Nama : Ferza Reyaldi
NIM : 09021281924060
Mata Kuliah : Data Mining

Tugas Pertemuan 3 (Pertemuan 13)
Pelajari & tuliskan penjelasan tentang metode atau algoritma K Means!

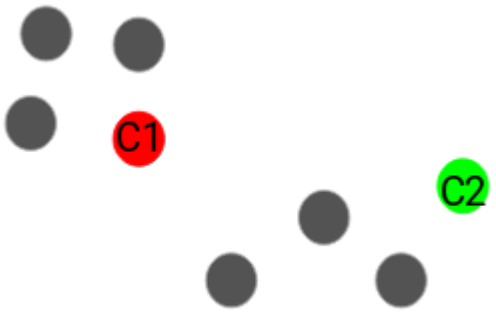
Jawab:
Algoritma K-Means adalah jenis *unsupervised learning*, yang digunakan untuk mengkategorikan data yang tidak berlabel, yaitu data tanpa kategori atau grup yang ditentukan. K-means adalah algoritma berbasis centroid, atau algoritma berbasis jarak, di mana kami menghitung jarak untuk menetapkan titik ke sebuah cluster. Di K-Means, setiap cluster dikaitkan dengan centroid.

Tujuan utama dari algoritma K-Means adalah untuk meminimalkan jumlah jarak antara titik dan masing-masing cluster centroid.

Cara Kerja Algoritma K-Means:

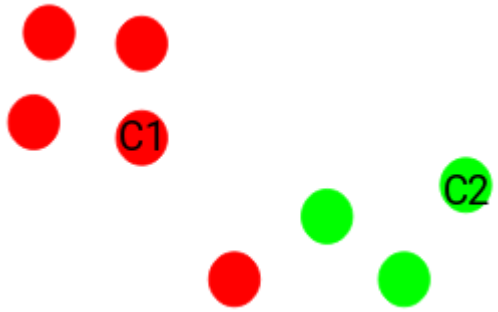
Misalkan diberi contoh bahwa dimiliki 8 poin dan akan diterapkan algoritma k-means untuk membuat cluster untuk poin-poin tersebut.

1) Pilih jumlah cluster k.
2) Pilih k titik acak dari data sebagai centroid.
   Misal diharapkan data terdiri 2 cluster, sehingga k = 2. Kemudian dipilih secara acak 2 lokasi centroid seperti gambar berikut.

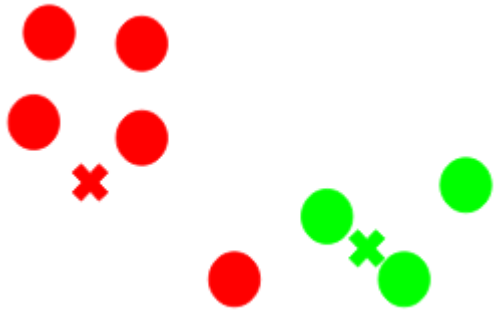


Lingkaran merah dan hijau mewakili centroid untuk cluster.

