Aplikasi Data Scientist

K-Means Clustering

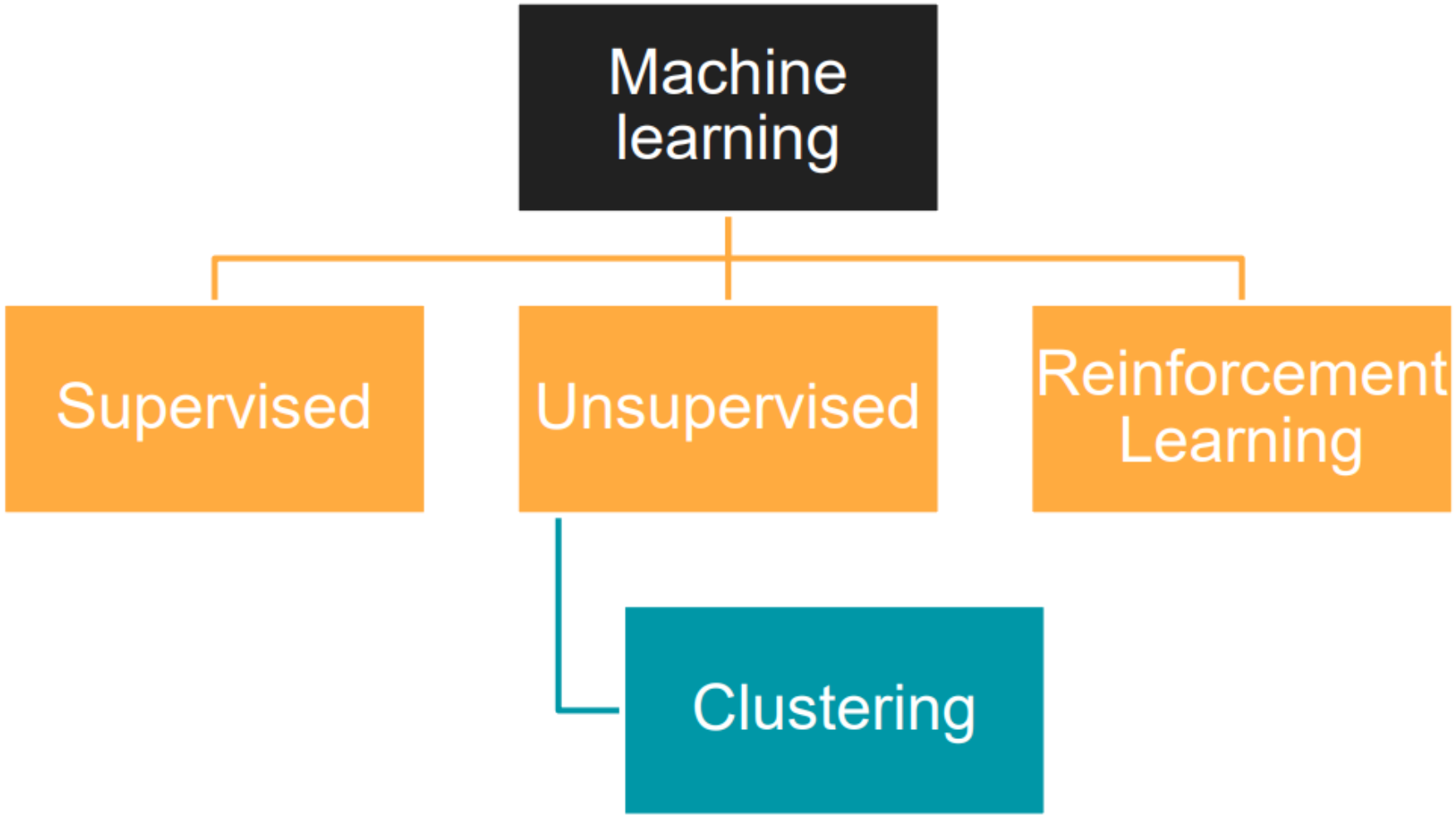
Ledy Elsera Astrianty, S.Kom., M.Kom

S1 Informatika – Univ. Teknologi Yogyakarta

Referensi:

