Departemen Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

PEMODELAN KLASIFIKASI

PERTEMUAN #1

PENGANTAR DAN K-MEANS

Bagus Sartono
bagusco@ipb.ac.id bagusco@gmail.com

DESKRIPSI MATA KULIAH

Deskripsi: Mata kuliah ini mendiskusikan beberapa algoritma dalam analisis data dan data mining untuk tujuan klasifikasi, yaitu menentukan kelas atau kelompok dari setiap amatan. Topik yang akan dibahas meliputi pendekatan un-supervised dan supervised, dengan penekanan lebih banyak pada yang kedua. Algoritma un-supervised yang akan dibahas adalah k-means, sedangkan algoritma supervised meliputi k-NN, regresi logistik, pohon klasifikasi, pengenalan support vector machine, dan naïve bayesian classifier. Juga akan didiskusikan pendekatan ensemble yaitu bagging, boosting, dan random forest. Tidak hanya algoritma yang akan dipelajari tetapi juga membahas proses evaluasi dan validasi model. Mata kuliah ini memiliki sks praktikum yang didalamnya akan mendiskusikan penggunaan software R.

PENGAJAR

Dr. Bagus Sartono

Dr. Anang Kurnia

SILABUS

- Pengantar
- Un-supervised Classification: k-means
- Pengantar mengenai supervised classification
- k-NN
- Penilaian kebaikan dugaan klasifikasi
- Analisis Diskriminan
- Regresi Logistik
- Pohon Klasifikasi
- Bagging
- Boosting
- Random Forest dan Rotation Forest
- Pengenalan Support Vector Machine
- Pengenalan Naïve Bayes Classifier

PENDEKATAN PEMBELAJARAN

- Integrasi antara ceramah teori dan praktek
- Memerlukan (dan menuntut) keaktifan mahasiswa
- Di setiap pertemuan mahasiswa membawa komputer/laptop, dan dosen akan menyiapkan data dan programnya
- Menggunakan R (dan SAS)

PENILAIAN

Komponen	Bobot
Keaktifan di Kelas	15%
Laporan Tugas #1	15%
Laporan Tugas #2	15%
Presentasi	10%
Laporan Tugas Akhir	25%
Kualitas Prediksi Tugas Akhir	20%

COURSE NOTE DAN DATA

http:….. akan diinfokan kembali kemudian

DATA YANG AKAN DIGUNAKAN

Data kredit bank di Jerman

Data kualitas wine buatan Portugal

Data tekstur

Pima Indians Diabetes Data Set

Lain-lain

DATA KREDIT BANK DI JERMAN

Ketika suatu bank memperoleh aplikasi pinjaman dari calon nasabah, maka selanjutnya bank akan membuat keputusan menyetujui atau tidak berdasarkan profil calon nasabah tersebut. Setiap calon nasabah akan dikategorikan menjadi dua yaitu:

- Good risk, yang berarti orang tersebut dianggap akan mampu membayar pinjaman dan selanjutnya aplikasi pinjaman itu disetujui
- Bad risk, yang berarti orang tersebut dianggap akan gagal membayar pinjaman dan selanjutnya aplikasi pinjaman itu ditolak

Pemodelan klasifikasi akan bergunanuntuk meminimumkan kerugian (dalam perspektif bank), yaitu dengan menentukan calon nasabah mana yang disetuji dan mana yang tidak. Manajer pinjaman biasanya akan mempertimbangkan kondisi demografi dan sosio-ekonomi calon nasabah sebelum membuat keputusan.

Data yang ada berisi informasi 1000 orang calon nasabah pinjaman yang terdiri atas 20 variabel.

DATA KUALITAS WINE

Data kedua ini berasal dari pengamatan terhadap sejumlah tipe wine merah (ada 1599 jenis) dan wine putih (ada 4898 jenis) yang semuanya produksi Portugal.

Harga dari wine sangat tergantung pada kualitas rasa yang bersifat abstrak. Biasanya ada sejumlah ahli rasa yang diminta mencicipi wine dan kemudian memberikan penilaian. Antar penilai bisa jadi sangat bervariasi opininya.