3) Tetapkan semua titik ke centroid cluster terdekat



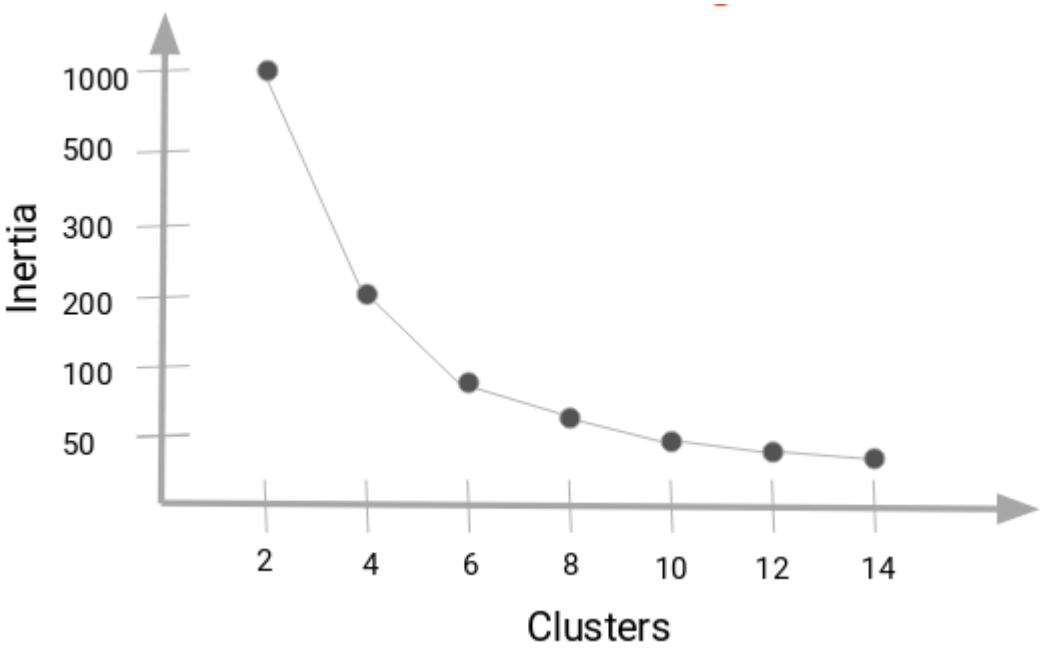
4) Hitung ulang centroid dari cluster yang baru terbentuk
   Setelah ditetapkan semua titik ke salah satu cluster, langkah selanjutnya adalah menghitung centroid dari cluster yang baru terbentuk:



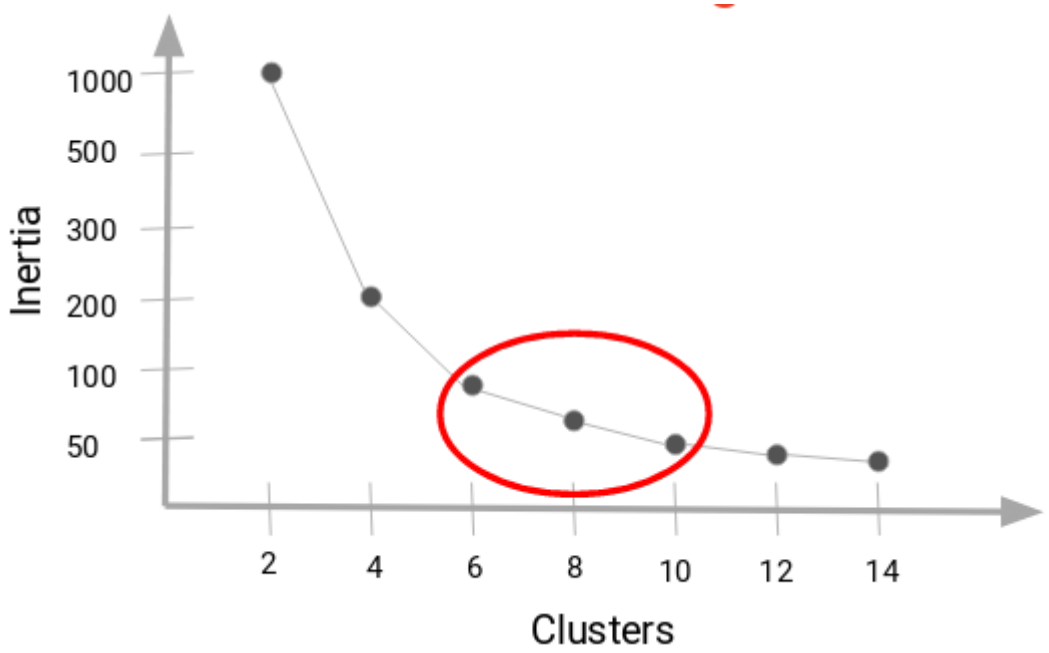
   Tanda silang berwarna merah dan hijau adalah centroid baru.

