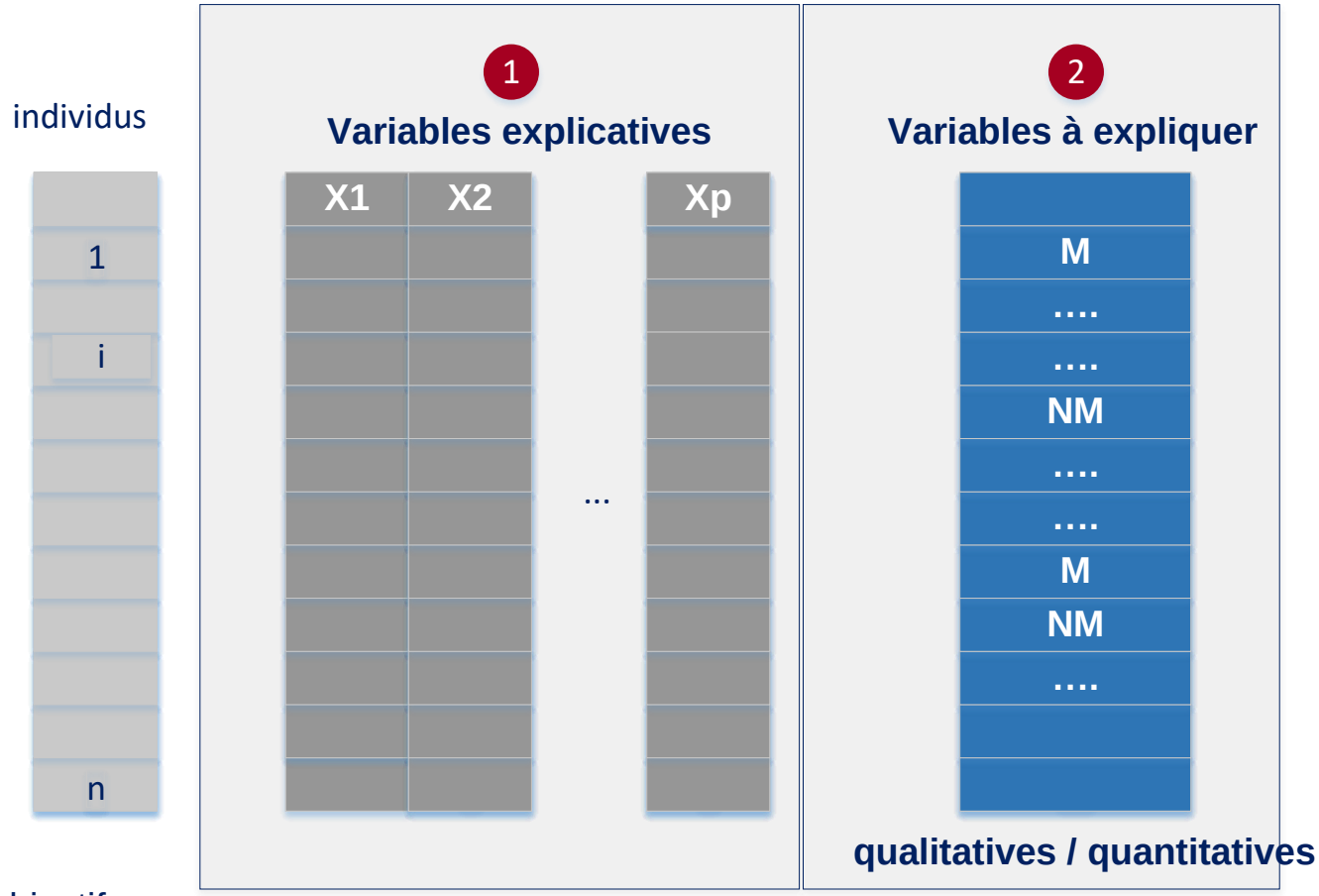
I – LA METHODE DES K PLUS PROCHES VOISINS (KNN)… ET VALIDATION CROISEE

1.1 le contexte

La méthode des k plus proches voisins (KNN) est une méthode de classification supervisée



- Exemple - objectifs

En fonction des données peut on prédire si un (nouveau patient) est atteint d'une pathologie ?

