MODUL PERTEMUAN 7

K-Means

A. Tujuan

1. Praktikan mampu memahami yang dimaksud dengan K-Means
2. Praktikan mampu mengimplementasikan K-Means ke dalam python
3. Praktikan mampu memahami cara kerja metode K-Means

B. Landasan Teori

Clustering

Clustering adalah pengelompokkan data-data menjadi beberapa cluster sehingga objek di dalam satu cluster memiliki banyak kesamaan dan memiliki banyak perbedaan dengan objek di cluster lain. Perbedaan dan persamaannya biasanya berdasarkan nilai atribut dari objek tersebut dan dapat juga berupa perhitungan jarak. Objek yang di dalam cluster mirip satu sama dengan yang lainnya, dan mempunyai perbedaan dengan objek dari cluster yang lain.

Terdapat dua pendekatan utama yaitu clustering dengan pendekatan partisi (non hirarki) dan clustering dengan pendekatan hirarki.

1. Clustering dengan pendekatan non hirarki

   Clustering ini mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada dimulai dengan menentukan terlebih dahulu jumlah cluster yang diinginkan (dua, tiga, atau yang lain). Setelah jumlah

cluster ditentukan, maka proses cluster dilakukan dengan tanpa mengikuti proses hirarki. Metode ini biasa disebut “Cluster K-Means”.

2. Clustering dengan pendekatan hirarki

   Clustering dilakukan melalui pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian diteruskan pada obyek yang lain dan seterusnya hingga cluster akan membentuk semacam ‘pohon’ dimana terdapat tingkatan (hirarki) yang jelas antar objek, dari yang paling mirip hingga yang paling tidak mirip. Alat yang membantu untuk memperjelas proses hirarki ini disebut “dendogram”.

K-Means Clustering

K-Means clustering merupakan salah satu unsupervised machine learning algorithms dengan melakukan analisa data dalam pemrosesan model tanpa supervisi dan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. K-Means clustering merupakan salah satu metode analisis cluster non hirarki yang berusaha untuk mempartisi objek yang ada ke dalam satu atau lebih cluster/kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokkan dalam satu cluster yang sama dan objek yang memiliki karakter berbeda dikelompokkan pada cluster lain

K-Means clustering bertujuan untuk meminimalisasikan objective function yang diset dalam proses clustering dengan cara meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya serta menemukan grup/kelompok dalam data, dengan jumlah grup/kelompok yang diwakili oleh variabel K (jumlah cluster yang diinginkan).

Kelebihan dari K-Means yaitu :

- Algoritma yang sederhana sehingga mudah dipahami
- Pemrosesan yang cepat
- Tersedia pada berbagai tools atau software
- Penerapan yang mudah
- Selalu memberikan hasil terlepas seperti apa datanya

Kekurangan dari K-Means yaitu :

- Hasilnya sensitif terhadap jumlah cluster (K)
- Sensitif terhadap inisialisasi “seed”
- Sensitif terhadap pencilan atau outlier
- Sensitif terhadap data dengan variabel yang memiliki skala berbeda
- Mengasumsikan setiap cluster berbentuk menyerupai lingkaran(spherical) dan kesulitan jika bentuk cluster yang memiliki bentuk berbeda

Menghitung Nilai K Optimal

Kinerja algoritma K-Means yang efisien sangat bergantung pada ukuran cluster yang dibentuk. Terdapat beberapa metode yang dapat digunakan dalam menemukan ukuran kluster atau nilai K pada algoritma K-Means Clustering, sebagai berikut.

a. Metode Elbow

Metode ini digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik. Metode ini memberikan ide/gagasan dengan cara memilih nilai kluster dan kemudian menambah nilai klaster tersebut untuk dijadikan model data dalam penentuan cluster terbaik. Dan selain itu persentase perhitungan yang dihasilkan menjadi pembanding antara jumlah cluster yang ditambah. Hasil persentase yang berbeda dari setiap nilai klaster dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai klaster pertama dengan nilai klaster kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai klaster tersebut yang terbaik.

Untuk mendapatkan perbandingannya adalah dengan menghitung SSE (*Sum of Square Error*) dari masing-masing nilai kluster. Karena semakin besar jumlah cluster K maka nilai SSE akan semakin kecil. Berikut merupakan perhitungan rumus SSE.