Kualitas wine juga tergantung pada beberapa hasil physicochemical tests yang dilakukan di laboratorium dengan memeriksa antara lain acidity, pH level, kandungan gula, serta kandungan beberapa senyawa kimia lain.

Akan sangat menarik kalau hasil pengukuran kualitas di laboratorium memiliki keterkaitan dengan penilaian subjektif oleh pakar rasa.

Data yang ada hanya berisi data penilaian wine putih, yang meliputi 12 karakteristik hasil uji laboratorium serta hasil penilaian ahli rasar pada skala 1 – 10, dengan 1 untuk kualitas rasa yang paling buruk dan 10 untuk rasa yang paling baik. Ada tiga penilai, dan nilai akhir diperoleh dari median ketiganya.

PIMA INDIANS DIABETES DATA SET

HTTP://ARCHIVE.ICS.UCI.EDU/ML/DATASETS/PIMA+INDIANS+DIABETES

Relevant Information: Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

Number of Instances: 768

Number of Attributes: 8 plus class

Attribute: (all numeric-valued)

- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

PENGANTAR

ANALISIS KLASIFIKASI

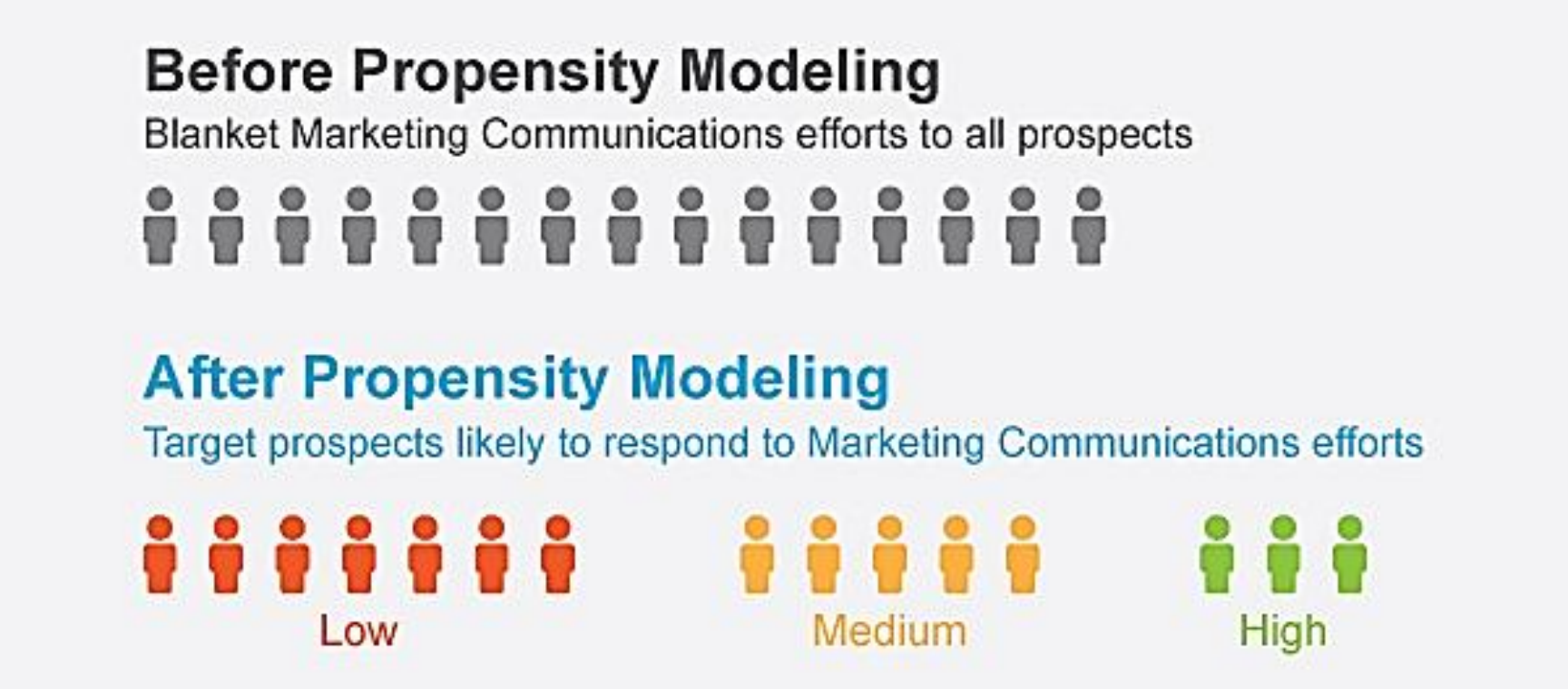
Tujuan analisis: menentukan keanggotaan grup/kelompok dari suatu individu

