

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?

Pembahasan

Untuk melakukan *k-means clustering* ini, saya akan membuat algoritma sendiri dengan menggunakan 2 titik *random* dan akan dilakukan sebanyak 10 kali. Berikut ini adalah *flowchart* dari algoritma tersebut:

```
## QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to '/tmp/runtime-rstudio-user'
## TypeError: Attempting to change the setter of an unconfigurable property.
## TypeError: Attempting to change the setter of an unconfigurable property.
```

```

# program untuk membuat titik sentroid secara random
random_titik = function(){
  list(
    sentroid_1 = runif(2,0,6),
    sentroid_2 = runif(2,0,6)
  )
}

# program untuk menghitung jarak
jarak = function(x1,x2){
  sb_1 = (x1[1] - x2[1])^2
  sb_2 = (x1[2] - x2[2])^2
  sqrt(sb_1 + sb_2)
}

# program untuk menghitung sentroid baru
new_sentroid = function(data){
  hit =
    data %>%
    group_by(cluster_no) %>%
    summarise(x = mean(x),
              y = mean(y)) %>%
    ungroup()
  output = list(sentroid_1 = c(hit$x[1],hit$y[1]),
                 sentroid_2 = c(hit$x[2],hit$y[2]))
  return(output)
}

```

```

}

# program untuk menghitung selisih sentroid baru dengan sentroid lama
konvergen_yn = function(){
  part1 = sentroid_baru$sentroid_1 - sentroid_1
  part2 = sentroid_baru$sentroid_2 - sentroid_2
  sqrt(sum(part1^2) + sum(part2^2))
}

```

```

# iterasi pertama
random = random_titik()
sentroid_1 = random$sentroid_1
sentroid_2 = random$sentroid_2

df$jarak_sentroid1 = NA
df$jarak_sentroid2 = NA

konvergensi = 1000
iter = 1
while(konvergensi > 0.0000005){
  # hitung jarak terhadap sentroid
  for(i in 1:nrow(df)){
    titik = c(df$x[i],df$y[i])
    df$jarak_sentroid1[i] = jarak(titik,sentroid_1)
    df$jarak_sentroid2[i] = jarak(titik,sentroid_2)
  }
  # memasukkan masing-masing titik ke cluster terdekat
  df =
    df %>%
    mutate(cluster_no = ifelse(jarak_sentroid1 < jarak_sentroid2,1,2))
  # menghitung sentroid baru
  sentroid_baru = new_sentroid(df)
  # menghitung konvergensi
  konvergensi = konvergen_yn()
  # assign ke dalam nama sentroid
  sentroid_1 = sentroid_baru$sentroid_1
  sentroid_2 = sentroid_baru$sentroid_2

  iter = iter + 1
}

```

```
df
```

```
##      titik   x   y jarak_sentroid1 jarak_sentroid2 cluster_no
## 1      p1 4.0 5.2      3.6490581      1.5582219         2
## 2      p2 2.1 3.9      2.1772976      1.0667969         2
## 3      p3 3.4 3.1      1.4884976      0.8173467         2
## 4      p4 2.7 2.0      0.2015564      1.9402892         1
## 5      p5 0.8 4.1      2.9674273      2.3765638         2
## 6      p6 4.6 2.9      2.2171209      1.7382143         2
## 7      p7 4.3 1.2      1.7322312      2.9128547         1
## 8      p8 2.2 1.0      0.9303897      3.0410616         1
## 9      p9 4.1 4.1      2.7056654      0.9581522         2
## 10     p10 1.5 3.0      1.6794716      1.8862809         1

```

```
centroid_1
```

```
## [1] 2.675 1.800
```

```
centroid_2
```

```
## [1] 3.166667 3.883333
```

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

Pertanyaan

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

Pembahasan

Entropi untuk masing-masing cluster dihitung sebagai berikut:

$$\begin{aligned}\text{Entropy 1} &= -\frac{1}{693} \log_2\left(\frac{1}{693}\right) - \frac{1}{693} \log_2\left(\frac{1}{693}\right) \\ &\quad - 0 - \frac{11}{693} \log_2\left(\frac{11}{693}\right) \\ &\quad - \frac{4}{693} \log_2\left(\frac{4}{693}\right) - \frac{676}{693} \log_2\left(\frac{676}{693}\right) \\ &= 0.200\end{aligned}$$

$$\begin{aligned}\text{Entropy 2} &= -\frac{27}{1562} \log_2\left(\frac{27}{1562}\right) - \frac{89}{1562} \log_2\left(\frac{89}{1562}\right) \\ &\quad - \frac{333}{1562} \log_2\left(\frac{333}{1562}\right) - \frac{827}{1562} \log_2\left(\frac{827}{1562}\right) \\ &\quad - \frac{253}{1562} \log_2\left(\frac{253}{1562}\right) - \frac{33}{1562} \log_2\left(\frac{33}{1562}\right) \\ &= 1.841\end{aligned}$$

$$\begin{aligned}\text{Entropy 3} &= -\frac{326}{949} \log_2\left(\frac{326}{949}\right) - \frac{465}{949} \log_2\left(\frac{465}{949}\right) \\ &\quad - \frac{8}{949} \log_2\left(\frac{8}{949}\right) - \frac{105}{949} \log_2\left(\frac{105}{949}\right) \\ &\quad - \frac{16}{949} \log_2\left(\frac{16}{949}\right) - \frac{29}{949} \log_2\left(\frac{29}{949}\right) \\ &= 1.696\end{aligned}$$

Sedangkan untuk *purity* dihitung dengan cara:

$$\begin{aligned}\text{Purity 1} &= \frac{676}{693} = 0.975 \\ \text{Purity 2} &= \frac{827}{1562} = 0.529 \\ \text{Purity 3} &= \frac{465}{949} = 0.490\end{aligned}$$

Total entropy dihitung sebagai berikut:

$$\text{Total entropy} = \frac{693 \times 0.200 + 1562 \times 1.841 + 949 \times 0.490}{3204} = 0.614$$

Total purity dihitung sebagai berikut:

$$\text{Total purity} = \frac{693 \times 0.975 + 1562 \times 0.529 + 949 \times 1.696}{3204} = 1.443$$

Berikut jika disajikan dalam bentuk tabel:

Table 3: Hasil Perhitungan Entropy dan Purity

cluster	entertainment	financial	foreign	metro	national	sports	Total	Entropy	Purity
#1	1	1	0	11	4	676	693	0.200	0.975
#2	27	89	333	827	253	33	1562	1.841	0.529
#3	326	465	8	105	16	29	949	1.696	0.490
Total	354	555	341	943	273	738	3204	0.614	1.443

Dari tabel di atas, kita bisa dapatkan informasi sebagai berikut:

Cluster #1 memiliki *purity* yang sangat tinggi dan *entropy* terendah. Artinya, cluster ini berhasil mengelompokkan data yang *unique* karakteristiknya (berasal dari satu atribut dominan). Berbeda dengan *cluster #2* dan *#3* yang tidak memiliki satu atribut yang dominan. Tapi secara keseluruhan, *cluster* yang dihasilkan sudah bisa memisahkan data menjadi 3 kelompok dengan karakteristik yang berbeda-beda.