



Departemen Statistika

Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

PEMODELAN KLASIFIKASI

PERTEMUAN #3

K-NEAREST NEIGHBOR

Bagus Sartono

bagusco@ipb.ac.id

bagusco@gmail.com

Supervised Classification

Data terdiri atas amatan-amatan yang berisi informasi mengenai:

- Keanggotaan Kelas/Grup
- Karakteristik amatan (sering disebut sebagai variabel, atribut, feature)

Informasi dari data digunakan untuk memperoleh “aturan” bagi penentuan keanggotaan kelas dari amatan lainnya nanti.

Supervised Classification

Terbagi atas:

- Metode yang berbasis model
- Metode yang tidak berbasis model

Metode yang berbasis model

- Model berupa fungsi matematis
- Model berupa aturan logika

Supervised Classification

Metode yang tidak berbasis model

- k-nearest neighbor

Metode yang berbasis model

- Regresi logistik
- Analisis diskriminan
- Classification tree
- SVM
- dll

K Nearest Neighbor

Nama lain:

- K-Nearest Neighbors
- Memory-Based Reasoning
- Example-Based Reasoning
- Instance-Based Learning
- Case-Based Reasoning
- Lazy Learning

KNN

Konsep dasar

- Menyimpan data training
- Mengklasifikasikan amatan baru berdasarkan kemiripan dengan amatan dalam data training
- Kelas yang dipilih adalah kelas dari amatan-amatan yang paling mirip (tetangga terdekatnya)