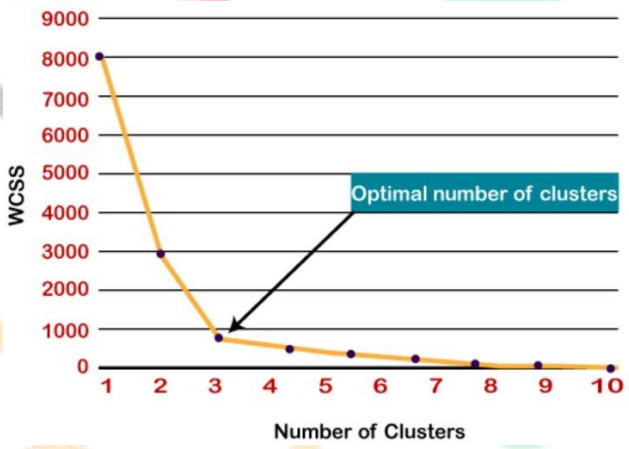
$$SSE = \sum_{K=1}^{K} \sum_{x_i e S_k} \|x_i - c_k\|_2^2$$

Algoritma metode Elbow dalam menentukan nilai K pada K-Means Clustering :

1. Mulai

2. Inisialisasi awal nilai K
3. Naikkan nilai K
4. Hitung hasil SSE dari tiap nilai K
5. Lihat hasil SSE dari nilai K yang turun secara drastis
6. Tetapkan nilai K yang berbentuk siku

Gambar berikut menunjukkan bagaimana grafik hasil metode elbow serta penentuan nilai K. Setelah dilihat akan ada beberapa nilai K yang mengalami penurunan paling besar dan selanjutnya hasil dari nilai K akan turun secara perlahan-lahan sampai hasil dari nilai K tersebut stabil. Misalnya, saat nilai K=2 ke K=3, terjadi penurunan, selanjutnya K=3 ke K=4 terjadi lagi penurunan drastis yang membentuk siku pada titik K=3. Oleh karena itu, nilai kluster K yang dianggap optimal adalah K=3.



b. Metode Silhouette

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan cluster, seberapa baik suatu objek ditempatkan dalam suatu kluster. Metode ini merupakan gabungan dari *metode cohesion dan separation*. *Cohesion* diukur dengan menghitung seluruh objek yang terdapat dalam sebuah kluster dan *separation* diukur dengan menghitung jarak rata-rata setiap objek dalam sebuah cluster dengan cluster terdekatnya.

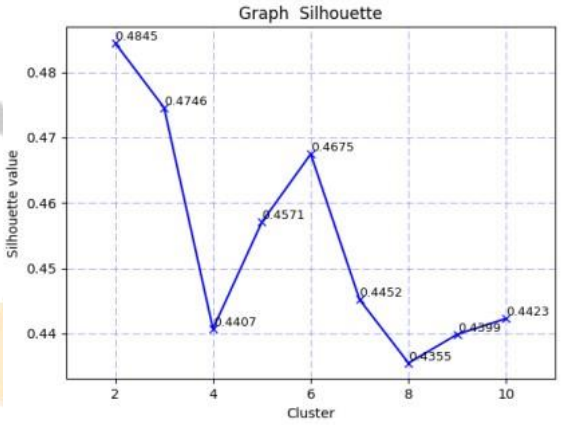
Nilai silhouette untuk keseluruhan data dengan jumlah klaster k, dapat didefinisikan sebagai sil (k) yang dihitung dengan persamaan berikut yakni rata-rata silhouette value untuk semua kluster.

$$\boldsymbol{sil\,(c) = sil\,(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(c_i)}$$

Keterangan:

sil (k) : nilai silhouette semua cluster
|k| : banyaknya *cluster k*
sil (c_i) : rata-rata nilai *silhouette*

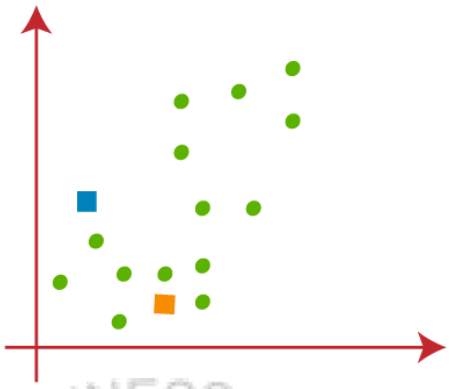
Gambar berikut menunjukkan bagaimana grafik hasil metode silhouette dalam penentuan nilai K. Dilakukan pengujian dengan jumlah K=2 sampai K=10. Berdasarkan hasil perhitungan metode *sillhoutter coefficient*, terlihat bahwa nilai *silhouette* maksimum ada pada K=2.



