PERTEMUAN 6

TEKNIK CLUSTERING MENGGUNAKAN R

TUJUAN PRAKTIKUM

Mahasiswa dapat melakukan menggunakan teknik *clustering* dengan menggunakan R

TEORI PENUNJANG

1. K-Means Clustering

K-Means merupakan salah satu metode pengelompokan data nonhierarki yang mempartisi data yang ada kedalam bentuk dua atau lebih kelompok. Metode ini mempartisi data kedalam kelompok sehingga data yang berkarakteristik sama dimasukkan kedalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan kedalam kelompok yang lain.

Pada R untuk melakukan pengklasteran dengan menggunakan k-means dapat dilakukan dilakukan dengan menggunakan fungsi kmeans (data, *k*), di mana:

- data : data yang akan di *clustering*
- *k* : jumlah kluster yang ingin dibentuk

dengan menggunakan R:

- Simpanlah data iris menjadi iris2
- Hilangkan atribut Species pada iris2
- Lakukanlah pengklasteran k-means dengan k=3, dan simpan hasilnya pada kmeans.result

```
> iris2 <- iris
> iris2$Species <- NULL
> (kmeans.result <- kmeans(iris2, 3))
```

- Maka akan muncul hasil berikut:

```
K-means clustering with 3 clusters of sizes 38, 50, 62

Cluster means:
  Sepal.Length Sepal.Width Petal.Length
1     6.850000    3.073684     5.742105
2     5.006000    3.428000     1.462000
3     5.901613    2.748387     4.393548
  Petal.Width
1    2.071053
2    0.246000
3    1.433871

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [22] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [43] 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3
 [64] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3
 [85] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1
[106] 1 3 1 1 1 1 1 1 3 3 1 1 1 1 3 1 3 1 3 1 1
[127] 3 3 1 1 1 1 1 3 1 1 1 1 3 1 1 1 3 1 1 1 3
[148] 1 1 3

Within cluster sum of squares by cluster:
[1] 23.87947 15.15100 39.82097
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"
[4] "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

1
2
3
4
5

Keterangan:

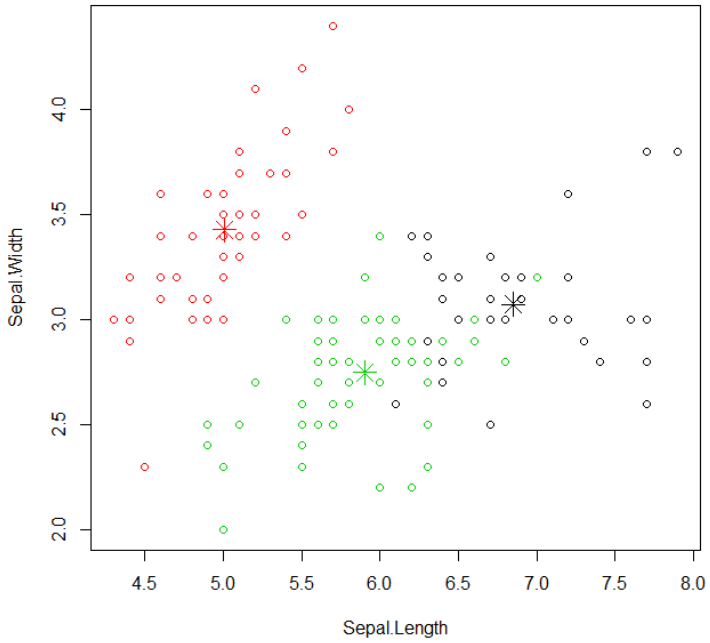
1: Jumlah data pada setiap kluster
2: Hasil rata-rata setiap atribut pada masing – masing kluster
3: Clustering vector yang menunjukkan cluster dari masing – masing data
4: Akurasi hasil kluster
5: Variabel yang terbentuk dari hasil kluster

- Bandingkanlah hasil pengklasteran kmeans dengan atribut Species pada data iris

```
> table(iris$Species, kmeans.result$cluster)
```

- Untuk melihat hasil plot antara kluster dengan pusat kluster dapat dilakukan dengan fungsi berikut:

```
> plot(iris2[c("Sepal.Length", "Sepal.Width")], col =
kmeans.result$cluster)
> points(kmeans.result$centers[,c("Sepal.Length",
"Sepal.Width")], col = 1:3,pch = 8, cex=2)
```



- Plot tersebut merupakan sebaran cluster berdasarkan nilai atribut Sepal Length dan Sepal Width, pada gambar tersebut tanda bintang menunjukkan pusat cluster untuk setiap kelas.

2. K-Medoids Clustering

Algoritma k-medoids hampir sama dengan algoritma k-means, di mana keduanya sama-sama mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok berdasarkan pendekatan partisi. Perbedaanya adalah jika pada k-means cluster diwakili dengan pusat dari cluster, sedangkan pada k-medoids, cluster diwakili oleh objek terdekat dari pusat cluster.

Pada R untuk melakukan pengklasteran dengan menggunakan k-means dapat dilakukan dilakukan dengan menggunakan fungsi pamk(data) dan pam(data, *k*). Perbedaan dari kedua

fungsi ini terletak dari jumlah cluster (*k*), di mana pada pamk(data) jumlah *k* tidak dapat diset sesuai dengan keinginan, sedangkan pada pam(data, *k*) nilai *k* yang diinginkan langsung dapat diset pada fungsi tersebut.

a. Menggunakan fungsi **pamk**

- Gunakanlah data iris2 yang telah dihilangkan atribut Speciesnya
- Lakukanlah pengklasteran dengan menggunakan fungsi pamk() dan simpan hasilnya pada pamk.result

```
> library(fpc)
> pamk.result <- pamk(iris2)
```

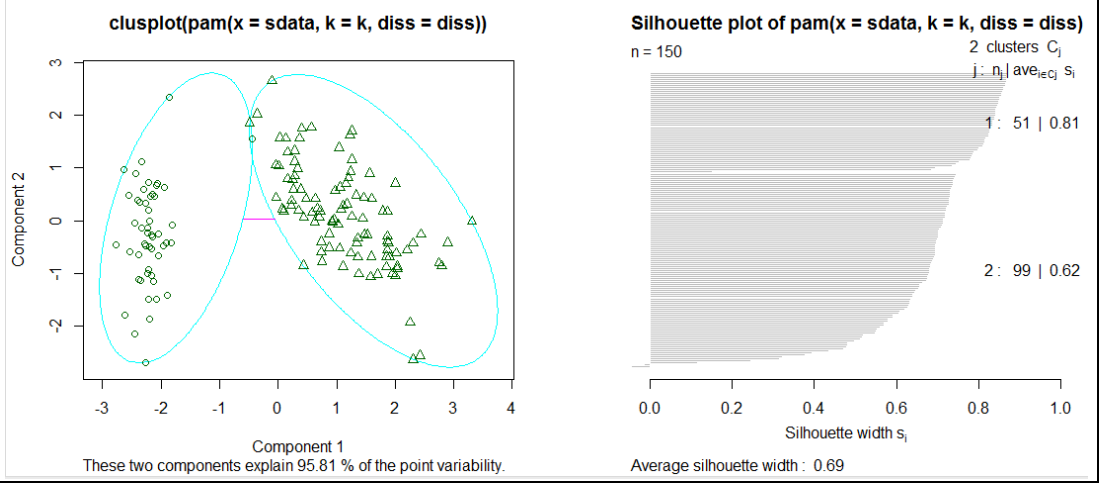
- Untuk menampilkan jumlah kluster yang terbentuk, bandingkanlah hasil pengklasteran k-medoids pamk() dengan atribut Species pada data iris

```
> pamk.result$nc
> table(pamk.result$pamobject$clustering, iris$Species)
   setosa versicolor virginica
1      50          1         0
2       0         49        50
```

- Dari confusion matrix tersebut terlihat bahwa pamk() menghasilkan hanya dua kelompok dengan satu kelompok didominasi oleh kelas setosa dan kelompok lainnya merupakan campuran dari kelas versicolor dan verginica.

- Untuk melihat plot dari hasil pamk.result gunakan sintax berikut:

```
> layout(matrix(c(1,2),1,2)) # 2 graphs per page
> plot(pamk.result$pamobject)
> layout(matrix(1)) # change back to one graph per page
```



- Gambar sebelah kiri adalah clusplot 2-dimensi (*clustering* plot) dari dua kelompok dan garis menunjukkan jarak antara *cluster*. Gambar sebelah kanan menunjukkan *silhouettes*. Dalam *silhouettes*, nilai S_i yang besar (hampir 1) menunjukkan bahwa pengamatan yang sesuai terkelompok sangat baik, nilai S_i yang mendekati 0 berarti pengamatan terletak di antara dua kelompok, dan pengamatan dengan S_i negatif

mungkin ditempatkan di cluster yang salah. Karena S_i rata-rata adalah masing-masing 0,81 dan 0,62 di silhouettes di atas, mengidentifikasikan bahwa dua kelompok ini baik.

b. Menggunakan fungsi **pam**

- Lakukanlah pengklasteran dengan menggunakan fungsi pam() dan simpan hasilnya pada pam.result

```
> pam.result <- pam(iris2, 3)
```

- Bandingkanlah hasil pengklasteran k-medoids pam() dengan atribut Species pada data iris

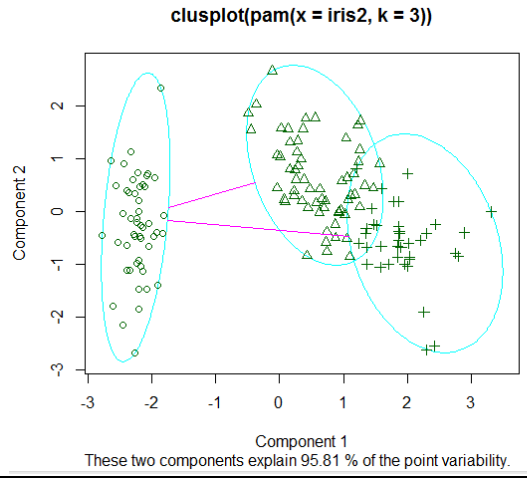
```
> table(pam.result$clustering, iris$Species)

   setosa versicolor virginica
1      50          0         0
2       0         48        14
3       0          2        36
```

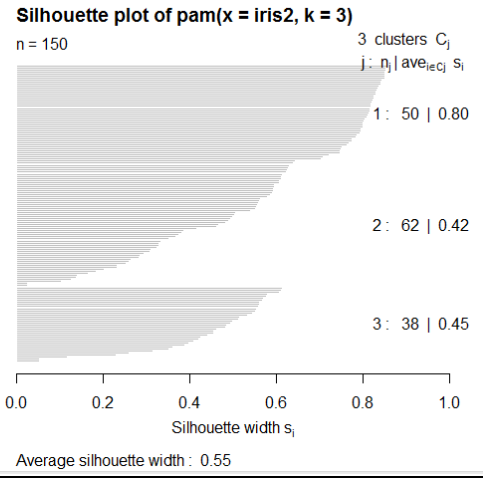
Dari hasil algoritma pam() diperoleh tiga klaster:

1) Kluster 1 adalah hanya teridiri atas jenis “setosa " dan dipisahkan mencolok dari dua lainnya.
2) Kluster 2 terdiri atas banyak “ versicolor" ditambah beberapa dari “ virginica ".
3) Mayoritas kluster 3 adalah “virginica" ditambah dengan dua dari “versicolor ".

- Untuk melihat plot dari hasil pam.result