Zhi-Hua Zhou, Shaowu Liu, "Machine Learning", Springer,2021

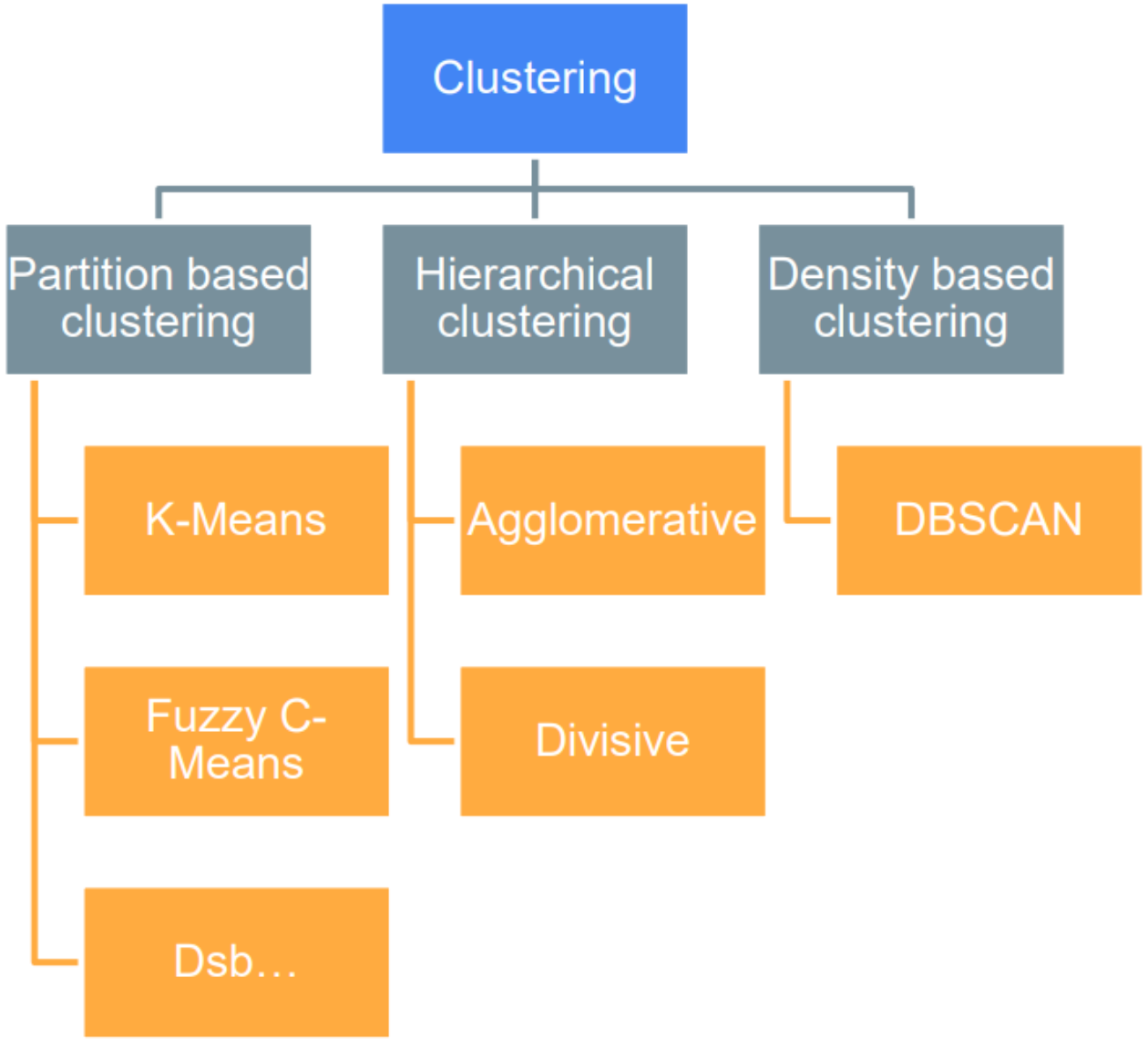
Related References: articles, books, Proceeding and Journals


Machine learning
Supervised
Unsupervised
Reinforcement Learning
Clustering

Teknik clustering melakukan pembelajaran mesin tanpa ada supervisi untuk menyelesaikan suatu masalah