La variable à expliquer est dans ce cas binaire : oui/non (deux classes – cas le plus simple)

Variables explicatives

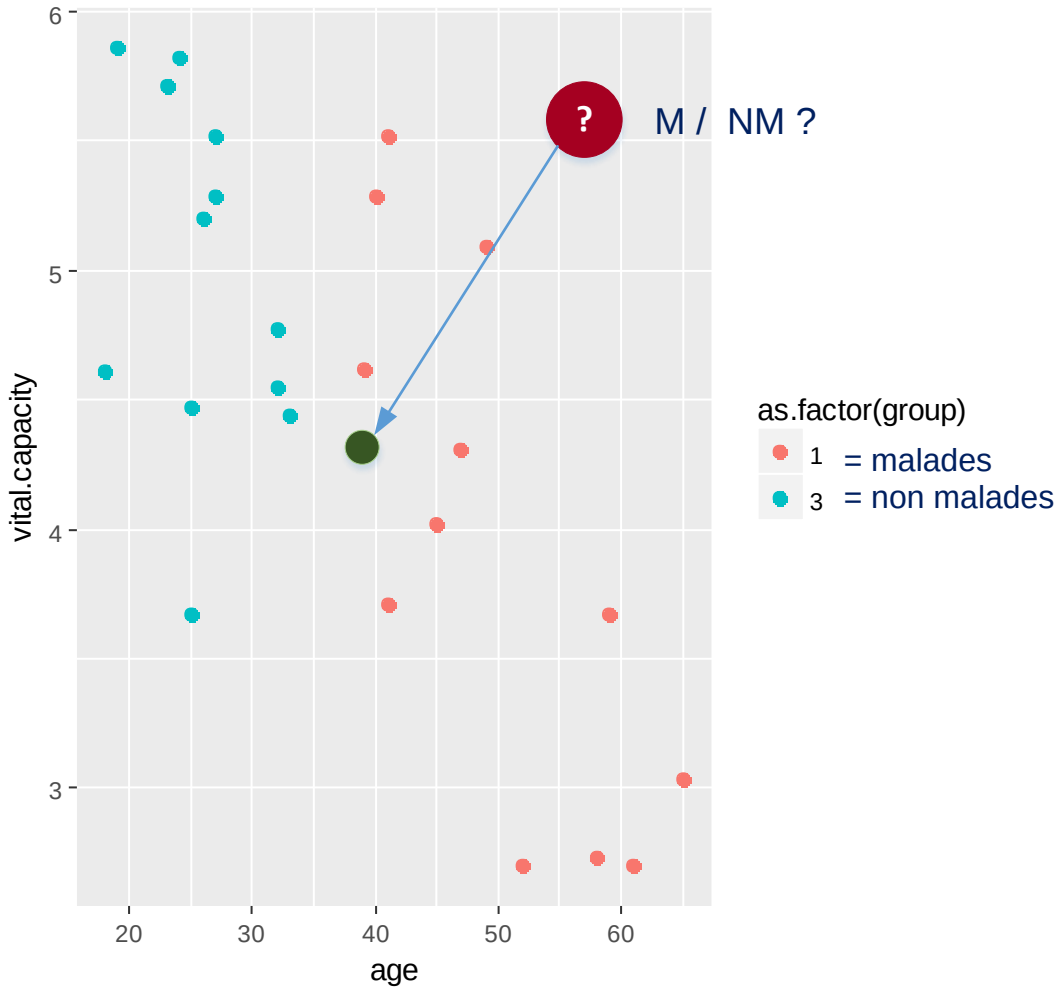
prédicteur

Age	CV

Variable à expliquer

M
….
….
NM
….
….
M
NM
….