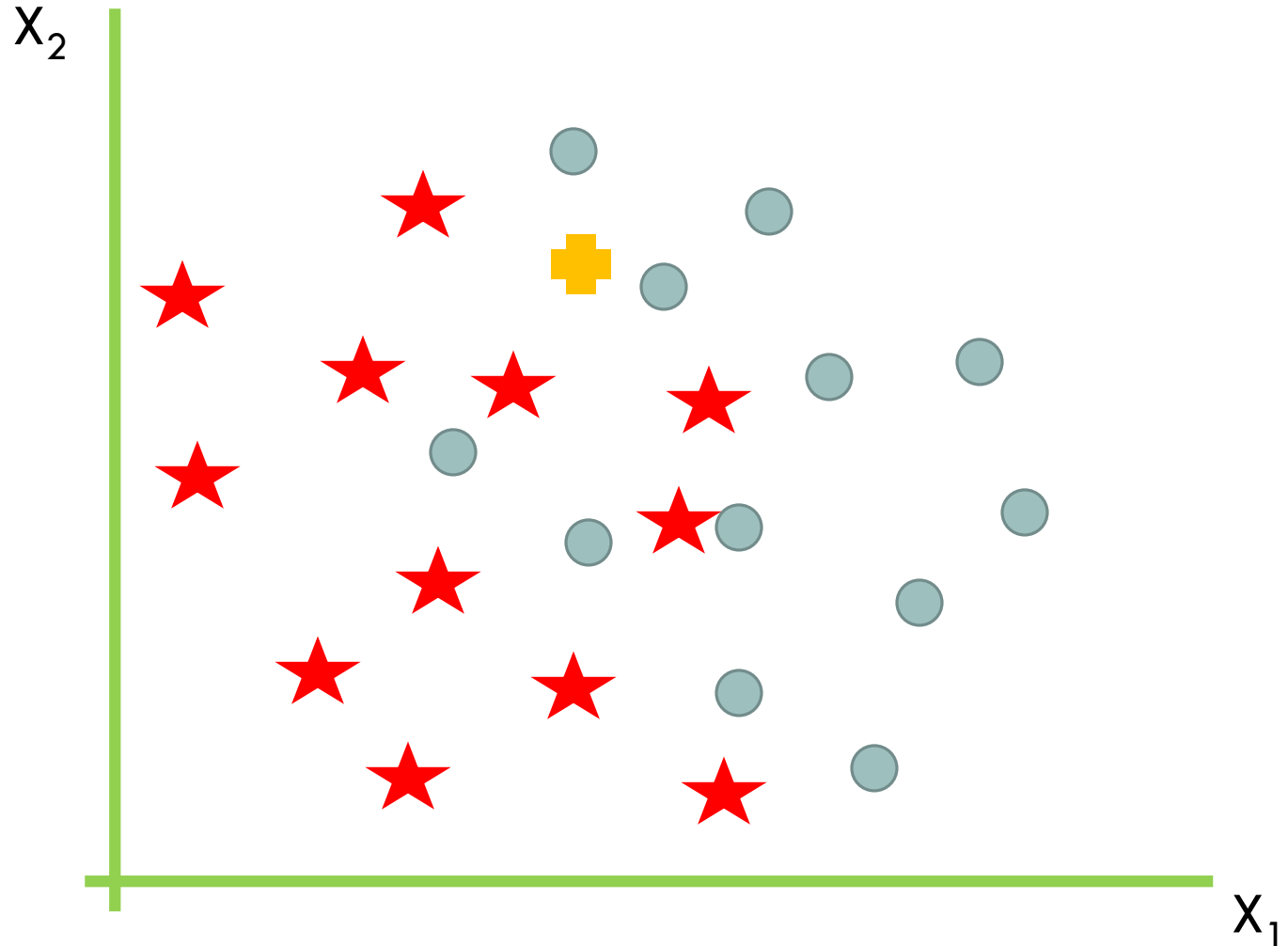
Ilustrasi: data training

X_2

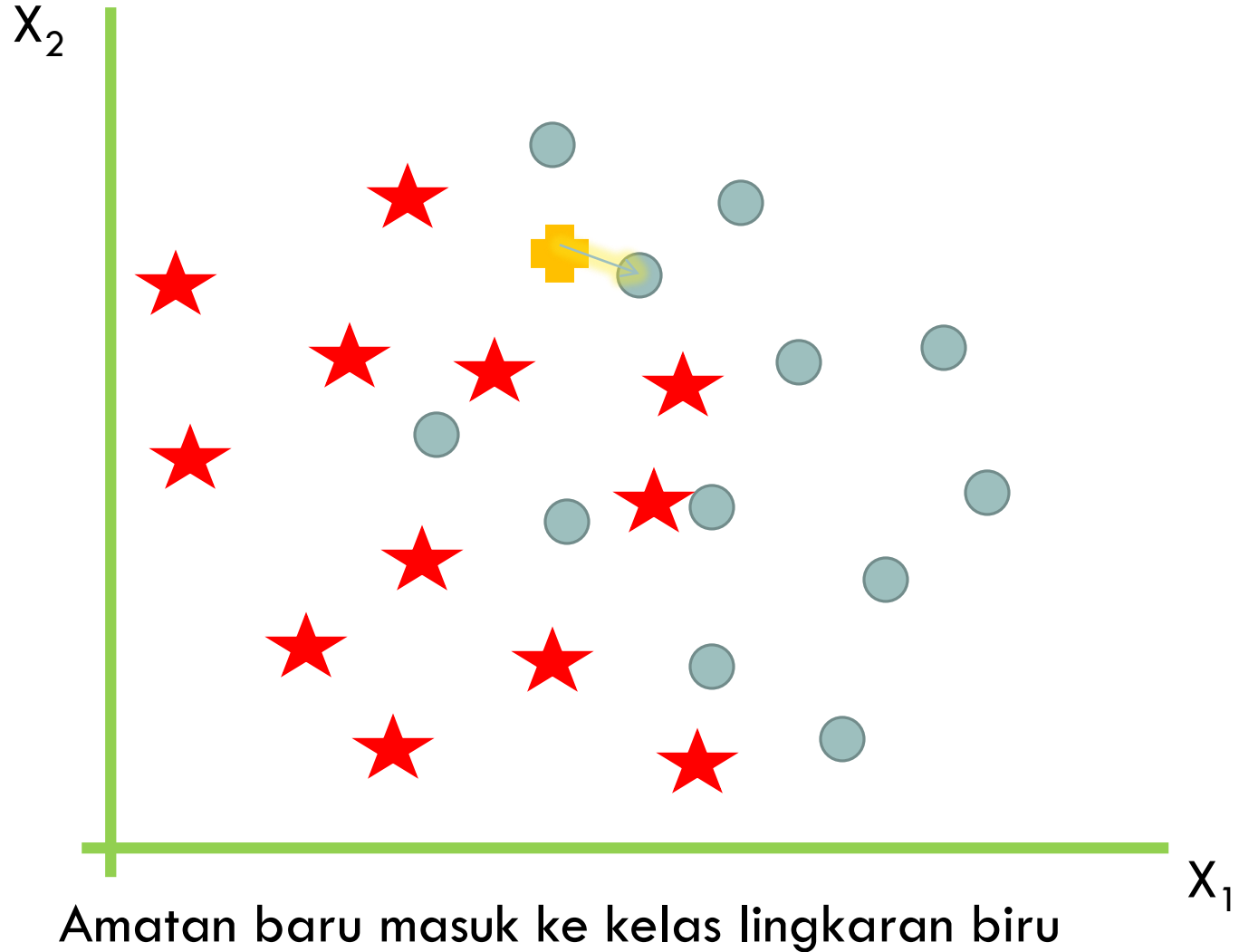


24 amatan

Ilustrasi: masuk kelas mana amatan baru + ini?

