



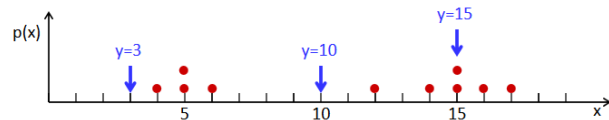
Module 4 - Technique non paramétrique

Solution des exercices

Exercice 2

Soit un ensemble de données $D = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$.

La représentation graphique de ces données est donnée par :



La densité de probabilité $p(x)$ peut être calculée selon :

$$p_M(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M \frac{1}{h_M^d} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_M}\right)$$

1. Cas où $h_M = 4$.

L'estimateur de Parzen au point $x = 3$ est donné par :

$$p(x = 3) = \frac{1}{10} \sum_{i=1}^{10} \frac{1}{4^1} \varphi\left(\frac{x - x_i}{4}\right)$$

$$p(x = 3) = \frac{1}{10 \times 4^1} \left[\underbrace{\varphi\left(\frac{3-4}{4}\right)}_{\left|\frac{-1}{4}\right| \leq 1/2} + \underbrace{\varphi\left(\frac{3-5}{4}\right)}_{\left|\frac{-2}{4}\right| \leq 1/2} + \underbrace{\varphi\left(\frac{3-5}{4}\right)}_{\left|\frac{-2}{4}\right| \leq 1/2} + \dots + \underbrace{\varphi\left(\frac{3-17}{4}\right)}_{\left|\frac{-14}{4}\right| > 1/2} \right]$$

$$P(x = 3) = \frac{1}{10 \times 4^1} [1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0]$$

$$P(x = 3) = \frac{3}{10 \times 4^1} = 0,075$$

L'estimateur de Parzen au point $x = 10$ est donné par :

$$p(x = 10) = \frac{1}{10} \sum_{i=1}^{10} \frac{1}{4^1} \varphi\left(\frac{x - x_i}{4}\right)$$

$$p(x = 10) = \frac{1}{10 \times 4^1} \left[\underbrace{\varphi\left(\frac{10-4}{4}\right)}_{\left|\frac{6}{4}\right| > 1/2} + \underbrace{\varphi\left(\frac{10-5}{4}\right)}_{\left|\frac{5}{4}\right| > 1/2} + \underbrace{\varphi\left(\frac{10-5}{4}\right)}_{\left|\frac{5}{4}\right| > 1/2} + \dots + \underbrace{\varphi\left(\frac{10-17}{4}\right)}_{\left|\frac{-7}{4}\right| > 1/2} \right]$$

$$P(x = 10) = \frac{1}{10 \times 4^1} [0 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0]$$

$$P(x = 10) = \frac{1}{10 \times 4^1} = 0,025$$

L'estimateur de Parzen au point $x = 15$ est donné par :

$$p(x = 15) = \frac{1}{10} \sum_{i=1}^{10} \frac{1}{4^1} \varphi\left(\frac{x - x_i}{4}\right)$$

$$p(x = 15) = \frac{1}{10 \times 4^1} \left[\underbrace{\varphi\left(\frac{15-4}{4}\right)}_{\left|\frac{11}{4}\right| > 1/2} + \underbrace{\varphi\left(\frac{15-5}{4}\right)}_{\left|\frac{10}{4}\right| > 1/2} + \underbrace{\varphi\left(\frac{15-5}{4}\right)}_{\left|\frac{10}{4}\right| > 1/2} + \dots + \underbrace{\varphi\left(\frac{15-17}{4}\right)}_{\left|\frac{-2}{4}\right| \leq 1/2} \right]$$

$$P(x = 15) = \frac{1}{10 \times 4^1} [0 + 0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1]$$

$$P(x = 15) = \frac{1}{10 \times 4^1} = 0,125$$

2. Cas où $h_M = 5$.

L'estimateur de Parzen au point $x = 3$ est donné par :