CV : capacité respiratoire vitale

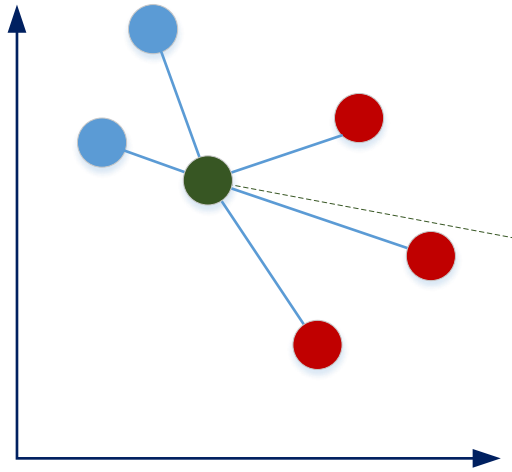


A quelle classe (M / NM) appartient le nouveau patient ?

1.2 Les objectifs

Prédire le plus simplement possible la classe auquel appartient le nouveau patient en utilisant la base de donnée initiale (…. dans un premier temps)

- Algorithme : Déterminer le (les) k plus proche(s) voisin(s) en calculant les distances



group	age	cv
3	32	4.55
3	33	4.44
1	39	4.62
1	40	5.29
1	41	5.52
?	34	4.95

- On utilisera la distance euclidienne (mais ils en existent d'autres – cf. chapitre suivant)

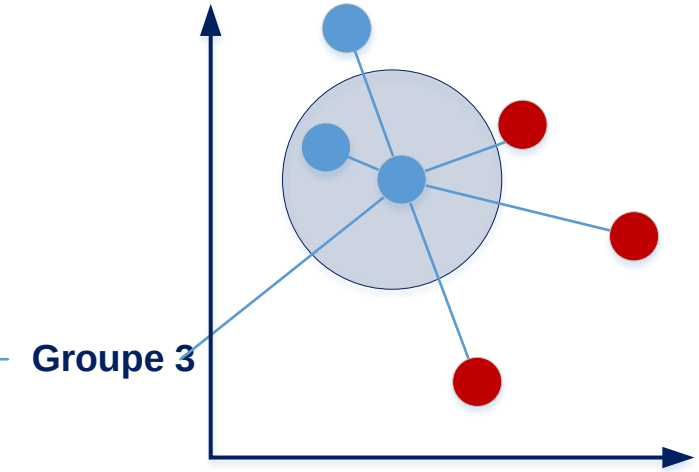
$$D_{AB} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$D_{i,5} = \sqrt{(41+34)^2 + (5.52 - 4.95)} = 7.02$$

	group	age	cv	Distance
1	3	32.00	4.55	2.04
2	3	33.00	4.44	1.12
3	1	39.00	4.62	5.01
4	1	40.00	5.29	6.01
5	1	41.00	5.52	7.02
i	?	34.00	4.95	

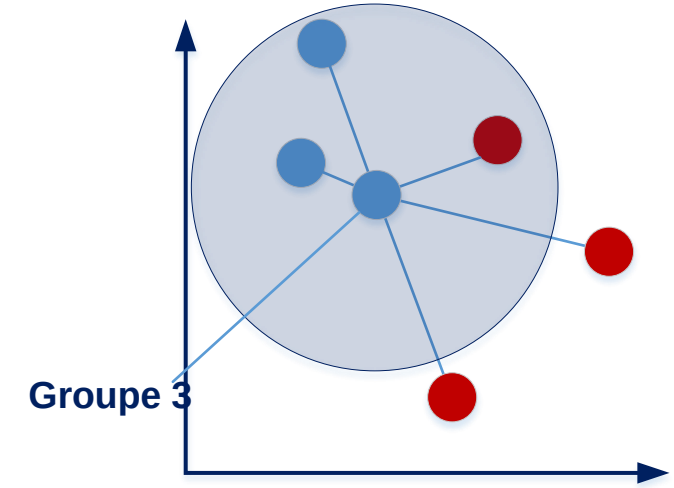
→ k = 1 (le plus proche voisin)

	group	age	cv	Distance
1	3	32.00	4.55	2.04
3	1	39.00	4.62	5.01
4	1	40.00	5.29	6.01
5	1	41.00	5.52	7.02



→ k = 3 (les 3 plus proches voisins)

	group	age	cv	Distance
1	3	32.00	4.55	2.04
2	3	33.00	4.44	1.12
4	1	40.00	5.29	6.01
5	1	41.00	5.52	7.02