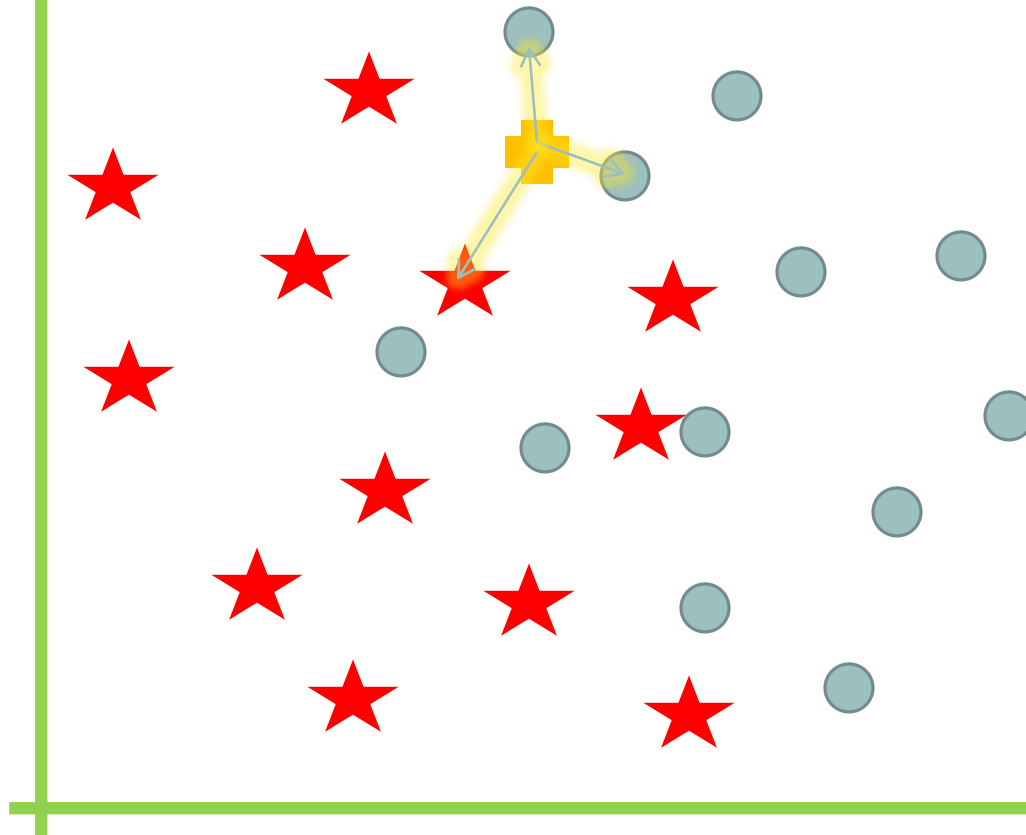


Ilustrasi: k-NN dengan $k = 1$



Ilustrasi: k-NN dengan $k = 3$

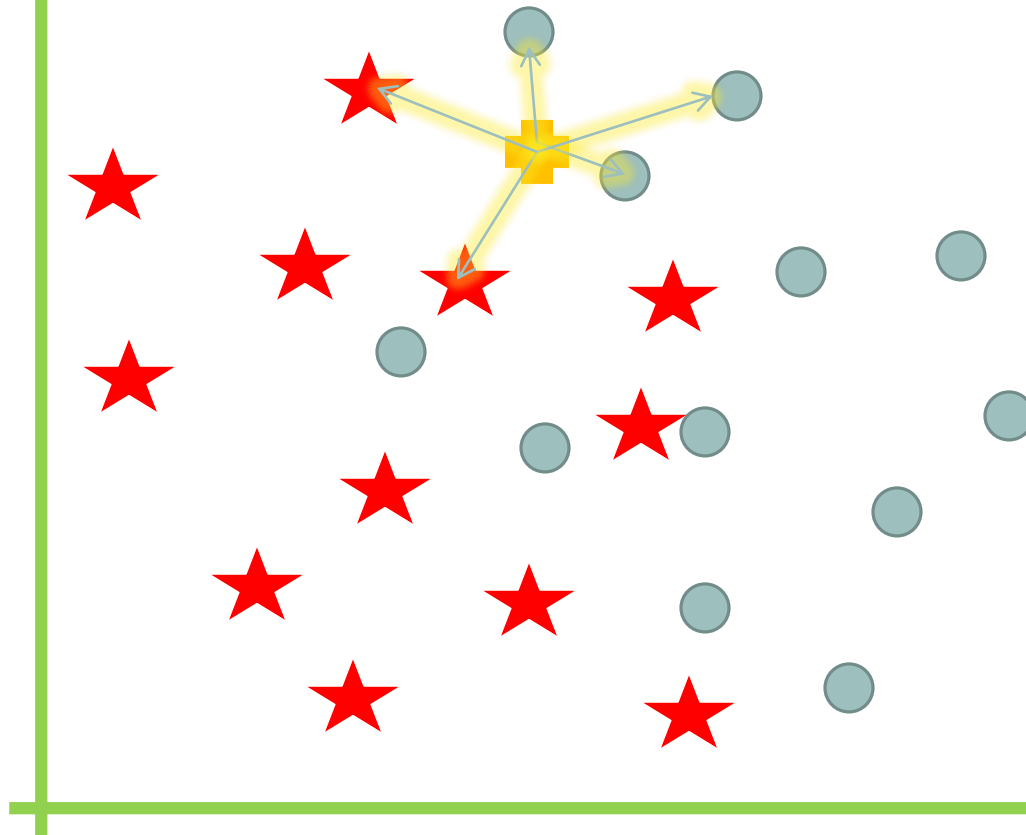
x_2



Amatan baru masuk ke kelas lingkaran biru, karena dari tiga tetangga, dua dari kelas tersebut

Ilustrasi: k-NN dengan $k = 5$

X_2



Amatan baru masuk ke kelas lingkaran biru, karena dari lima tetangga, tiga dari kelas tersebut

Perhatikan ilustrasi berikut

Age	Loan	Default	Distance
25	\$40,000	N	102000
35	\$60,000	N	82000
45	\$80,000	N	62000
20	\$20,000	N	122000
35	\$120,000	N	22000
52	\$18,000	N	124000
23	\$95,000	Y	47000
40	\$62,000	Y	80000
60	\$100,000	Y	42000
48	\$220,000	Y	78000
33	\$150,000	Y	8000
48	\$142,000	?	