Tipe metode

- Unsupervised, tidak terdapat informasi mengenai kelompok/grup dari amatan pada data yang digunakan. Analisis dilakukan untuk menentukan keanggotaan grup dari amatan tersebut. Sering juga dikenal sebagai analisis gerombol (clustering, cluster analysis)

- Supervised, data memiliki informasi mengenai kelompok/grup sesungguhnya dari amatan. Analisis dilakukan untuk menentukan pembeda antar grup, dan aturan pembeda tersebut dapat dimanfaatkan untuk menentukan keanggotaan dari amatan lain yang tidak ada dalam data.

KEGUNAAN ANALISIS KLASIFIKASI




GEOGRAPHIC
DEMOGRAPHIC
MARKET SEGMENTATION
BEHAVIORAL
PSYCHOGRAPHIC

K-MEANS

Andaikan terdapat *n* buah amatan $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

Masing-masing amatan akan dikelompokkan ke dalam satu dari *k* buah kelompok. Besaran *k* umumnya jauh lebih kecil dibandingkan *n*.

Andaikan $\mathbf{c}_1, \ldots, \mathbf{c}_k$ adalah centroid dari *k* buah kelompok.

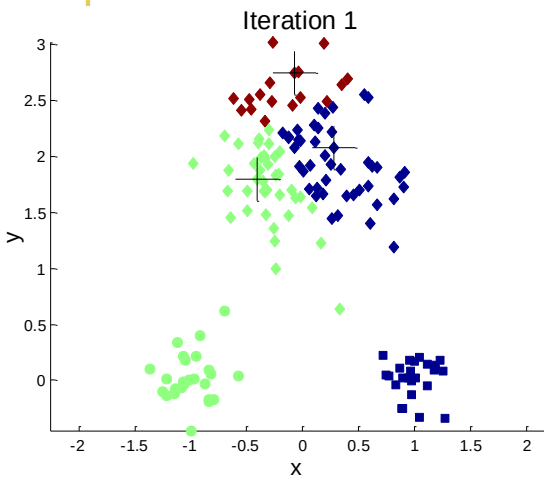
Intuisi dari pengelompokan adalah bahwa amatan akan dimasukkan ke kelompok *j* jika amatan tersebut memiliki jarak paling dekat dengan $\mathbf{c}_j$ dibandingkan dengan centroid lainnya.