⇒ 2 individus appartiennent au groupe 3

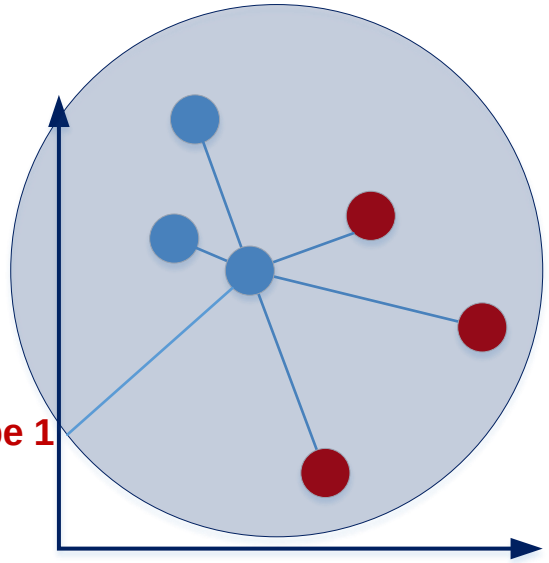
⇒ 1 individus appartiennent au groupe 1

Les éléments les plus proches appartiennent majoritairement au groupe 3

→ L'individus i est alors classé dans le groupe majoritaire c.a.d le groupe 3 (Non malade)

⇒ k = 5 (les 3 plus proches voisins)

	group	age	cv	Distance
1	3	32.00	4.55	2.04
2	3	33.00	4.44	1.12
3	1	39.00	4.62	5.01
4	1	40.00	5.29	6.01
5	1	41.00	5.52	7.02
i	?	34.00	4.95	



- 2 individus appartiennent au groupe 3
- 3 individus appartiennent au groupe 1

Les éléments les plus proches appartiennent majoritairement au groupe 1

→ L'individus i est alors classé dans le groupe majoritaire c.a.d le groupe 1 (malade)

Procédure

A. Trouver les k plus proches observations

B. Utiliser une règle de décision à la majorité pour classer la nouvelle observation

Comment choisir le « meilleur » k ?

Il s'agit d'un « compromis » . La valeur de k pourra être sera choisie en utilisant une validation croisée (cf. suite du cours)

- La détermination de la similarité est fondée sur la mesure des distances

$$D(x_i, x_j) = \left[\sum_{k=1}^{d} \left(x_{ik} - x_{jk} \right)^r \right]^{\frac{1}{r}}$$ Distance de Minkowski

  - r = 1 : Distance de Manhattan (norme L1)

$$D(x_i, x_j) = \sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|$$

  - r = 2 : Distance de Euclidienne (norme L2)

Sélectionnée par défaut

$$D(x_i, x_j) = \left[\sum_{k=1}^{d} \left(x_{ik} - x_{jk} \right)^2 \right]^{\frac{1}{2}} = \sqrt{(x_{i,1} - x_{j,1})^2 + (x_{i,k} - x_{j,k})^2 + \ldots + (x_{i,d} - x_{j,d})^2}$$

Les distances de Minkowski sont dépendantes des mesures utilisées par les variables. Autrement dit, les variables ayant de grandes valeurs vont dominer les distances. On normalise les distances pour donner le même poids à chaque variable)

- Variation min-max $x_{i,k} = \dfrac{x_{i,k} - \min(x_{.k})}{\max(x_{.k}) - \min(x_{.k})}$ → variation relative (en %) entre le min et le max