Euclidean Distance

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Perhatikan ilustrasi berikut

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	0.61	?	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

Beberapa isu dalam analisis kNN

Bagaimana menghitung kemiripan antar amatan?

- Jarak apa yang digunakan? Euclid? Mahalanobis? Lainnya?
- Perlu pembakuan data? Penskalaan variabel?

Berapa banyak tetangga (k)?

- Saran:
 - Gunakan nilai ganjil
 - Lakukan validasi atau validasi silang

Ilustrasi: data

Data → ilustrasiknn.txt

Terdiri atas dua variabel x1 dan x2

Berisi data dari dua kelompok, yang diindikasikan oleh kolom 'class'

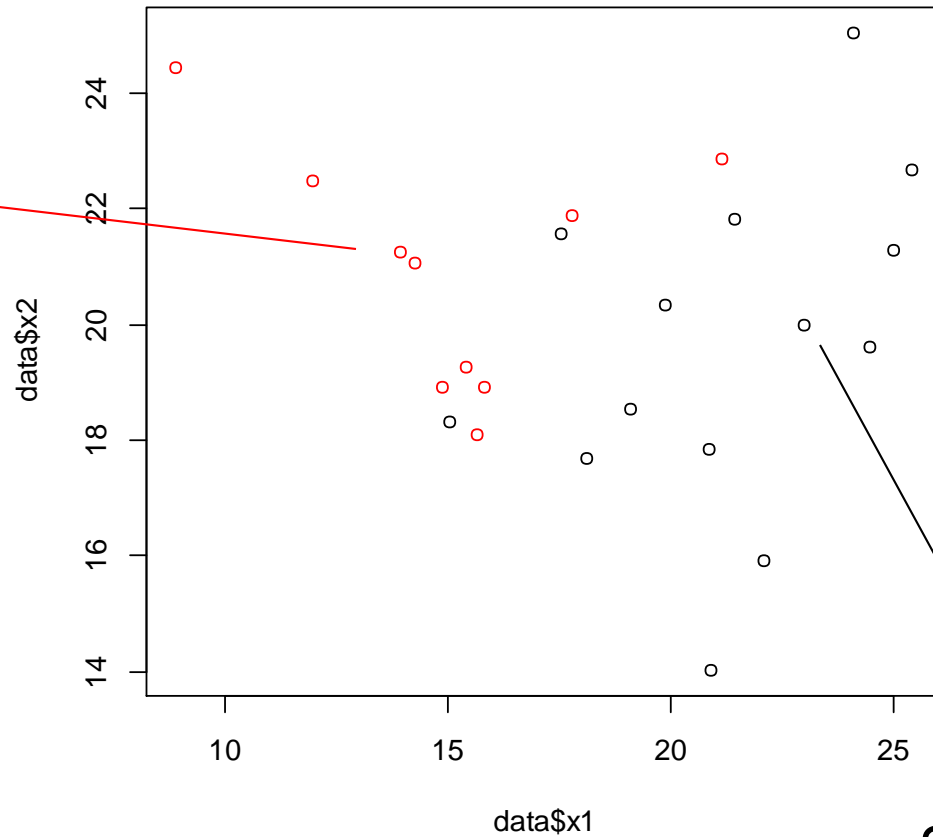
- class = 1, 14 amatan
- class = 2, 10 amatan

```
data <- read.table("D:/bagusco/Kuliah S2 --- Pemodelan  
Klasifikasi/ilustrasiknn.txt", header=TRUE)
```