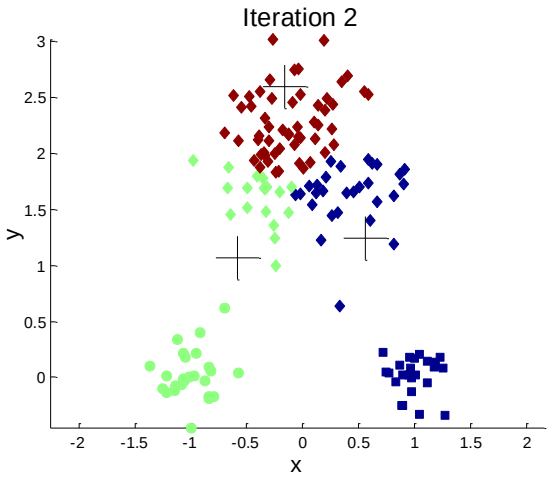
ALGORITMA

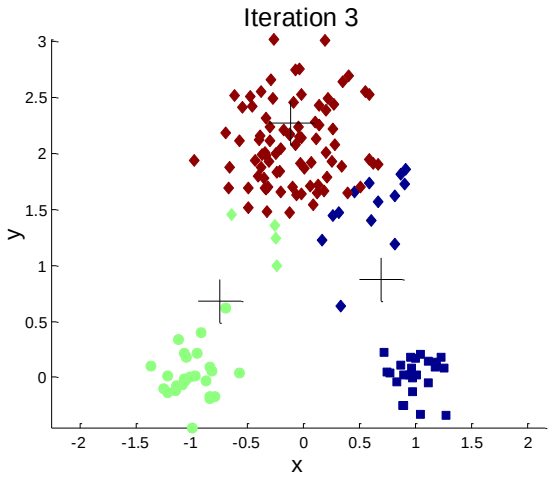
1. Tentukan secara acak $\mathbf{c}_1, \ldots, \mathbf{c}_k$
2. Hitung jarak dari setiap $\boldsymbol{x}_i$ ke $\mathbf{c}_j$
3. Masukkan amatan ke-*i* ke dalam kelompok ke-*j* jika d_{ij} adalah yang paling kecil dibandingkan $d_{ij'}$
4. Perbarui $\mathbf{c}_1, \ldots, \mathbf{c}_k$ dengan menghitung rata-rata dari semua $\boldsymbol{x}_i$ yang menjadi anggota kelompok masing-masing
5. Kembali ke tahap 2 sampai konvergen

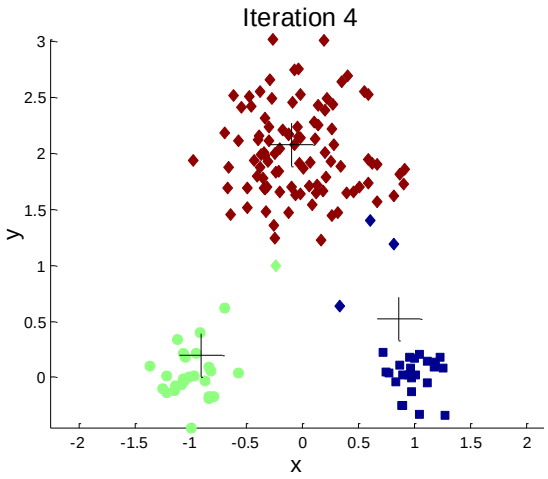
How do we decide when to stop?

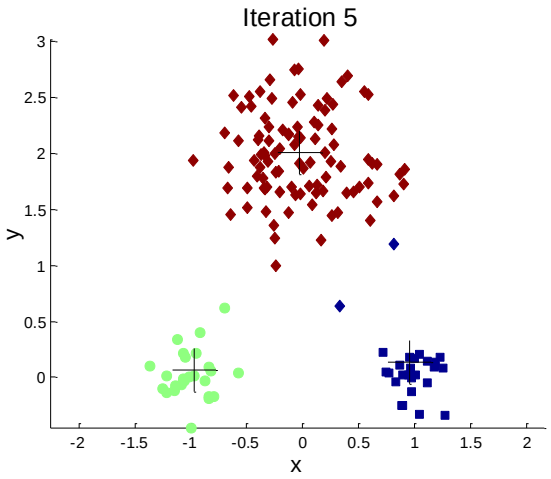
One criterion for stopping is if we observe the assignment functions in the two iterations are exactly the same. If the assignment function doesn't change anymore, then the centroid won't change either (and vice versa).

