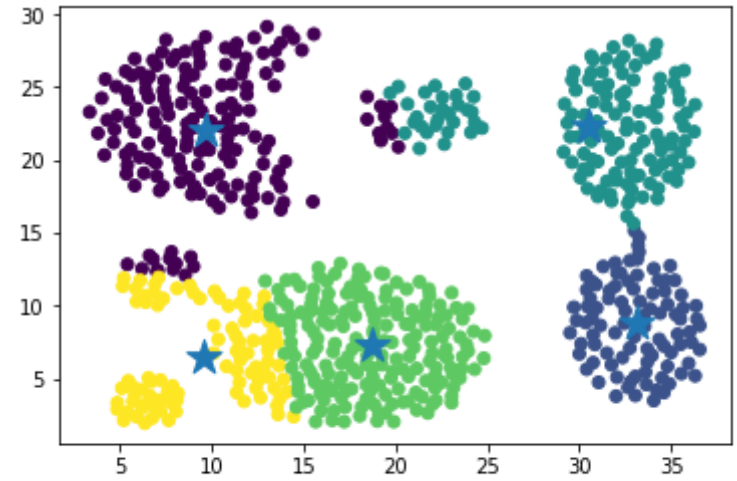
Laporan K-Means Machine Learning

Kartini Nurfalah
IF-39-01
1301154577

K-means merupakan bagian dari ***Partitional Clustering,*** dimana setiap data di *assign* dalam satu cluster.

1. Data biasa direpresentasikan dalam ruang Euclid |R^p → (x1, x2, x3, ..., xn)
2. Inputan dalam K-means ada 2 yaitu :
   a. Data
   b. Jumlah k
3. Output yang dihasilkan dari algoritma K-means juga ada 2, yaitu :
   a. Informasi data xi anggota cluster cj
   b. Nilai centroid (banyak sesuai dengan jumlah k)
4. Algoritma K-means :
   i. pilih k data sebagai inisial centroid (bisa dilakukan dengan generate random dari data dan tidak)
   ii. repeat
   iii. from k (bentuklah k) dengan cara assign setiap data dengan centroid terdekat
   iv. update centroid dari setip cluster
   v. until centroid tidak berubah (bisa diasumsikan berhenti ketika perubahan = 10^-9)
   a. Line iii.
   Assign setiap data ke centroid terdekat dengan cara cek jarak tiap data ke semua centroid
   Cek jarak dengan euclidian distance = || xi - cj ||2
   b. Line iv.
   Cara update centroid yaitu dengan mencari nilai mean dari setiap cluster
   cj = 1/|cj| * SUM(xi) | i E cj
5. Nilai k dalam inisialisasi awal K-means tidak ada ketentuannya, nilai k optional sesuai kebutuhan kita. Minimal k=2 maksimum k= jumlah data
6. Untuk mengecek bagus atau tidaknya hasil clustering yang kita buat bisa dilakukan dengan menggunakan perhitungan Sum Square Euclidian (SSE)
   ~~SSE = Sum(SUM(||cj - xi||22)~~
7. Nilai SSE di K-means dijamin sama atu turun tidak pernah naik dari iterasi sebelumnya, **tetapi** K-means tidak menjamin centroid tidak best of the best artinya nilai centroid bisa local minimum bukan global optimum.
   Solusi : Untuk mendapatkan hasil global maksimum maka bisa jalankan K-means berkali kali, pilih nilai SSE yang paling kecil.
8. Perhitungan Euclid digunakan untuk mencari disimilarity → semakin tidak mirip maka nilai yang dihasilkan semakin besar
9. Space running time complexity : O(N) → Big O (N)
10. Mungkinkah suatu centroid tidak memiliki anggota?
    a. Jika Centroid diambil dari data, maka minimal ada 1 anggota yaitu centroid itu sendiri
    b. Jika centroid di random bukan dari data maka awal iterasi mungkin tidak memiliki anggota bahkan hingga iterasi berikutnya. Jika hasil akhir masih tetap tidak memiliki anggota maka hilangkan centroid (salah satu k) yang tidak memiliki anggota.

11. Pada K-means kita menggunakan clustering centroid base, dimana setiap data yang ada di cluster itu akan lebih dekat dengan pusat(center) cluster itu dari pada ke cluster lain.
12. Train Phase
    SSE digunakan untuk mencari nilai k terbaik karena k-means tidak menjamin hasil yang global minimum tetapi hanya menjamin nilai SSE turun.




```
SSE :  2937.9685129943873
```

1. Test Phase