Ilustrasi: plot tebaran data

```
plot(data$x1, data$x2, col=data$class)
```

class = 2

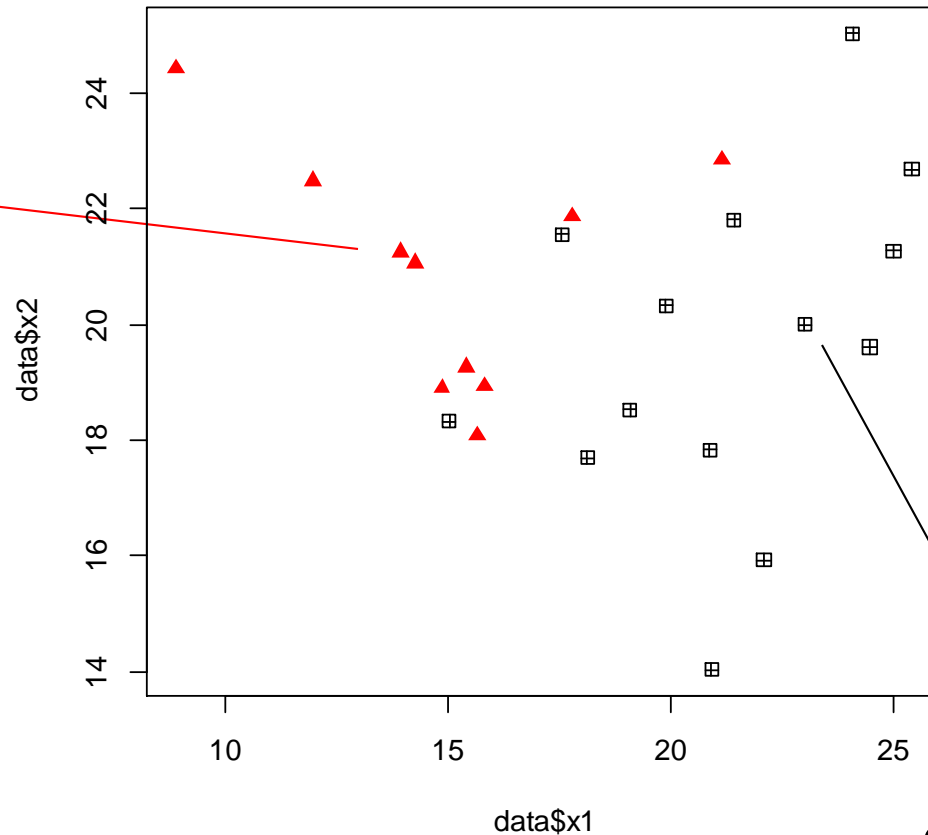


class = 1

Ilustrasi: plot tebaran data

```
plot(data$x1, data$x2, col=data$class,  
     pch=ifelse(data$class>1,17,12))
```

class = 2



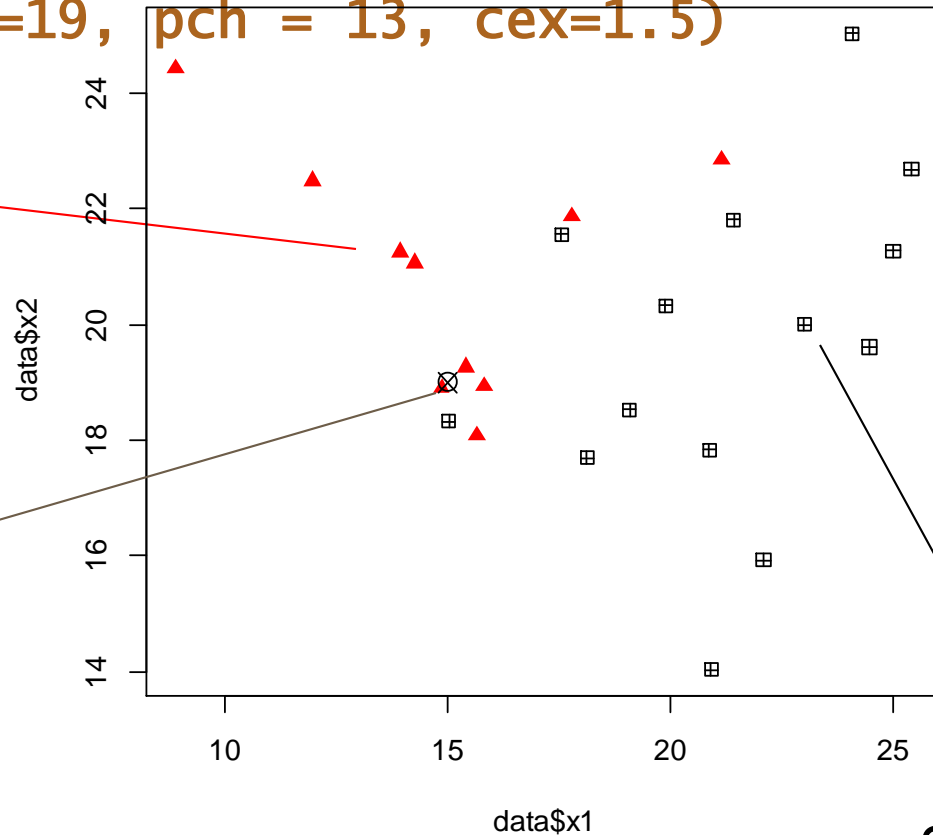
class = 1

Ilustrasi: mengidentifikasi kelas dari amatan baru,
x1=15, x2=19

```
> plot(data$x1, data$x2, col=data$class,  
      pch=ifelse(data$class>1,17,12))  
> points(x=15, y=19, pch = 13, cex=1.5)
```

class = 2

amatan baru



class = 1

Ilustrasi: mengidentifikasi kelas dari amatan baru, **x1=15, x2=19**

```
training <- data[,1:2]
kelas <- as.factor(data[,3])
maudiprediksi <- c(15,19)

library(class)
prediksi <- knn(training, maudiprediksi, kelas, k = 5)
prediksi
```

