikan atau hingga jumlah iterasi maksimum tercapai. Proses ini memastikan bahwa cluster yang terbentuk stabil dan tidak berubah lagi.

6. Evaluasi

   Setelah proses clustering selesai, kita dapat mengevaluasi hasil K-Means dengan menggunakan metrik seperti Silhouette Score untuk mengukur kualitas cluster yang terbentuk.

Rizqika Mulia Pratama
13522126

2. Perbandingan model buatan sendiri dan library

Berdasarkan hasil evaluasi, model implementasi sendiri menunjukkan nilai Silhouette Score lebih tinggi (0.29) dibandingkan dengan model K-Means dari scikit-learn (0.22). Silhouette Score adalah metrik yang digunakan untuk mengevaluasi seberapa baik objek dalam satu cluster lebih mirip satu sama lain dibandingkan dengan objek di cluster lain. Nilai yang lebih tinggi menunjukkan bahwa cluster yang terbentuk lebih baik dan lebih terpisah.

Hasil yang lebih baik ini mungkin disebabkan karena penggunaan K-Means++, hal ini dapat memberikan hasil yang lebih baik dibandingkan dengan inisialisasi acak yang mungkin digunakan oleh scikit-learn.