- Z-score $x_{i,k} = \dfrac{x_{i,k} - \bar{x}_{.k}}{\sqrt{\text{var}(x_{.k})}}$

4.1 Classification en fonction du k plus proches voisins

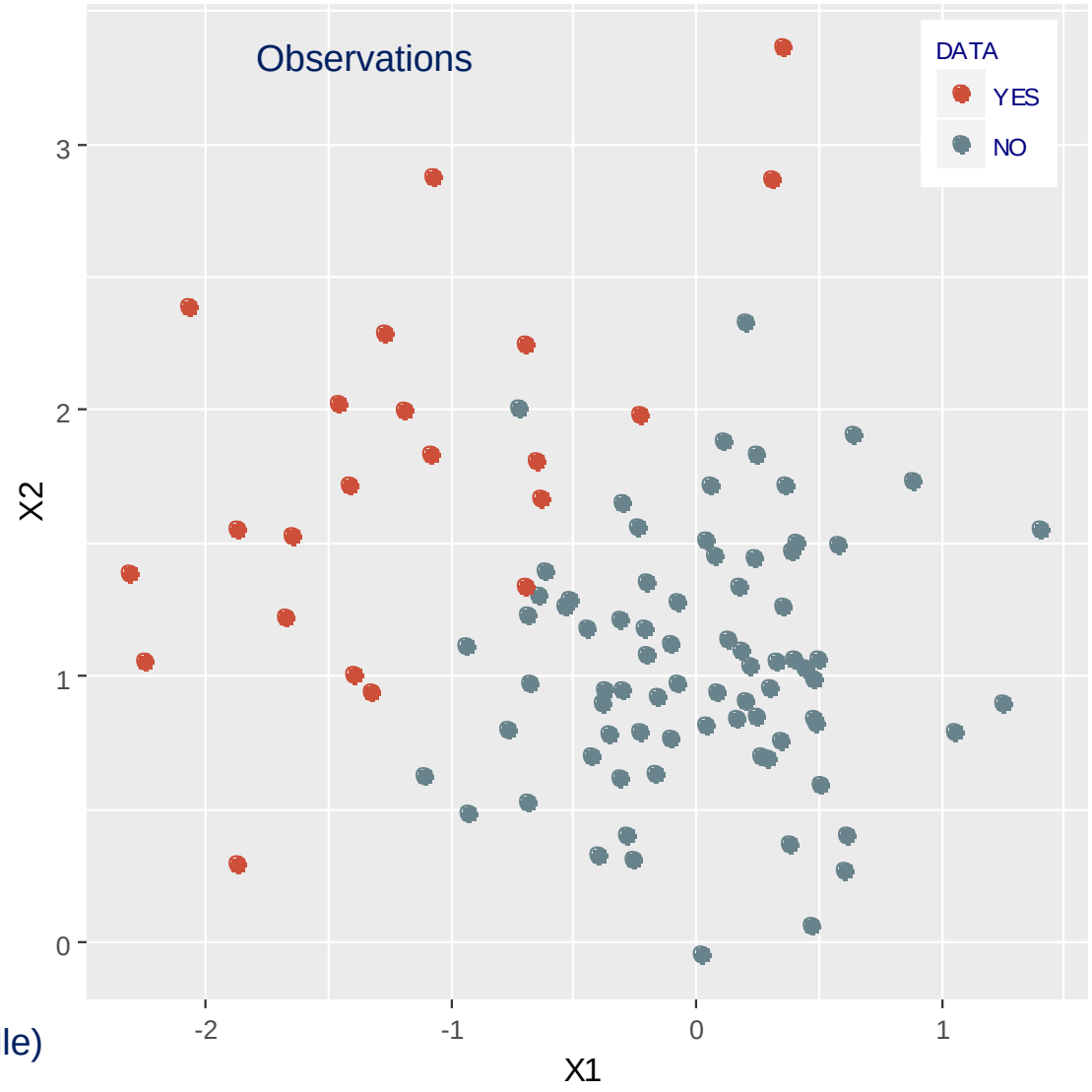
Deux variables explicatives X1 et X2 (quantitatives)
dosages biologiques (**données centrées**)

Yes = Malades = 22
No = Non Malades = 78

Total = 100 observations

X1	X2	
		M
		....
		....
		NM
		....
		....
		M
		NM
		....