Clustering juga dapat dianalogikan sebagai tugas mengidentifikasi subkelompok dalam data sedemikian rupa sehingga titik data dalam subkelompok yang sama (cluster) sangat mirip sedangkan titik data dalam cluster yang berbeda sangat berbeda

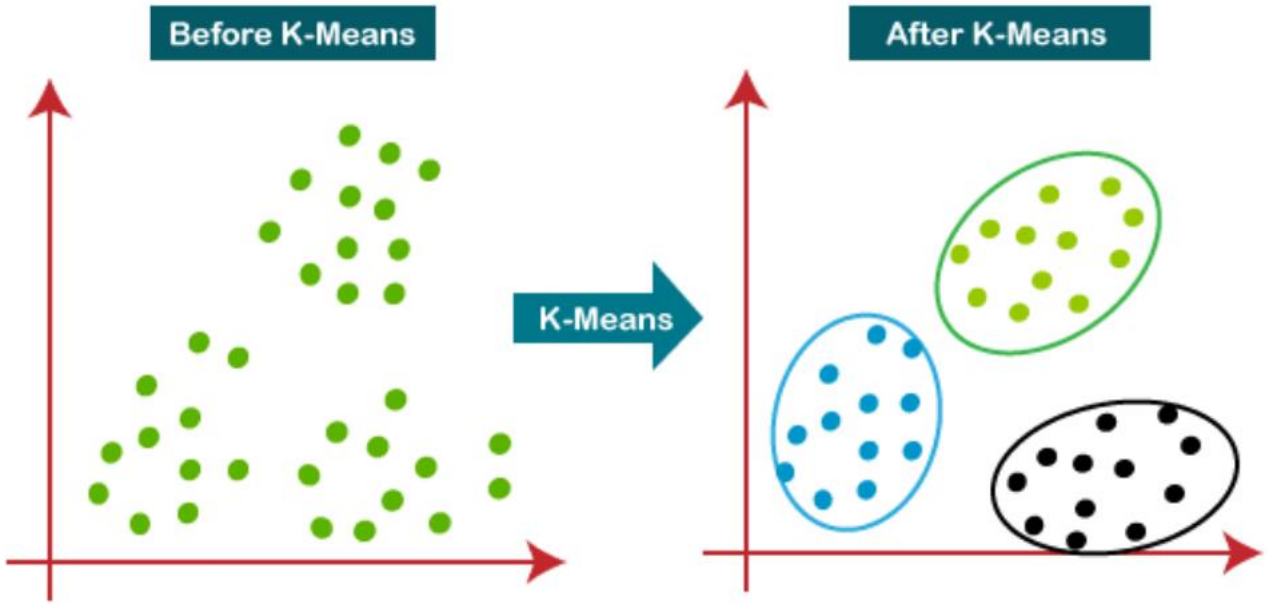
Keputusan tentang ukuran kemiripan dapat ditentukan melalui jumlah centroid masing-masing kelompok dan jaraknya.


Clustering
Partition based clustering
Hierarchical clustering
Density based clustering
K-Means
Fuzzy C-Means
Dsb…
Agglomerative
Divisive
DBSCAN

K-Means

- Salah satu algoritma clustering yang bersifat iteratif yang mencoba untuk mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster) di mana setiap titik data hanya dimiliki oleh satu kelompok

