

Untitled

Mohammad Rizka Fadhli

5/19/2022

SOAL DAN PEMBAHASAN

Soal I

Diberikan 10 buah titik data sebagai berikut:

Table 1: Data Soal I

titik	x	y
p1	4.0	5.2
p2	2.1	3.9
p3	3.4	3.1
p4	2.7	2.0
p5	0.8	4.1
p6	4.6	2.9
p7	4.3	1.2
p8	2.2	1.0
p9	4.1	4.1
p10	1.5	3.0

- Lakukan klusterisasi dari data tersebut dengan menggunakan algoritma *k-means* dengan jumlah partisi $K = 2$ sebanyak 10 kali.
- Tentukan sentroid awal (secara *random*) yang berbeda setiap melakukan klusterisasi.
- *Stopping criteria* untuk klusterisasi bisa ditentukan sendiri (tidak harus sampai tidak ada perubahan sentroid)

Pertanyaan

1. Tuliskan hasil akhir kluster yang didapat untuk setiap klusterisasi!
2. Hitung nilai *average SSE* untuk masing-masing hasil klusterisasi!
3. Hitung nilai *average Silhouette Coefficient* untuk masing-masing hasil klusterisasi!
4. Dari hasil *SSE* dan *Silhouette Coefficient*, menurut Anda, hasil klusterisasi mana yang memberikan hasil terbaik? Berikan alasannya!
5. Apakah algoritma *K-means* sudah memberikan hasil yang baik? Apa yang dapat dilakukan agar hasil klusterisasi lebih baik?

Soal II

Diberikan *confusion matrix* sebagai berikut:

Table 2: Data Soal II

cluster	entertainment	financial	foreign	metro	national	sports	Total
#1	1	1	0	11	4	676	693
#2	27	89	333	827	253	33	1562
#3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

Hitung nilai *entropy* dan *purity* untuk matriks tersebut! Berikan analisis untuk hasil yang didapat!