5) Ulangi langkah 3 dan 4 sampai memenuhi *stopping criteria clustering*, yaitu sebagai berikut:
   - Centroid dari cluster yang baru terbentuk tidak berubah
   - Poin tetap di cluster yang sama
   - Jumlah maksimum iterasi tercapai

Bagaimana Cara Memilih nilai K (Jumlah Cluster) yang Tepat di K-Means Clustering?
Satu hal yang dapat dilakukan adalah memplot grafik, juga dikenal sebagai *elbow curve*, di mana sumbu x akan mewakili jumlah cluster dan sumbu y akan menjadi matrik evaluasi. Katakanlah inersia untuk saat ini.



Nilai cluster dimana penurunan nilai inersia ini menjadi konstan dapat dipilih sebagai nilai cluster yang tepat.



Berdasarkan grafik diatas dapat dipilih sejumlah cluster antara 6 dan 10 dengan pertimbangan biaya komputasi saat memutuskan jumlah cluster. Semakin tinggi jumlah cluster, biaya komputasi juga akan meningkat.

Pros:

- Sederhana.
- Fleksibel.
- Cocok untuk kumpulan data besar.

- Efisien, bagus dalam mensegmentasi kumpulan data yang besar. Efisiensinya tergantung pada bentuk cluster. K-means bekerja dengan baik di cluster hyper-spherical.
- Kompleksitas waktu adalah linier dalam jumlah objek data sehingga meningkatkan waktu eksekusi. Tidak perlu lebih banyak waktu dalam mengklasifikasikan karakteristik serupa dalam data seperti algoritma hierarkis.
- Tight clusters, dibandingkan dengan algoritma hierarkis, k-means menghasilkan cluster yang lebih ketat terutama dengan cluster globular.
- Mudah diinterpretasikan.
- Biaya komputasi k-means lebih cepat dan efisien, O(K*n*d).
- Analisis K-means meningkatkan akurasi pengelompokan dan memastikan informasi tentang domain masalah tertentu tersedia. Modifikasi algoritma k-means berdasarkan informasi ini meningkatkan akurasi cluster.
- Spherical clusters, Mode clustering ini bekerja dengan baik ketika berhadapan dengan spherical cluster. Ini beroperasi dengan asumsi distribusi fitur bersama karena setiap cluster berbentuk bola. Semua fitur atau karakter cluster memiliki varians yang sama dan masing-masing tidak tergantung pada yang lain.

Cons:

- Tidak ada Kumpulan kluster yang tidak optimal.
- Kurangnya konsistensi, K-means clustering memberikan hasil yang bervariasi pada proses yang berbeda dari suatu algoritma. Sebuah pilihan acak pola cluster menghasilkan hasil clustering yang berbeda mengakibatkan inkonsistensi.
- Uniform effect, Menghasilkan cluster dengan ukuran yang seragam bahkan ketika data input memiliki ukuran yang berbeda.
- Cara data diurutkan dalam membangun algoritma memengaruhi hasil akhir kumpulan data.
- Sensitivitas terhadap skala, Mengubah atau mengubah skala dataset baik melalui normalisasi atau standardisasi akan sepenuhnya mengubah hasil akhir.
- Ketika berhadapan dengan dataset yang besar, melakukan teknik dendrogram akan membuat komputer crash karena banyak beban komputasi dan batasan Ram.
- Hanya dapat dilakukan pada data numerik.
- Beroperasi dalam asumsi, Teknik pengelompokan K-means mengasumsikan bahwa cluster bola dan setiap cluster memiliki jumlah yang sama untuk pengamatan. Asumsi bola harus dipenuhi. Algoritma tidak dapat bekerja dengan cluster dengan ukuran yang tidak biasa.
- Diharuskan menentukan jumlah cluster (K) di awal algoritma.
- Sulit untuk memprediksi nilai-k atau jumlah cluster. Juga sulit untuk membandingkan kualitas cluster yang dihasilkan.