$$p(x = 3) = \frac{1}{10} \sum_{i=1}^{10} \frac{1}{5^1} \varphi\left(\frac{x - x_i}{4}\right)$$

$$p(x = 3) = \frac{1}{10 \times 5^1} \left[\underbrace{\varphi\left(\frac{3-4}{5}\right)}_{|\frac{-1}{5}| \leq 1/2} + \underbrace{\varphi\left(\frac{3-5}{5}\right)}_{|\frac{-2}{5}| \leq 1/2} + \underbrace{\varphi\left(\frac{3-5}{5}\right)}_{|\frac{-2}{5}| \leq 1/2} + \dots + \underbrace{\varphi\left(\frac{3-17}{5}\right)}_{|\frac{-14}{5}| > 1/2} \right]$$

$$P(x = 3) = \frac{1}{10 \times 5^1} [1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0]$$

$$P(x = 3) = \frac{3}{10 \times 5^1} = 0,06$$

L'estimateur de Parzen au point $x = 10$ est donné par :

$$p(x = 10) = \frac{1}{10} \sum_{i=1}^{10} \frac{1}{5^1} \varphi\left(\frac{x - x_i}{5}\right)$$

$$p(x = 10) = \frac{1}{10 \times 5^1} \left[\underbrace{\varphi\left(\frac{10-4}{5}\right)}_{|\frac{6}{5}| > 1/2} + \underbrace{\varphi\left(\frac{10-5}{5}\right)}_{|\frac{5}{5}| > 1/2} + \underbrace{\varphi\left(\frac{10-5}{5}\right)}_{|\frac{5}{5}| > 1/2} + \dots + \underbrace{\varphi\left(\frac{10-17}{5}\right)}_{|\frac{-7}{5}| > 1/2} \right]$$

$$P(x = 10) = \frac{1}{10 \times 5^1} [0 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0]$$

$$P(x = 10) = \frac{1}{10 \times 5^1} = 0,025$$

L'estimateur de Parzen au point $x = 15$ est donné par :

$$p(x = 15) = \frac{1}{10} \sum_{i=1}^{10} \frac{1}{5^1} \varphi\left(\frac{x - x_i}{5}\right)$$

$$p(x = 15) = \frac{1}{10 \times 5^1} \left[\underbrace{\varphi\left(\frac{15-4}{5}\right)}_{|\frac{11}{5}| > 1/2} + \underbrace{\varphi\left(\frac{15-5}{5}\right)}_{|\frac{10}{5}| > 1/2} + \underbrace{\varphi\left(\frac{15-5}{5}\right)}_{|\frac{10}{5}| > 1/2} + \dots + \underbrace{\varphi\left(\frac{15-17}{5}\right)}_{|\frac{-2}{5}| \leq 1/2} \right]$$

$$P(x = 15) = \frac{1}{10 \times 5^1} [0 + 0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1]$$

$$P(x = 15) = \frac{1}{10 \times 5^1} = 0,02$$

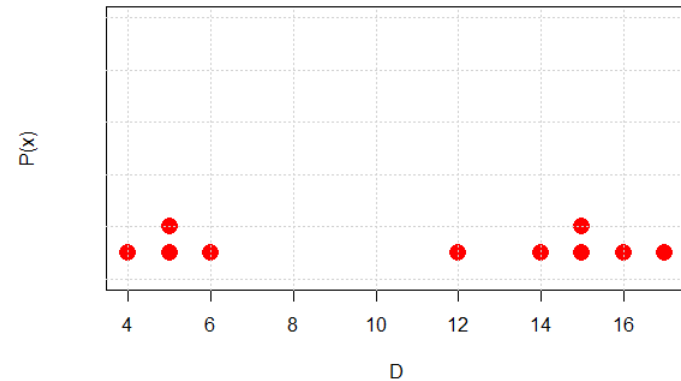
3. Les densité de probabilité changent selon la largeur de la fenêtre de Parzen, c'est-à-dire selon la valeur de h .

