Algoritma K-Means

Muh.Nurtanzis Sutoyo
mr.iyes@yahoo.co.id

K-Means merupakan algoritma yang umum digunakan untuk clustering dokumen. Prinsip utama K-Means adalah menyusun k prototype atau pusat massa (centroid) dari sekumpulan data berdimensi. Sebelum diterapkan proses algoritma K-means, dokumen akan di preprocessing terlebih dahulu. Kemudian dokumen direpresentasikan sebagai vektor yang memiliki term dengan nilai tertentu.

Algoritma k-means merupakan algoritma yang membutuhkan parameter input sebanyak k dan membagi sekumpulan n objek kedalam k cluster sehingga tingkat kemiripan antar anggota dalam satu cluster tinggi sedangkan tingkat kemiripan dengan anggota pada cluster lain sangat rendah. Kemiripan anggota terhadap cluster diukur dengan kedekatan objek terhadap nilai mean pada cluster atau dapat disebut sebagai centroid cluster.

- Konsep dasar dari K-Means adalah pencarian pusat cluster secara iteratif.
- Pusat cluster ditetapkan berdasarkan jarak setiap data ke pusat cluster.
- Proses clustering dimulai dengan mengidentifikasi data yang akan dicluster, x_{ij} (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan dicluster dan m adalah jumlah variabel.
- Pada awal iterasi, pusat setiap cluster ditetapkan secara bebas (sembarang), c_{kj} (k=1,...,K; j=1,...,m).
- Kemudian dihitung jarak antara setiap data dengan setiap pusat cluster.
- Untuk melakukan penghitungan jarak data ke-i (X_i) pada pusat cluster ke-k (C_k), diberi nama (d_{ik}), dapat digunakan formula Euclidean, yaitu:

$$d_{ij} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj})^2} \quad (1)$$

- Suatu data akan menjadi anggota dari cluster ke-J apabila jarak data tersebut ke pusat cluster ke-J bernilai paling kecil jika dibandingkan dengan jarak ke pusat cluster lainnya.
- Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap cluster.

- Nilai pusat cluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data yang menjadi anggota pada cluster tersebut, dengan rumus:

$$c_{kj} = \frac{\Sigma_{h=1}^{p} y_{hj}}{p}; y_{hj} = x_{hj} \in clusterke - k \tag{2}$$

Algoritma K-Means

- Tentukan jumlah cluster (K), tetapkan pusat cluster sembarang.
- Hitung jarak setiap data ke pusat cluster.
- Kelompokkan data ke dalam cluster yang dengan jarak yang paling pendek.
- Hitung pusat cluster.
- Ulangi langkah 2 - 4 hingga sudah tidak ada lagi data yang berpindah ke cluster yang lain.

Contoh Kasus

Diketahui angka Kemampuan Dasar Berhitung dan angka Hasil Belajar (IPK) 15 orang mahasiswa seperti terlihat pada tabel. Mahasiswa-mahasiswa tersebut akan dikelompokkan berdasarkan Kemampuan Dasar Berhitung dan angka Hasil Belajar (IPK) menjadi tiga kelompok. Dimana proses pengelompokkan menggunakan metode K-Means.

No	Kemampuan Berhitung	Hasil Belajar
1	6.0	2.92
2	6.7	3.07
3	7.4	3.22
4	6.7	2.93
5	9.2	3.03
6	7.4	3.29
7	9.3	3.28
8	4.5	2.72
9	6.4	2.92
10	8.5	3.49
11	6.9	3.08
12	5.8	2.83
13	6.3	3.18
14	6.4	3.20
15	3.9	3.29

1. Misalkan kita akan pengelompokkan data tersebut menjadi 3 cluster, K = 3. Misalkan pusat cluster kita tetapkan sembarang, C_1 = (3.9, 2.7); C_2 = (6.6, 3.1); dan C_3 = (9.3, 3.5).