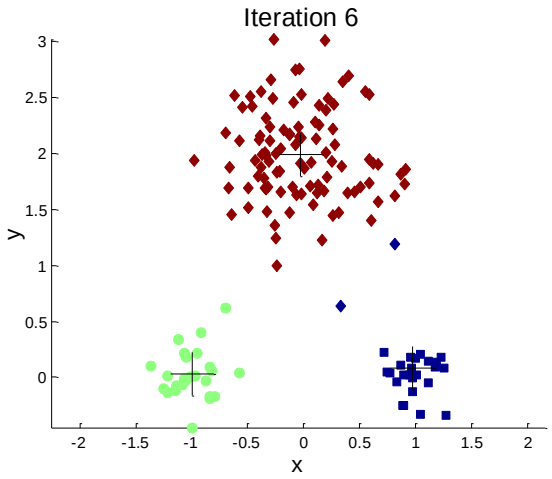

Iteration 1
x
y


Iteration 2
x
y


Iteration 3
x
y


Iteration 4
x
y


Iteration 5
x
y


Iteration 6
x
y

ILUSTRASI SEDERHANA

Misal ada 9 amatan yang masing-masing bernilai:

1, 2, 2, 3, 4, 4, 8, 8, 10

ingin dikelompokkan menjadi dua grup

Centroid awal c1 = 1 dan c2 = 2

Iterasi 1

- Keanggotaan : {1}, {2, 2, 3, 4, 4, 8, 8, 10}
- Centroid baru : c1 = 1, c2 = 5.125