Exercice 3

1. Représentation graphique des données :

```
1 D <- c(4,5,5,6,12,14,15,15,16,17)
2 y <- c(1,1,2,1, 1, 1, 1, 2, 1, 1)
3 plot(D,y,ylim=c(0, 10), pch = 16, cex = 2, yaxt='n',ylab
4      ='P(x)',
5      col="red",main="Représentation des points")
6 grid()
```

Représentation des points



2. Calcul des densités pour $h = 4$:

```

1
2 # Fonction de calcul de la densite
3 densite <- function(Points,x,h) {
4   densite <- 0
5   Tempo <- 0
6   for (i in 1:length(Points))
7   {
8     if ( (abs(x-Points[i])/h) > 0.5)
9     {
10      Tempo <- Tempo +0
11    }
12    else{
13      Tempo <- Tempo +1
14    }
15  }
16  densite<-(1/(length(Points)*h^1)) * Tempo
17 }
18
19 # Declaration du vecteur de points
20 D <- c(4,5,5,6,12,14,15,15,16,17)
21
22 # Calcul de la valeur de la densite p(x) pour x=3 et h=4
23 den<-densite(D,3,4)
24 print(den)
25
26 # Calcul de la valeur de la densite p(x) pour x=10 et h=4
27 den<-densite(D,10,4)
28 print(den)
29
30 # Calcul de la valeur de la densite p(x) pour x=15 et h=4
31 den<-densite(D,15,4)
32 print(den)

```

3. Calcul des densités pour $h = 10$:

```

1
2 # Fonction de calcul de la densite
3 densite <- function(Points,x,h) {
4   densite <- 0
5   Tempo <- 0
6   for (i in 1:length(Points))
7   {
8     if ( (abs(x-Points[i])/h) > 0.5)
9     {
10      Tempo <- Tempo +0
11    }
12    else{
13      Tempo <- Tempo +1
14    }

```

```

15   }
16   densite<-(1/(length(Points)*h^1)) * Tempo
17 }
18
19 # Declaration de vecteur de points
20 D <- c(4,5,5,6,12,14,15,15,16,17)
21
22 # Calcul de la valeur de la densite p(x) pour x=3 et h=5
23 den<-densite(D,3,5)
24 print(den)
25
26 # Calcul de la valeur de la densite p(x) pour x=10 et h=5
27 den<-densite(D,10,5)
28 print(den)
29
30 # Calcul de la valeur de la densite p(x) pour x=15 et h=5
31 den<-densite(D,15,5)
32 print(den)

```

Exercice 4

1. Chargement de la base de données *iris* :

```

1 # Charger la base de donnees 'iris'
2 load(iris)

```

2. Énumération des variables :

```

1 # statistiques sur les donnees
2 summary(iris)
3
4 # Nombre d'elements par classe
5 summary(iris$Species)
6
7 # Nombre de lignes
8 NbrL<-nrow(iris)
9 sprintf("Le nombre d'echantillons de la base de donnees
   Iris = %d",NbrL)