Amatan dengan x1=15, x2=19 masuk ke class = 2

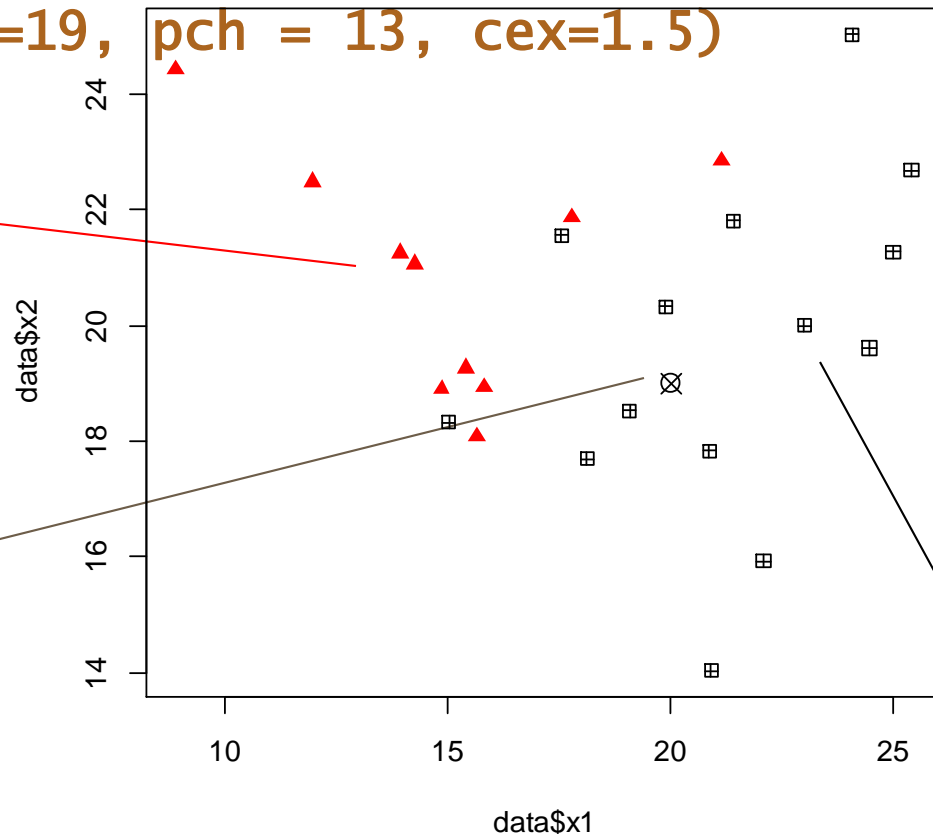
Ilustrasi: mengidentifikasi kelas dari amatan baru,
x1=20, x2=19

```
> plot(data$x1, data$x2, col=data$class,  
      pch=ifelse(data$class>1,17,12))
```

```
> points(x=20, y=19, pch = 13, cex=1.5)
```

class = 2

amatan baru



class = 1

Ilustrasi: mengidentifikasi kelas dari amatan baru, **x1=20, x2=19**

```
training <- data[,1:2]
kelas <- as.factor(data[,3])
maudiprediksi <- c(20,19)

library(class)
prediksi <- knn(training, maudiprediksi,
  kelas, k = 5)
prediksi
```

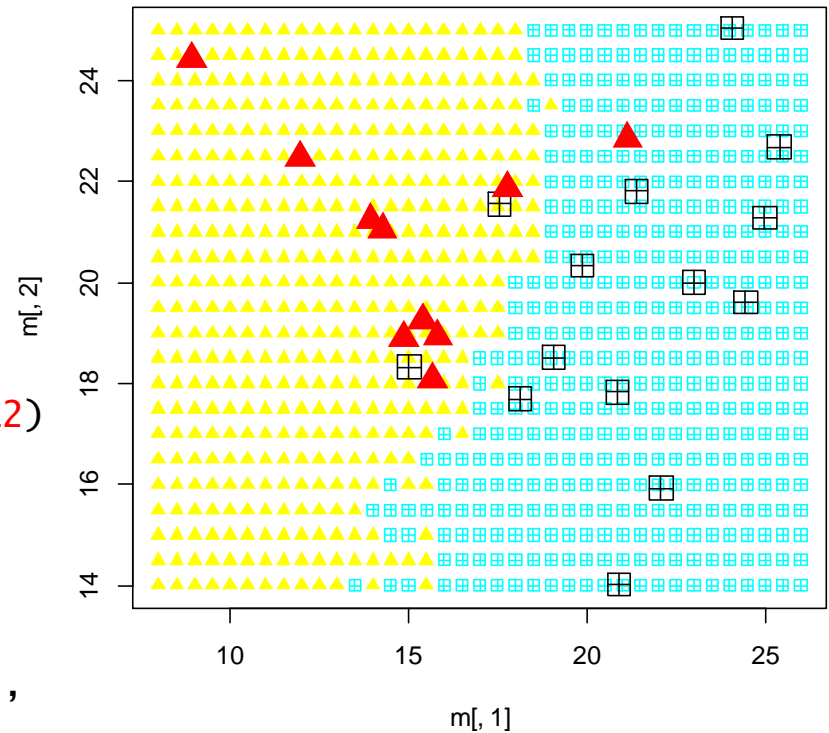
Amatan dengan x1=20, x2=19 masuk ke class = 1