Une variable explicative (catégorielle)
annotation



- **K = 30**

Observations

KNN	YES	NO
YES	12	0
NO	10	78
	22	78

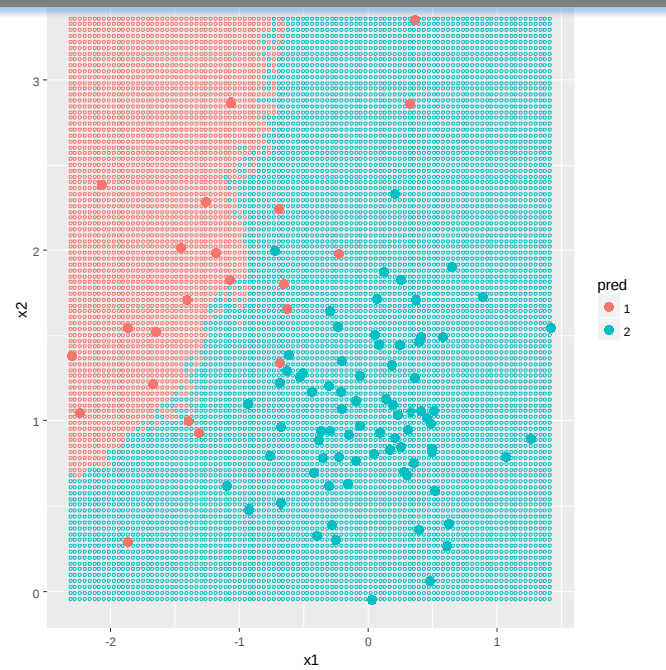
10 mal classées

12 bien classées

0 mal classés

78 bien classées

Taux d'erreur de classement = 10%



- **K = 10**

Observations

KNN	YES	NO
YES	15	1
NO	7	77
	22	78

7 mal classées

12 bien classées

1 mal classée

78 bien classées

Taux d'erreur de classement = (7 + 1)/100 *100 = 8%

- **K = 5**

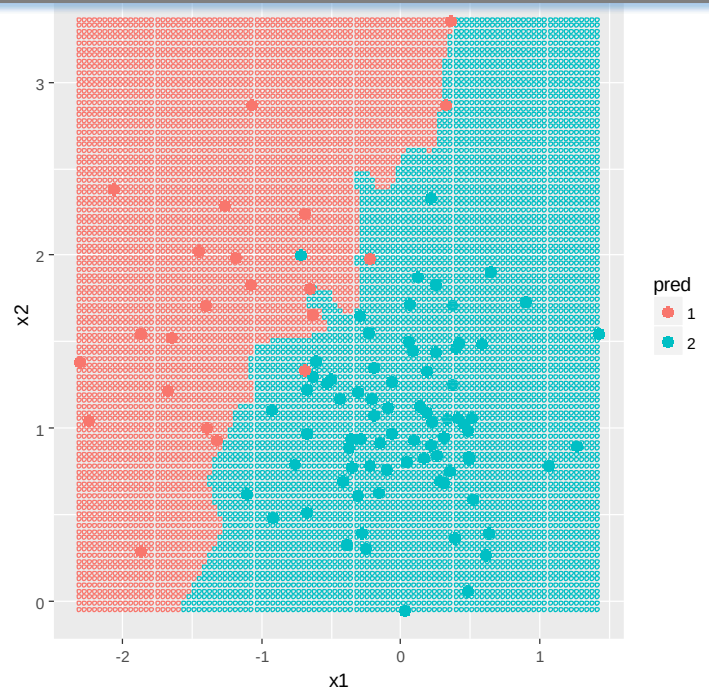
Observations

KNN	YES	NO
YES	18	1
NO	4	77