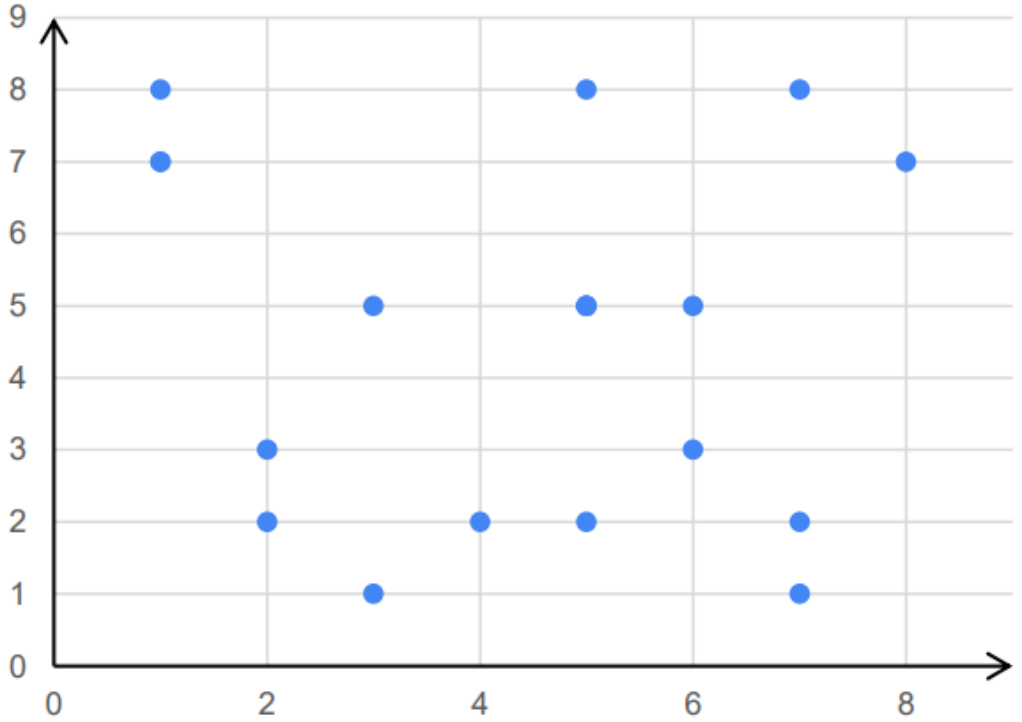
- K-Means mencoba membuat titik data intra-cluster semirip mungkin sambil dengan titik data yang lain pada satu cluster



Langkah Kerja K-Means

1. Memilih jumlah cluster awal (K) yang ingin dibuat

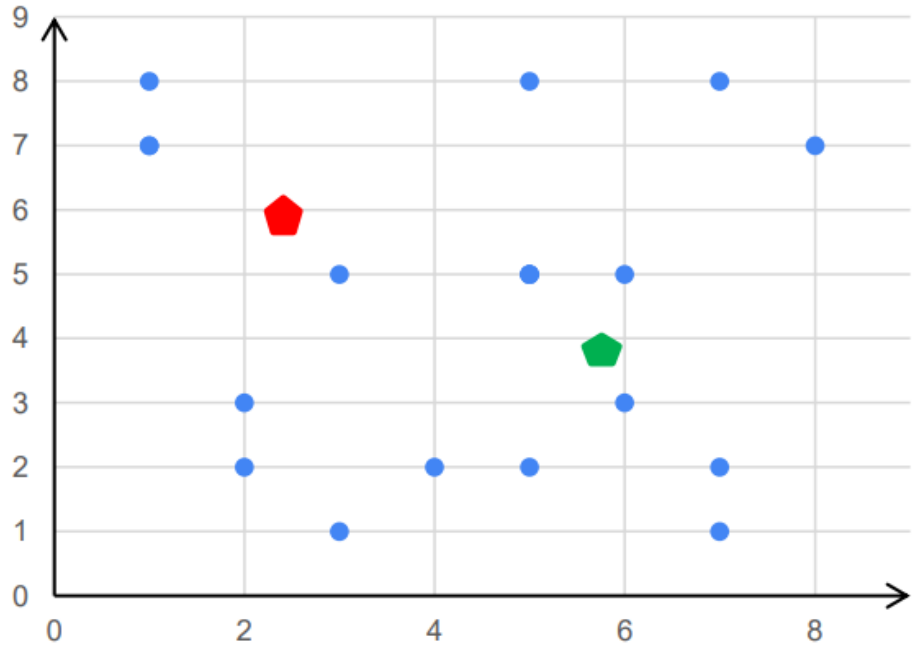
Sebagai contoh terdapat data 2 dimensi seperti yang ditampilkan dalam grafik, langkah pertama adalah memilih jumlah kluster. Misal kita pilih untuk membaginya ke dalam 2 kluster.