Ilustrasi: mengidentifikasi batas antar kelas berdasarkan knn dengan **k = 12**

```
> m <- NULL
> a <- seq(8, 26, by = 0.5)
> b <- seq(14, 25, by = 0.5)

> for (i in a){
  for (j in b) {
    m <- rbind(m, c(i, j))
  }
}
prediksi <- knn(training, m, kelas, k = 12)

> plot(m[,1], m[,2],
       col=ifelse(prediksi=="1",
                  "cyan", "yellow"),
       pch=ifelse(prediksi=="2", 17, 12))
> points(data$x1, data$x2, col=data$class,
        pch=ifelse(data$class>1, 17, 12),
        cex=2)
```

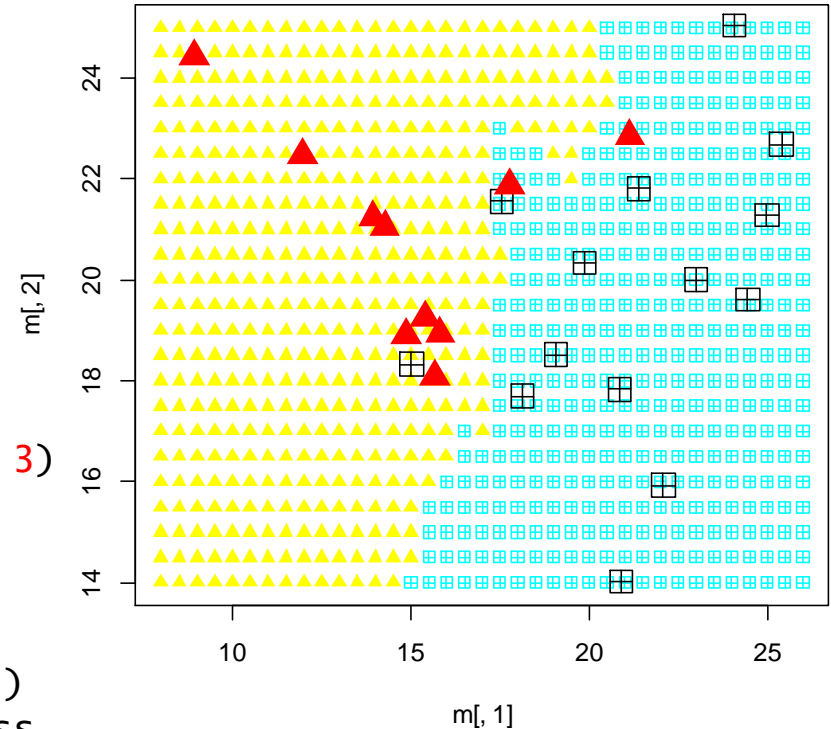


Ilustrasi: mengidentifikasi batas antar kelas berdasarkan knn dengan $k = 3$

```
> m <- NULL
> a <- seq(8, 26, by = 0.5)
> b <- seq(14, 25, by = 0.5)

> for (i in a){
  for (j in b) {
    m <- rbind(m, c(i, j))
  }
}
prediksi <- knn(training, m, kelas, k = 3)

> plot(m[,1], m[,2],
       col=ifelse(prediksi=="1",
                  "cyan","yellow"),
       pch=ifelse(prediksi=="2",17,12))
> points(data$x1, data$x2, col=data$class,
        pch=ifelse(data$class>1,17,12),
        cex=2)
```