Algoritma K-Means Clustering

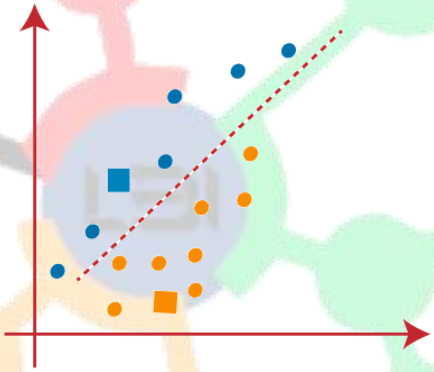
1. Tentukan jumlah cluster
2. Pilih titik acak sebanyak K

   Titik ini akan menjadi titik centroid awal



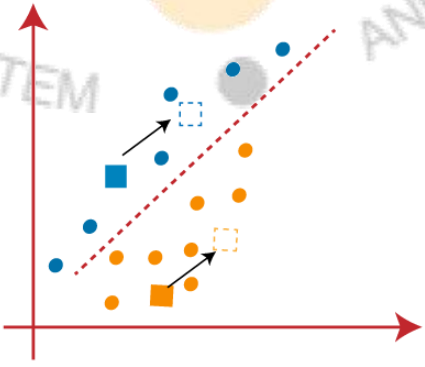
3. Berikan label pada semua data berdasarkan titik centroid terdekat

   Label yang diberikan mengikuti titik centroid dari setiap cluster yang dapat dihitung dengan berbagai metode, seperti Euclidean Distance.

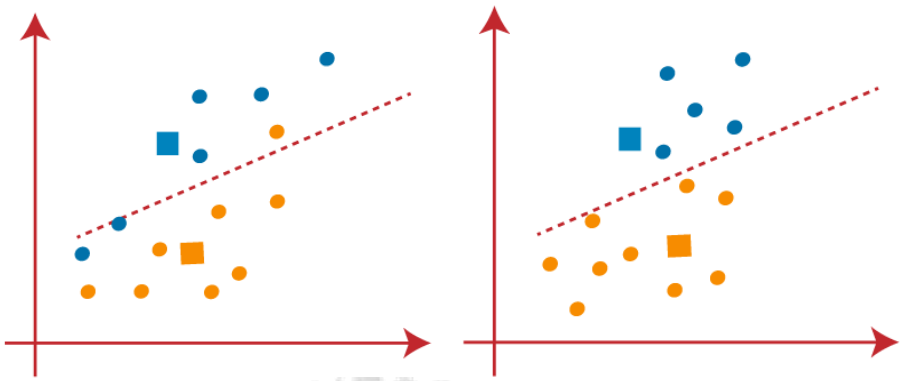


4. Tentukan titik centroid baru berdasarkan cluster yang terbentuk

   Titik centroid awal berpindah ke lokasi baru berdasarkan cluster yang terbentuk.

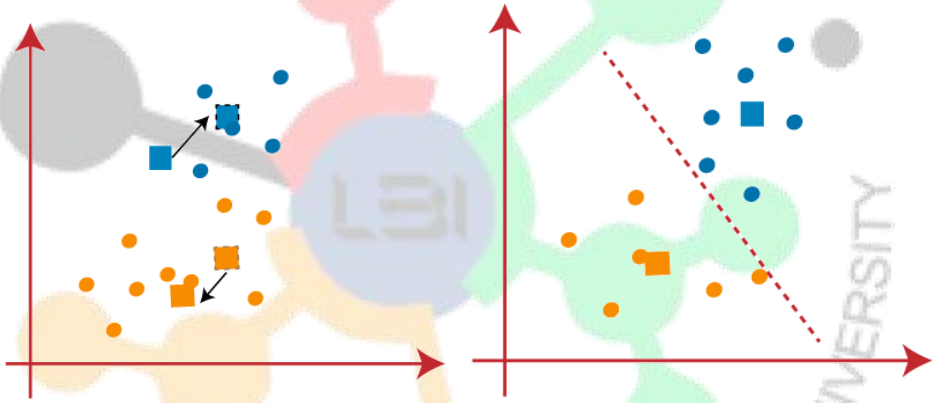


5. Lakukan pelabelan ulang data berdasarkan jarak terhadap centroid baru



6. Ulangi langkah 4 dan 5 hingga tidak ada pergerakan lagi

   Lakukan perulangan untuk mencari centroid baru hingga tidak ada lagi perpindahan centroid pada setiap cluster.



Berikut hasil akhir setelah tidak ditemukan perpindahan pada centroid di setiap cluster