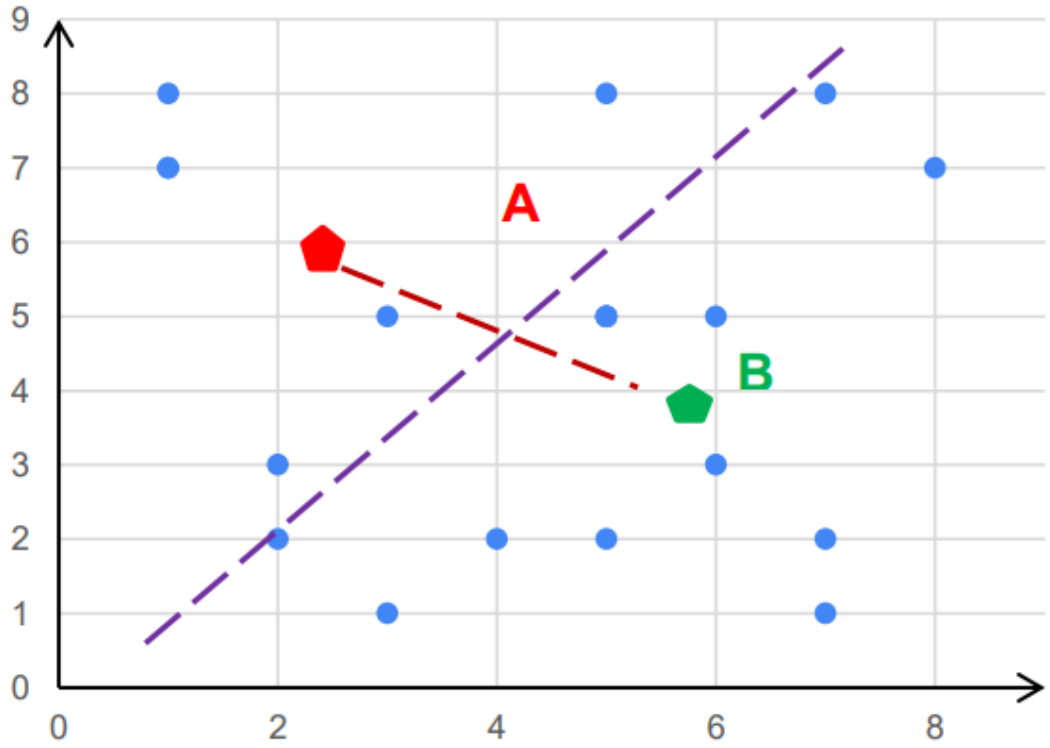
Langkah Kerja K-Means

2. Memilih titik secara random sebanyak K buah, di mana titik ini akan menjadi pusat (centroid) dari masing-masing kelompok (clusters).

Langkah kedua adalah menentukan titik pusatnya. Dalam satu klaster terdapat satu titik pusat atau yang disebut dengan centroid. Penentuan awal posisi titik pusat ini bebas, karena nantinya algoritma K-Means akan merubah posisi tiap titik hingga dicapai solusi paling optimal. Pada Gambar 2, titik merah mewakili pusat dari kluster 1, dan biru untuk kluster 2. Dengan demikian, maka masingmasing data point akan memilih titik pusat (centroid) yang paling dekat. Jika sudah dipilih maka data point tersebut akan menjadi bagian dari klusternya