```

```

> summary(iris)
  Sepal.Length    Sepal.width    Petal.Length    Petal.width    Species
Min.   :4.300    Min.   :2.000    Min.   :1.000    Min.   :0.100    setosa   :50
1st Qu.:5.100    1st Qu.:2.800    1st Qu.:1.600    1st Qu.:0.300    versicolor:50
Median :5.800    Median :3.000    Median :4.350    Median :1.300    virginica :50
Mean   :5.843    Mean   :3.057    Mean   :3.758    Mean   :1.199
3rd Qu.:6.400    3rd Qu.:3.300    3rd Qu.:5.100    3rd Qu.:1.800
Max.   :7.900    Max.   :4.400    Max.   :6.900    Max.   :2.500

> summary(iris$Species)
  setosa versicolor virginica
      50         50         50

> NbrL<-nrow(iris)
> sprintf("Le nombre d'échantillons de la base de données Iris = %d",NbrL)
[1] "Le nombre d'échantillons de la base de données Iris = 150"

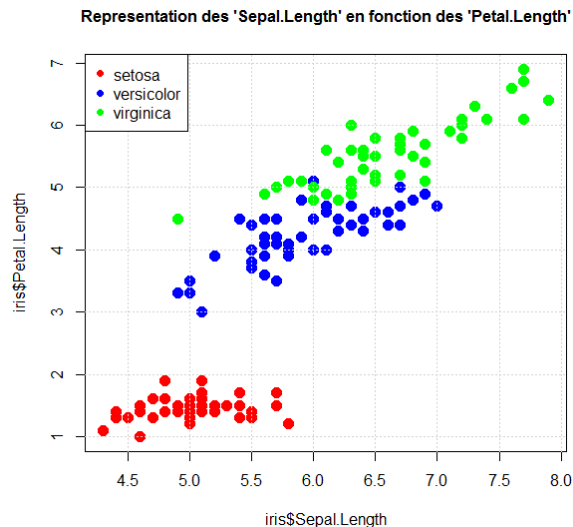
```

3. Visualisation des données des trois groupes en utilisant les caractéristiques *Sepal.Length* et *Petal.Length* :

```

1 #Fixer les couleurs des trois classes
2 COLORS = c("red", "blue", "green")
3
4 plot( x = iris$Sepal.Length, y = iris$Petal.Length,
5       col = COLORS[ as.numeric( iris$Species ) ],pch=19,
6       cex=1.5,
7       main="Représentation des 'Sepal.Length' en fonction
8             des 'Petal.Length'",cex.main=1)
9 grid()
10 #Ajouter une légende
11 legend( "topleft", pch = 19, col = COLORS, legend =
12         levels( iris$Species ) )

```



4. Formation d'un ensemble d'entraînement (80%) et un ensemble de test (20%) :

```

1 library(caTools)
2 # Fixer le seed a 147
3 set.seed(147)
4
5 # Diviser la base de données en base d'entraînement et en
6   base de test
7 sample_iris <- sample.split(iris$Species, SplitRatio =
8                             0.80)
9 iris_train <- subset(iris, sample_iris == TRUE)
10 iris_test  <- subset(iris, sample_iris == FALSE)

```

5. La classification à l'aide de l'estimateur des k-plus voisins en prenant k = 3 :

```

1 library(class)
2 knn.3 <- knn(iris_train[,c(1,3)], iris_test[,c(1,3)],
3             iris_train$Species, k=3)
4
5 # Taux de classification de la base de test
6 100 * sum(iris_test$Species == knn.3)/length(iris_test$
7          Species)
8
9 # Matrice de confusion
10 t(table(knn.3, iris_test$Species))

```

Le taux de classification de la base de test est : 83,33%

La matrice de confusion correspondante :

	Pred-Classe		
	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	6	4
virginica	0	1	9

6. La classification à l'aide de l'estimateur des k-plus voisins en prenant k = 7 :

```

1 library(class)
2 knn.7 <- knn(iris_train[,c(1,3)], iris_test[,c(1,3)],
3             iris_train$Species, k=7)
4 # Taux de classification de la base de test
5 100 * sum(iris_test$Species == knn.7)/length(iris_test$
6 Species)
7 # Matrice de confusion
8 t(table(knn.7 ,iris_test$Species))

```

Le taux de classification de la base de test est : 96,66%

La matrice de confusion correspondante :

		Pred-Classe		
		setosa	versicolor	virginica
Classe	setosa	10	0	0
	versicolor	0	9	1
	virginica	0	0	10

7. On remarque qu'une classification à l'aide de l'estimateur des k-plus voisins en prenant $k = 7$ nous donne le meilleur taux de classification (96,66%).

K=3	K=7
83,33%	96,66%