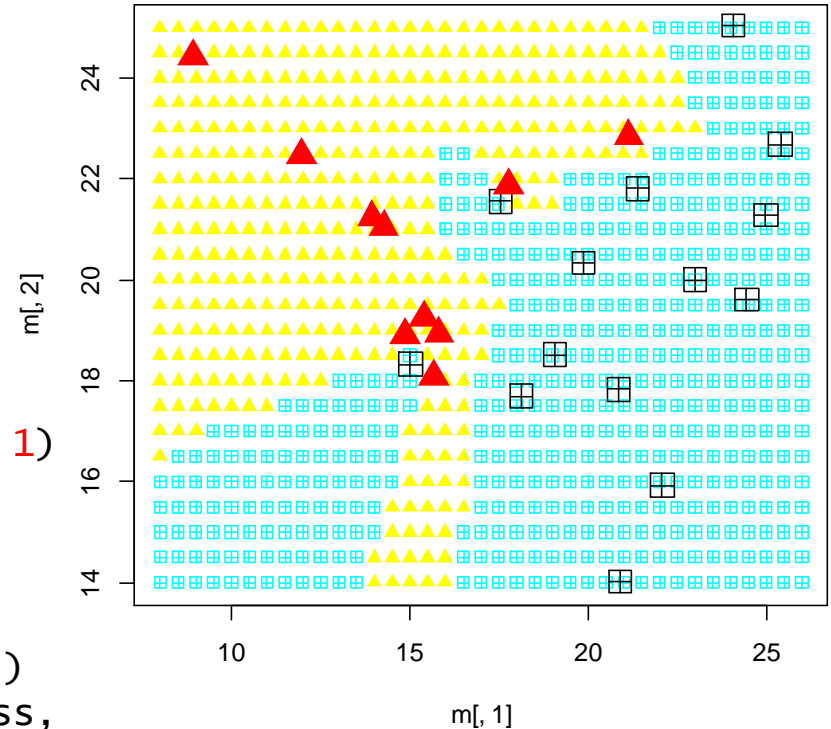


Ilustrasi: mengidentifikasi batas antar kelas berdasarkan knn dengan $k = 1$

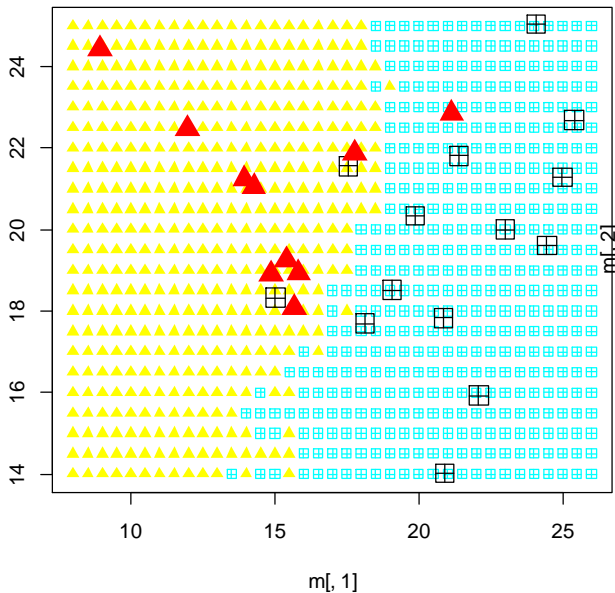
```
> m <- NULL
> a <- seq(8, 26, by = 0.5)
> b <- seq(14, 25, by = 0.5)

> for (i in a){
  for (j in b) {
    m <- rbind(m, c(i, j))
  }
}
prediksi <- knn(training, m, kelas, k = 1)

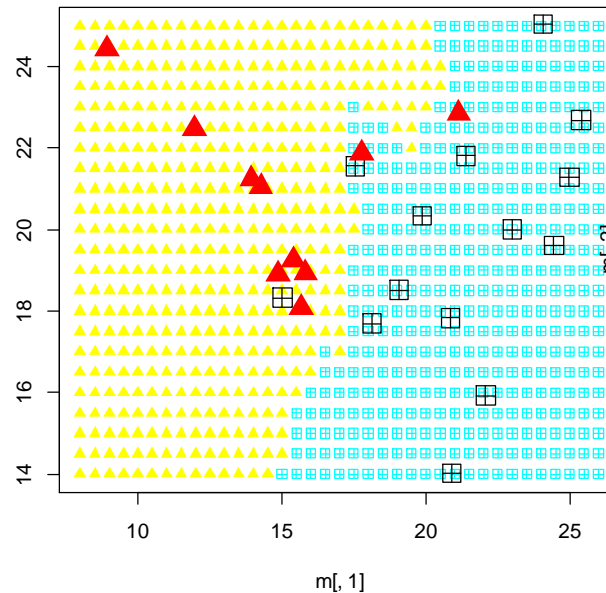
> plot(m[,1], m[,2],
       col=ifelse(prediksi=="1",
                  "cyan", "yellow"),
       pch=ifelse(prediksi=="2", 17, 12))
> points(data$x1, data$x2, col=data$class,
        pch=ifelse(data$class>1, 17, 12),
        cex=2)
```



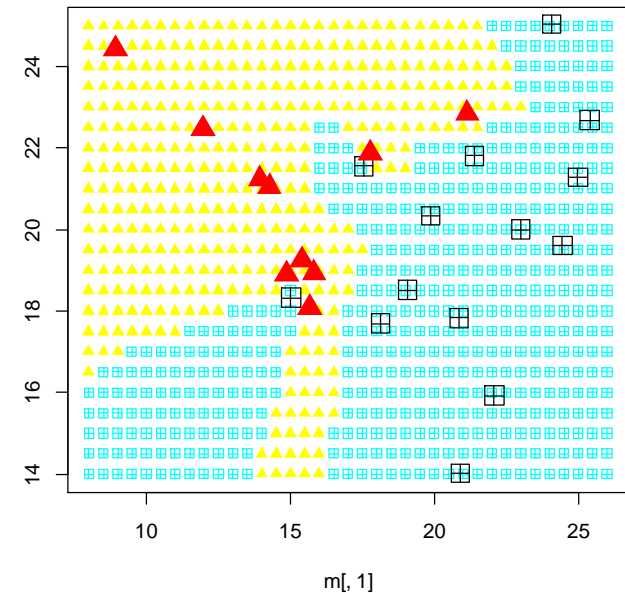
Perbedaan batas kelas hasil kNN dengan k berbeda-beda



k=12



k=3



k=1

Mana yang lebih baik?

Gunakan validasi atau validasi silang

Latihan: Data Kualitas Wine

Data ini berasal dari pengamatan terhadap sejumlah tipe 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 raser 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.

Latihan

Gunakan data “white_wine2.csv”

Gunakan hanya 2 variabel prediktor:

- DENSITY
- ALCOHOL

Kelas kualitas:

Jika skor kualitas $> 6 \rightarrow$ Baik

Skor kualitas $< 6 \rightarrow$ Kurang