Iterasi 2

- Keanggotaan : {1, 2, 2, 3}, {4, 4, 8, 8, 10}
- Centroid : c1 = 2, c2 = 6.8

Iterasi 3

- Keanggotaan : {1, 2, 2, 3, 4, 4}, {8, 8, 10}
- Centroid : c1 = 2.67 c2 = 8.67

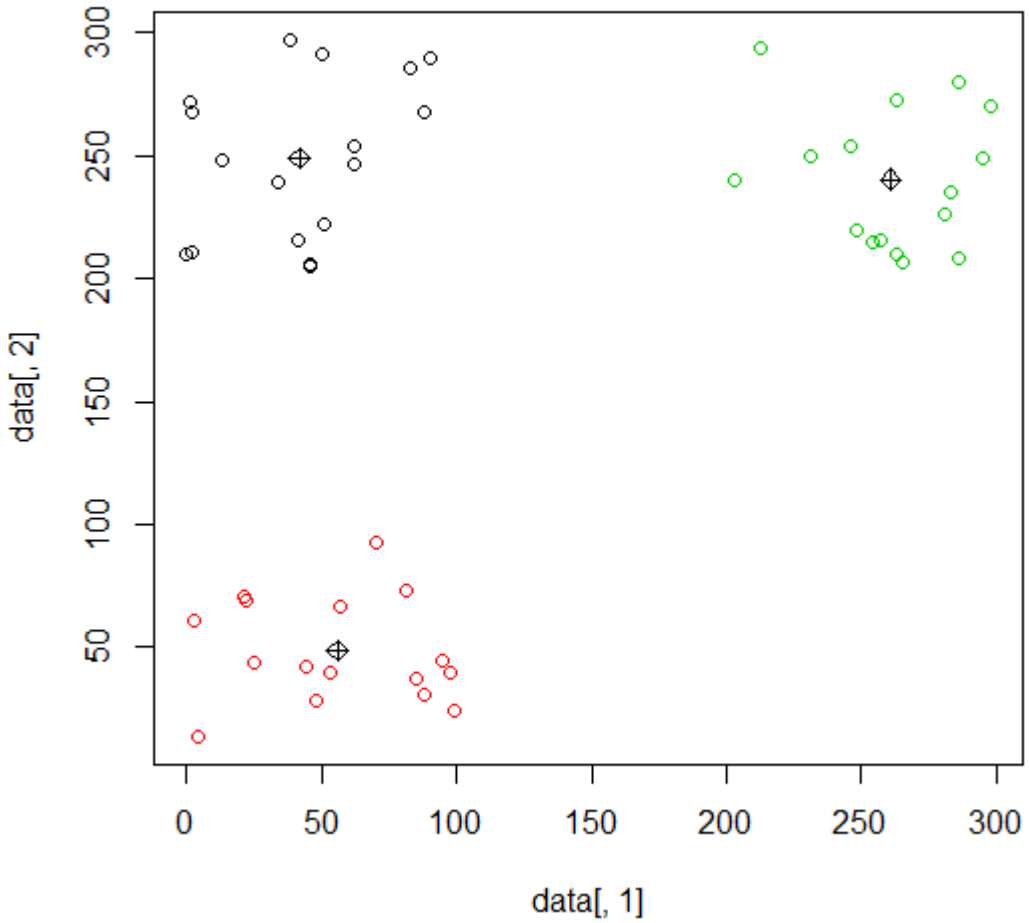
Iterasi 4

- Keanggotaan : {1, 2, 2, 3, 4, 4}, {8, 8, 10}
- Centroid : c1 = 2.67 c2 = 8.67

Stop

```
setwd ("D:/bagusco/Kuliah S2 --- Pemodelan Klasifikasi/data")
data <- read.csv("ilustrasikm.csv")

cluster <- kmeans(data, 3)
plot(data[,1], data[,2], col=cluster$cluster)
points(cluster$centers, pch=9)
```



Departemen Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

PEMODELAN KLASIFIKASI

PERTEMUAN #2

MASIH TENTANG K-MEANS

Bagus Sartono
bagusco@ipb.ac.id bagusco@gmail.com

Penentuan banyaknya cluster

Dua alternatif

- Ditentukan oleh peneliti
- Didasarkan pada data

Pada studi segmentasi, banyaknya cluster melambangkan banyaknya segmen

Preferable approach: *"let the data speak"*

- Didasarkan pada statistik tertentu
- Memungkinkan terjadinya perdebatan dan ketidakpastian

Ilustrasi situasi dimana banyaknya cluster ditentukan oleh peneliti

A retailer wants to identify several shopping profiles in order to activate new and targeted retail outlets

The budget only allows him to open three types of outlets

A partition into three clusters follows naturally, although it is not necessarily the optimal one.

Fixed number of clusters and (*k*-means) non hierarchical approach

Ilustrasi situasi dimana biarkan data yang menentukan banyaknya cluster

Clustering of shopping profiles is expected to detect a new market niche.

For market segmentation purposes, it is less advisable to constrain the analysis to a fixed number of clusters

- A hierarchical procedure allows to explore all potentially valid numbers of clusters
- For each of them there are some statistical diagnostics to pinpoint the best partition.
- What is needed is a *stopping rule* for the hierarchical algorithm, which determines the number of clusters at which the algorithm should stop.

Statistical tests are not always univocal, leaving some room to the researcher's experience and arbitrariness

Statistical rigidities should be balanced with the knowledge gained from and interpretability of the final classification.

Determining the optimal number of cluster from hierarchical methods

Graphical

- scree diagram → lihat loncatan jarak penggabungan yang paling besar

Statistical

- Within sum of squares
- Silhouette coefficient
- Pseudo F

Silhouette Coefficient

For each observation i, the *silhouette coef s(i)* is defined as follows:

Put a(i) = average dissimilarity between i and all other points of the cluster to which i belongs (if i is the *only* observation in its cluster, *s(i) := 0* without further calculations).

For all *other* clusters C, put *d(i,C)*= average dissimilarity of i to all observations of C.

The smallest of these *d(i,C)* is *b(i) := min*$_C$ *d(i,C)*, and can be seen as the dissimilarity between i and its "neighbor" cluster, i.e., the nearest one to which it does *not* belong.