2. Hitung jarak setiap data terhadap setiap pusat cluster. Misalkan untuk menghitung jarak data pertama (No.1) dengan pusat cluster pertama adalah:

$$d_{11} = \sqrt{(6.0 - 3.9)^2 + (2.92 - 2.7)^2} = 2.109$$

Jarak data pertama (No.1) dengan pusat cluster kedua adalah:

$$d_{12} = \sqrt{(6.0 - 6.6)^2 + (2.92 - 3.1)^2} = 0.627$$

Jarak data pertama (No.1) dengan pusat cluster ketiga adalah:

$$d_{13} = \sqrt{(6.0 - 9.3)^2 + (2.92 - 3.5)^2} = 3.348$$

Hasil perhitungan jarak selengkapnya adalah:

No	Kemampuan Berhitung	Hasil Belajar	Jarak C_1	Jarak C_2	Jarak C_3
1	6.0	2.92	2.109	0.627	3.348
2	6.7	3.07	2.821	0.105	2.633
3	7.4	3.22	3.535	0.808	1.919
4	6.7	2.93	2.807	0.201	2.659
5	9.2	3.03	5.309	2.601	0.470
6	7.4	3.29	3.546	0.821	1.910
7	6.0	3.28	2.173	0.625	3.306
8	4.5	2.72	0.600	2.135	4.861
9	6.4	2.92	2.508	0.272	2.955
10	8.5	3.49	4.664	1.938	0.800
11	6.9	3.08	3.021	0.301	2.434
12	5.8	2.83	1.903	0.845	3.561
13	6.3	3.18	2.443	0.309	3.016
14	6.4	3.20	2.545	0.221	2.914
15	3.9	3.29	0.570	2.706	5.403

3. Suatu data akan menjadi anggota dari suatu cluster yang memiliki jarak terkecil dari pusat clusternya. Misalkan untuk data pertama, jarak terkecil diperoleh pada cluster kedua, sehingga data pertama akan menjadi anggota dari cluster kedua. Demikian juga untuk data kedua (No.2), jarak terkecil ada pada cluster kedua, maka data tersebut akan masuk pada cluster kedua. Posisi cluster selengkapnya adalah:

No	Kemampuan Berhitung	Hasil Belajar	Jarak C_1	Jarak C_2	Jarak C_3
1	6.0	2.92		*	
2	6.7	3.07		*	
3	7.4	3.22		*	
4	6.7	2.93		*	
5	9.2	3.03			*
6	7.4	3.29		*	
7	9.3	3.28		*	
8	4.5	2.72	*		
9	6.4	2.92		*	
10	8.5	3.49			*
11	6.9	3.08		*	
12	5.8	2.83		*	
13	6.3	3.18		*	
14	6.4	3.20		*	
15	3.9	3.29	*		

Berdasarkan hasil penggolongan tersebut, diperoleh anggota cluster pertama ada 2, cluster kedua ada 10, dan cluster ketiga ada 3.

4. Hitung pusat cluster baru. Untuk cluster pertama, ada 2 data yaitu data ke-8 dan data ke-15, sehingga:

$$C_{11} = \frac{4.5 + 3.9}{2} = 4.2;$$

$$C_{12} = \frac{2.72 + 2.72}{2} = 2.72$$

Untuk cluster kedua, ada 10 data yaitu data ke-1, data ke-2, data ke-3, data ke-4, data ke-6, data ke-9, data ke-11, data ke-12, data ke-13 dan data ke-14, sehingga:

$$C_{21} = \frac{6.0 + 6.7 + 7.4 + 6.7 + 7.4 + 6.4 + 6.9 + 5.8 + 6.3 + 6.4}{10} = 6.6;$$

$$C_{22} = \frac{2.92 + 3.07 + 3.22 + 2.93 + 3.29 + 2.92 + 3.08 + 2.83 + 3.18 + 3.20}{10} = 3.06$$

Untuk cluster ketiga, ada 3 data yaitu data ke-5, data ke-7, dan data ke-10 sehingga:

$$C_{31} = \frac{9.2 + 9.3 + 8.5}{3} = 9;$$

$$C_{32} = \frac{3.03 + 3.28 + 3.49}{3} = 3.26$$

5. Ulangi menghitung jarak setiap data terhadap setiap pusat cluster yang baru. Hasil perhitungan jarak selengkapnya terlihat pada tabel berikut.