```
> layout(matrix(c(1,2),1,2)) # 2 graphs per page
> plot(pam.result)
> layout(matrix(1)) # change back to one graph per page
```





3. Hierachical Clustering

Hierachical Clustering adalah algoritma klustering yang mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan.

Pada R untuk melakukan pengklasteran dengan menggunakan *hierachical clustering* dapat dilakukan dilakukan dengan menggunakan fungsi hclust(dist(data), method="*linkType*") di mana:

- data: data yang akan di *clustering.*
- *linkType:* untuk menentukan tipe penentuan pemilihan jarak yang dipakai.

Langkah-langkah hierarchical clustering menggunakan R:

- Ambil 40 data dari dataset iris, simpan pada irisSample dan hapus variabel Species

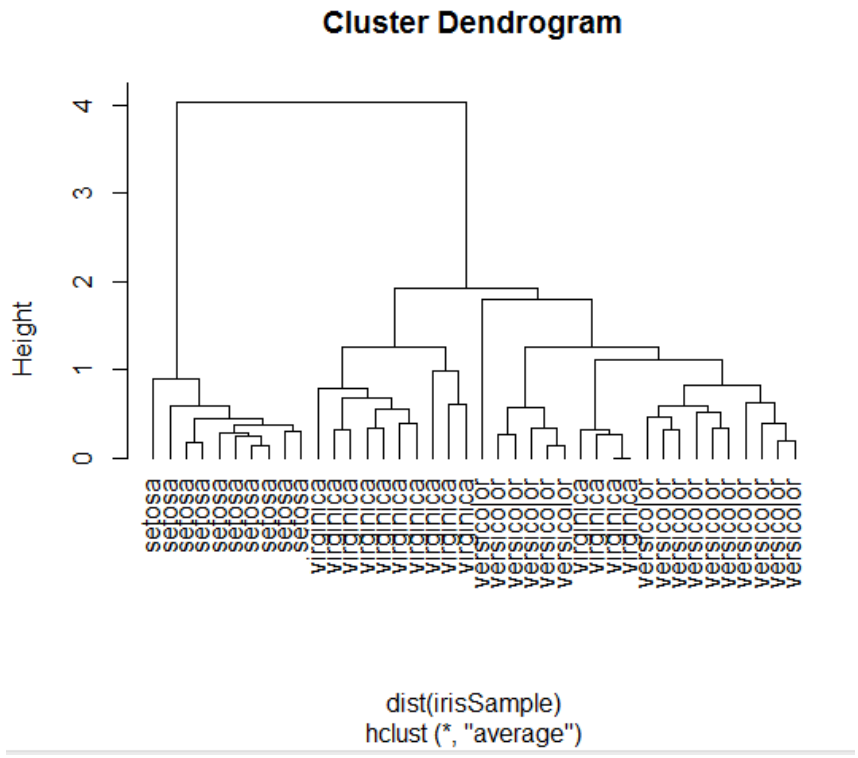
```
> idx <- sample(1:dim(iris)[1], 40)
> irisSample <- iris[idx,]
> irisSample$Species <- NULL
```

- Lakukanlah pengklasteran *hierachical clustering* dengan menggunakan metode *average linked* dan simpan hasilnya pada hc

```
> hc <- hclust(dist(irisSample), method="ave")
```

- Untuk melihat hasil *clustering plot* dengan pelabelan berdasarkan atribut Species pada data iris

```
> plot(hc, hang = -1, labels=iris$Species[idx])
```



- Dari plot hirarki tersebut dapat dilihat bahwa jenis setosa memiliki perbedaan yang cukup jauh disbanding jenis virginica dan versicolor. Ada beberapa jenis virginica yang mirip dengan jenis versicolor.

Catatan: Clustering merupakan metode unsupervised di mana atribut kelas tidak diketahui. Evaluasi cluster umumnya dilihat dari jarak intra dan inter cluster. Pada modul ini,

perbandingan antar atribut kelas dan hasil clustering hanya sebagai metode asesment, pada praktiknya di lapangan atribut kelas tidak akan diketahui.

LAPORAN PENDAHULUAN

1. Jenis *clustering* apa saja yang Anda ketahui?
2. Apa saja kelebihan dan kekurangan dari setiap jenis *clustering?*

MATERI PRAKTIKUM

1. K-Means Clustering pada R
2. K-Medoids Clustering pada R
3. Hierachical Clustering pada R

DAFTAR PUSTAKA

1. Yanchang Zhao. 2012. *R and Data Mining: Examples and Case Studies*. http://www.rdatamining.com/docs/RDataMining.pdf
2. Han, J. (2006) Data Mining: Concepts and Technique. [Internet]. [diunduh 2014 Mar 24]. Tersedia pada: http://www.cs.uiuc.edu/homes/hanj/bk2/slidesindex.htm