Finally,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

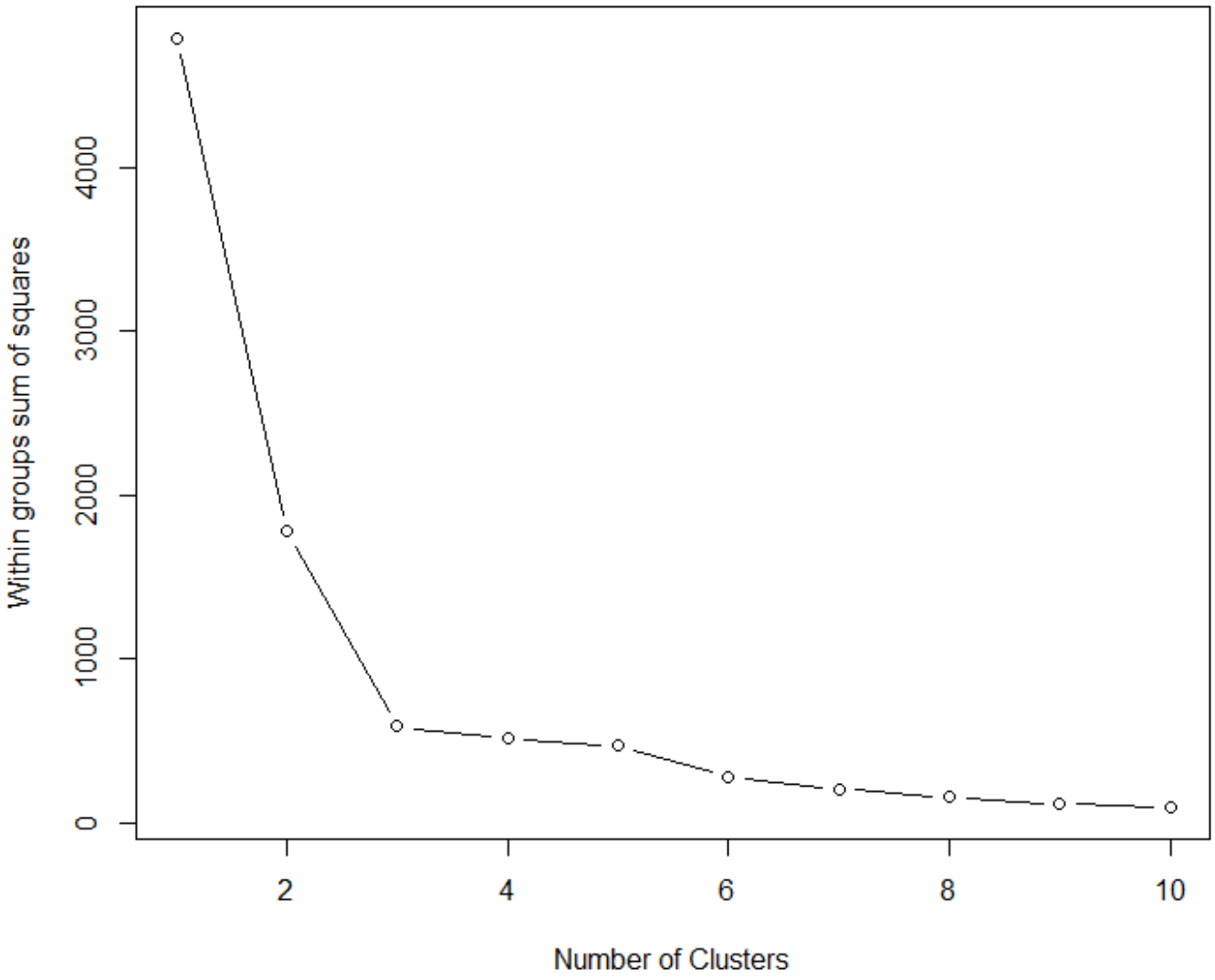
```
ilustrasi <- read.csv("D:/bagusco/bagusco/Tulisan BSO/Cluster
Analysis/ilustrasi2a.csv", header=T, sep=";")
head(ilustrasi)


hasilgerombol <- kmeans(ilustrasi, centers=3, iter.max =10)
hasilgerombol$cluster

hasilgerombol$tot.withinss

wssplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")}

wssplot(ilustrasi, nc=10)
```


Within groups sum of squares
0
1000
2000
3000
4000
2
4
6
8
10
Number of Clusters

```
library("cluster")
jarak <- as.matrix(dist(ilustrasi))

hasilgerombol <- kmeans(ilustrasi, centers=3, iter.max =10)
sil.3 <-
mean(silhouette(hasilgerombol$cluster,dmatrix=jarak)[,3])

hasilgerombol <- kmeans(ilustrasi, centers=4, iter.max =10)
sil.4 <-
mean(silhouette(hasilgerombol$cluster,dmatrix=jarak)[,3])

c(sil.3, sil.4)
```

```
> c(sil.3, sil.4)
[1] 0.6172319 0.4626237
```