No	Kemampuan Berhitung	Hasil Belajar	Jarak C_1	Jarak C_2	Jarak C_3
1	6.0	2.92	1.811	0.617	3.348
2	6.7	3.07	2.524	0.100	2.633
3	7.4	3.22	3.238	0.815	1.919
4	6.7	2.93	2.508	0.167	2.659
5	9.2	3.03	5.009	2.600	0.470
6	7.4	3.29	3.250	0.831	1.910
7	9.3	3.28	5.130	2.708	0.210
8	4.5	2.72	0.300	2.128	4.861
9	6.4	2.92	2.209	0.246	2.955
10	8.5	3.49	4.368	1.947	0,800
11	6.9	3.08	2.723	0.300	2.434
12	5.8	2.83	1.603	0.833	3.561
13	6.3	3.18	2.149	0.321	3.016
14	6.4	3.20	2.251	0.241	2.914
15	3.9	3.29	0.300	2.721	5.454

6. Posisi cluster selengkapnya terlihat pada tabel berikut.

No	Kemampuan Berhitung	Hasil Belajar	Jarak C_1	Jarak C_2	Jarak C_3
1	6.0	2.92		*	
2	6.7	3.07		*	
3	7.4	3.22		*	
4	6.7	2.93		*	
5	9.2	3.03			*
6	7.4	3.29		*	
7	9.3	3.28			**
8	4.5	2.72	*		
9	6.4	2.92		*	
10	8.5	3.49			*
11	6.9	3.08		*	
12	5.8	2.83		*	
13	6.3	3.18		*	
14	6.4	3.20		*	
15	3.9	3.29	*		

Terlihat masih ada 1 data yang berubah posisi dari kondisi semula, yaitu data ke-7 . Sehingga perlu dihitung pusat cluster baru.

7. Hitung pusat cluster baru sebagaimana pada langkah ke-4, sehingga diperoleh:

$$C_{11} = 4.2; C_{12} = 2.72$$

$$C_{21} = 6.6; C_{22} = 3.06$$

$$C_{31} = 9; C_{32} = 3.26$$

8. Ulangi menghitung jarak setiap data terhadap setiap pusat cluster yang baru. Hasil perhitungan jarak selengkapnya terlihat pada tabel berikut.

No	Kemampuan Berhitung	Hasil Belajar	Jarak C_1	Jarak C_2	Jarak C_3
1	6.0	2.92	1.811	0.617	3.348
2	6.7	3.07	2.524	0.100	2.633
3	7.4	3.22	3.238	0.815	1.919
4	6.7	2.93	2.508	0.167	2.659
5	9.2	3.03	5.009	2.600	0.470
6	7.4	3.29	3.250	0.831	1.910
7	9.3	3.28	5.130	2.708	0.210
8	4.5	2.72	0.300	2.128	4.861
9	6.4	2.92	2.209	0.246	2.955
10	8.5	3.49	4.368	1.947	0,800
11	6.9	3.08	2.723	0.300	2.434
12	5.8	2.83	1.603	0.833	3.561
13	6.3	3.18	2.149	0.321	3.016
14	6.4	3.20	2.251	0.241	2.914
15	3.9	3.29	0.300	2.721	5.454

9. Posisi cluster selengkapnya terlihat pada tabel berikut.

No	Kemampuan Berhitung	Hasil Belajar	Jarak C_1	Jarak C_2	Jarak C_3
1	6.0	2.92		*	
2	6.7	3.07		*	
3	7.4	3.22		*	
4	6.7	2.93		*	
5	9.2	3.03			*
6	7.4	3.29		*	
7	9.3	3.28			*
8	4.5	2.72	*		
9	6.4	2.92		*	
10	8.5	3.49			*
11	6.9	3.08		*	
12	5.8	2.83		*	
13	6.3	3.18		*	
14	6.4	3.20		*	
15	3.9	3.29	*		

Terlihat bahwa posisi data sudah tidak mengalami perubahan, sehingga proses iterasi sudah dapat dihentikan. Jika hasil cluster ditampilkan dalam bentuk gambar akan terlihat seperti gambar berikut.