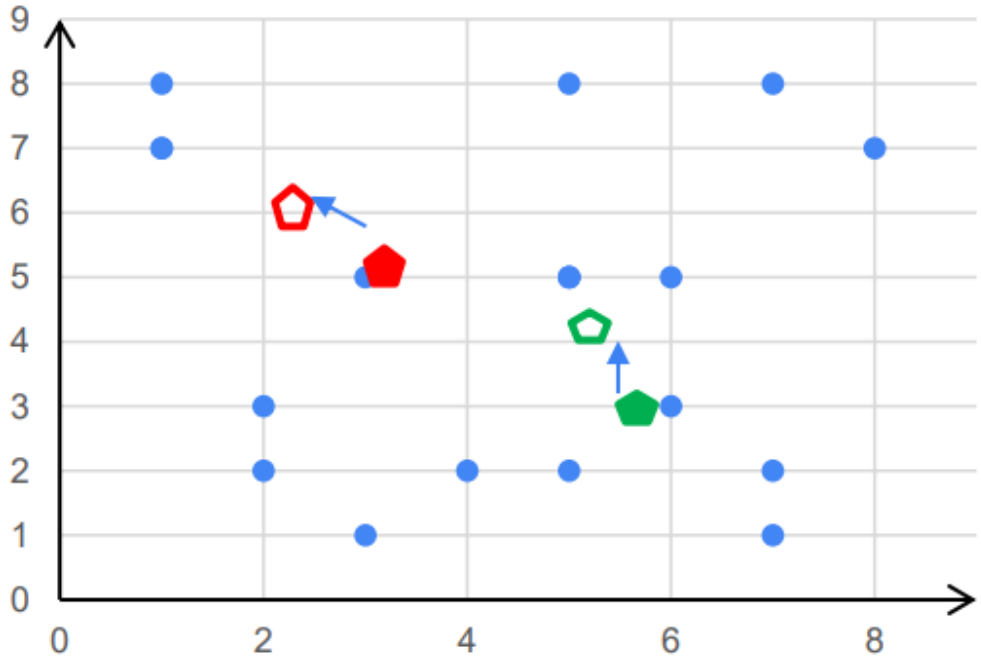
Langkah Kerja K-Means



3. Dari dataset yang kita miliki, buat dataset yang terdekat dengan titik centroid sebagai bagian dari cluster tersebut. Sehingga secara total akan terbentuk clusters sebanyak K buah.

Berdasarkan pengelompokan pada langkah ke dua, maka setiap titik data saat telah tergabung dalam salah satu kluster. Titik data yang diwakili dengan symbol 'x' berwarna merah masuk ke kluster 1, dan symbol 'x' berwarna biru masuk ke kluster 2, seperti pada berikut:

Langkah Kerja K-Means



4. Lakukan kalkulasi, dan tempatkan pusat centroid yang baru untuk setiap cluster-nya. Langkah ini dilakukan untuk menemukan centroid yang paling tepat untuk maisng-masing klaster.

Langkah ke-empat adalah melakukan penghitungan sesuai algoritma K-Means, yaitu mencari posisi titik pusat yang paling sesuai untuk sestiap klasternya berdasarkan penghitungan jarak terdekat. Penghitungan jarak masing-masing titik data ke pusat klaster dapat menggunakan metode Euclidean distance.

Langkah Kerja K-Means

5. Dari dataset yang kita miliki ambil titik centroid terdekat, sehingga dataset tadi menjadi bagian dari cluster tersebut. Jika masih ada data yang berubah kelompok (pindah cluster), kembali ke langkah 4. Jika tidak, maka cluster yang terbentuk sudah baik.

Langkah terakhir dari algoritma K-Means adalah melakukan pengecekan pada titik pusat yang telah ditentukan sebelumnya. Pilih titik pusat terdekat, dan masuk ke dalam kluster tersebut. Jika masih ada perpindahan kluster, kembali ke langkah 4. Algoritma K-Means akan terus mencari titik pusatnya, sampai pembagian datasetnya optimum dan posisi titik pusat tidak berubah lagi

Euclidean Distance

Euclidean distance adalah perhitungan jarak dari 2 buah titik dalam Euclidean space. Euclidean space diperkenalkan oleh Euclid, seorang matematikawan dari Yunani. untuk mempelajari hubungan antara sudut dan jarak. Euclidean ini berkaitan dengan Teorema Phytagoras dan biasanya diterapkan pada 1, 2 dan 3 dimensi

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Contoh:

Terdapat titik x1 bernilai 10 dan x2 bernilai -5. Maka kurangkan x1-x2 = 10-(-5) = 15. Lalu 15 kuadratkan sehingga memperoleh hasil 225. Langkah terakhir adalah diakarkan 2, sebagaimana rumusnya, sehingga kembali ke hasil 15

Optimasi K-Means

Terdapat salah satu **faktor krusial** baik tidaknya metode ini adalah saat **menentukan jumlah klusternya** (nilai K). Karena hasil pengemlompokan akan menghasilkan analisa yang berbeda untuk jumlah klaster yang berbeda juga. Jika terlalu sedikit K (misal 2), maka pembagian kluster menjadi cepat, namun mungkin ada informasi tersembunyi yang tidak terungkap. Jika K=8, maka terlalu banyak kluster. Mungkin akan terlalu sulit untuk membuat analisa atau memilih dukungan keputusan dari hasil cluster.

Thank You

- Ledy Elsera Astrianty
- ledyelsera@gmail.com