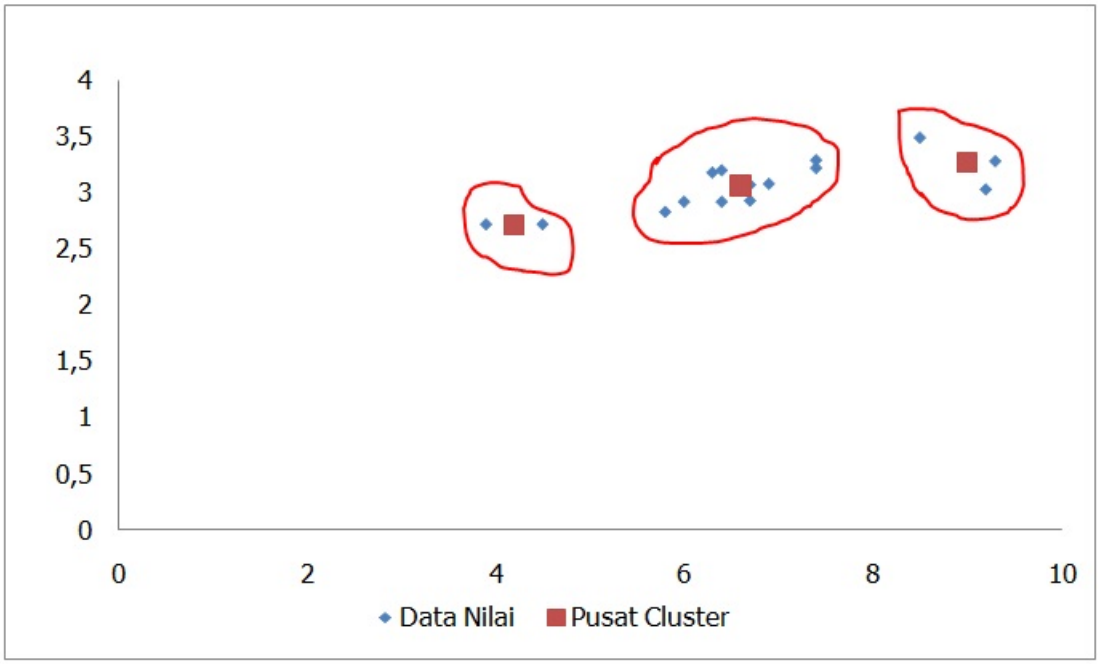

Figure 1: Hasil Cluster

Hasil akhir yang diperoleh adalah 3 cluster, dengan:

- Cluster pertama memiliki pusat (4.2, 2.72) yang dapat diartikan sebagai kelompok mahasiswa yang memiliki kemampuan dasar berhitung dan hasil belajar **rendah**. Ada 2 mahasiswa yang termasuk dalam kelompok ini.
- Cluster kedua memiliki pusat (6.6, 3.06) yang dapat diartikan sebagai kelompok mahasiswa yang memiliki kemampuan dasar berhitung dan hasil belajar **sedang**. Ada 10 mahasiswa yang termasuk dalam kelompok ini.
- Cluster ketiga memiliki pusat (9.0, 3.27) yang dapat diartikan sebagai kelompok mahasiswa yang memiliki kemampuan dasar berhitung dan hasil belajar **tinggi**. Ada 3 negara yang termasuk dalam kelompok ini.

References

Kusumadewi, Sri dkk. 2003. *Artificial Intelligence*. Graha Ilmu: Yogyakarta.

Russell S dan Norvig P. 2009. *Artificial Intelligence: A Modern Approach (3rd edition)*. New Jersey: Prentice Hall.

===Semoga Bermanfaat===$^{mr.iyes}$