	22	78

4 mal classées — 18 bien classées

1 mal classée — 78 bien classées

Taux d'erreur de classement = (4 + 1)/100 *100 = 5%



- **K = 3**

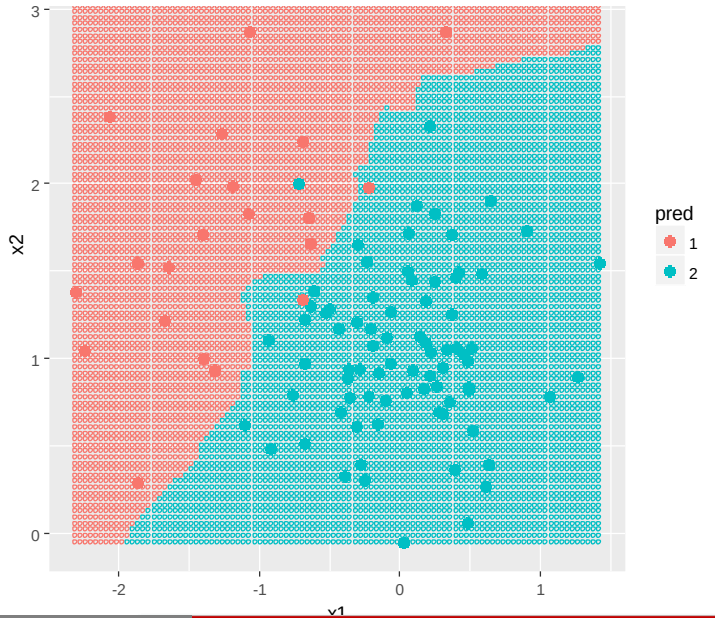
Observations

KNN	YES	NO
YES	20	1
NO	2	77
	22	78

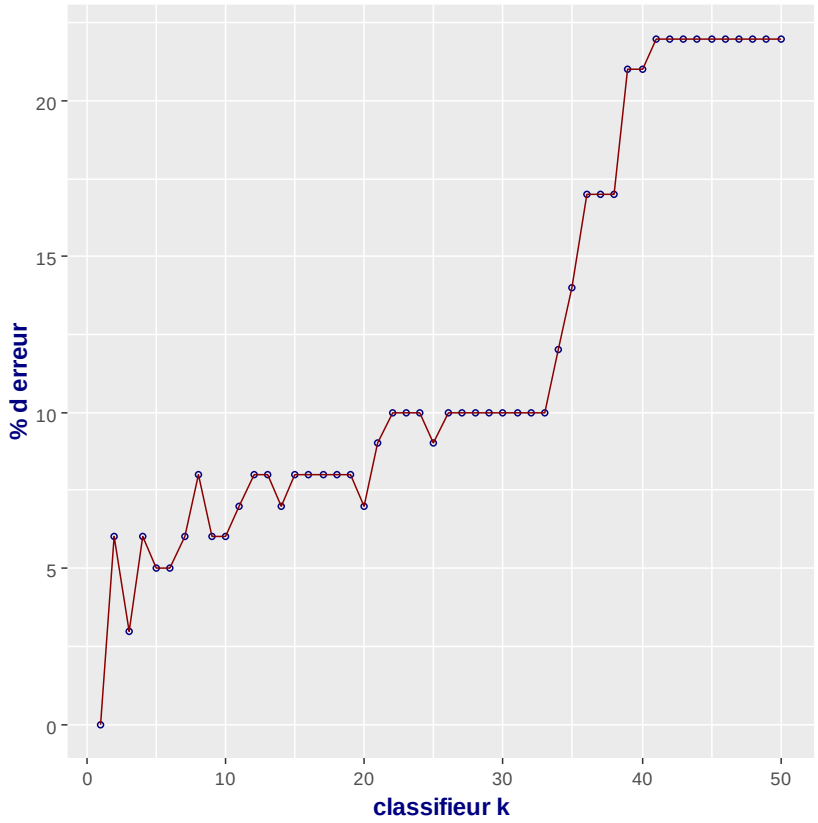
2 mal classées — 20 bien classées

1 mal classée — 78 bien classées

Taux d'erreur de classement = (2 + 1)/100 *100 = 3%



4.2 Classification et généralisation

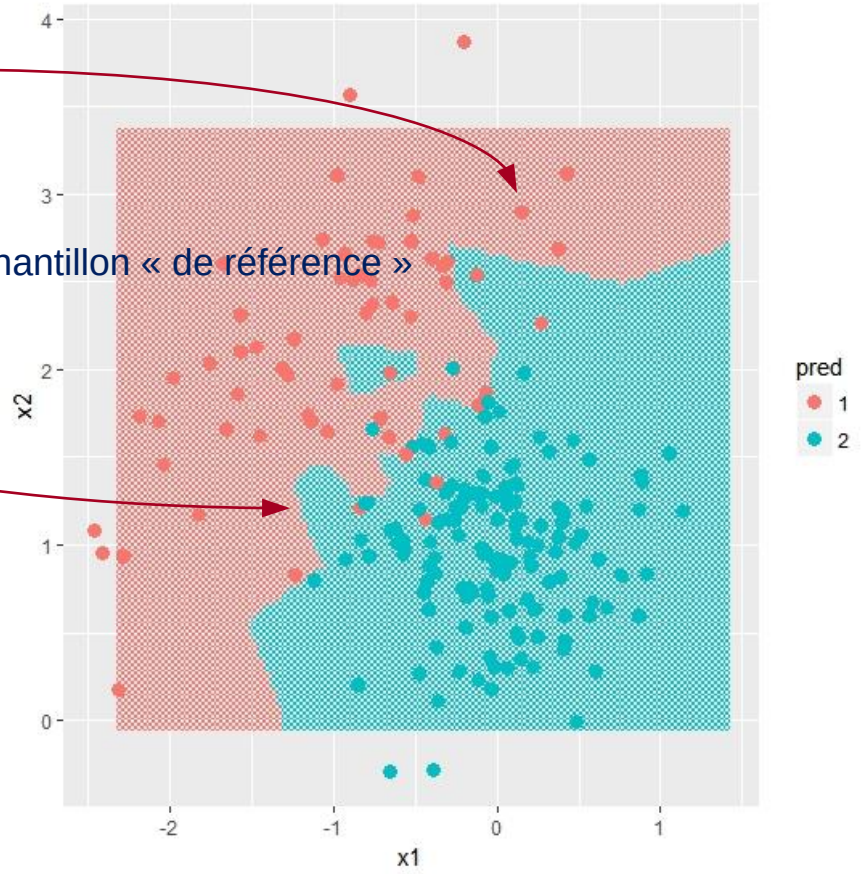


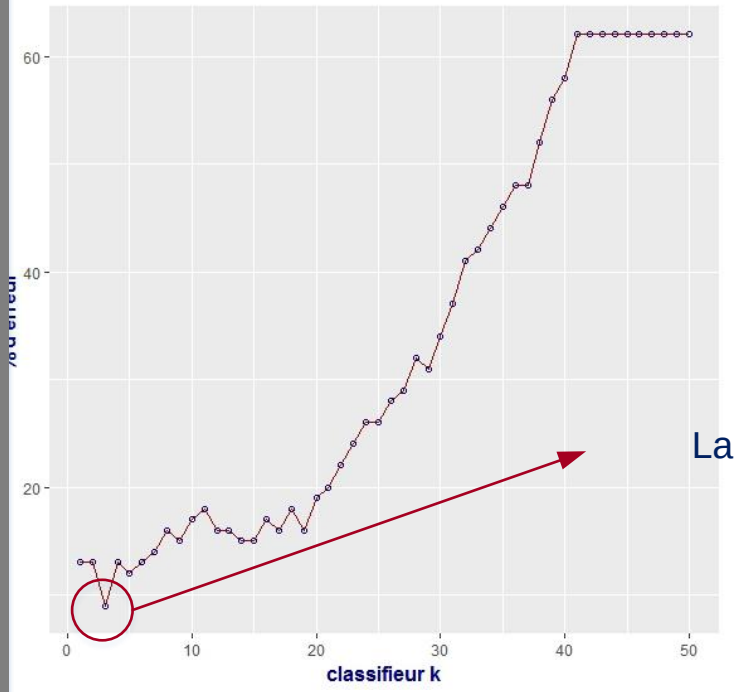
Pourcentage de mal classés en fonction de k

Dans notre exemple, c'est à dire avec l'échantillon étudié (qui sert de référence), la valeur qui sépare au mieux les malades des non malades est k = 1. Si l'on étudie un autre échantillon en prenant comme référence la classification obtenue avec KNN, le k optimal sera t'il toujours égal à 1 ?

Nouvel échantillon

Classification « optimale » (k = 1) obtenue avec l'échantillon « de référence »





La valeur optimale de k n'est plus égale à 1 mais vaut 3 !!!

Quelle sera la validité du modèle (k = 1) lorsqu'il s'agit de l'appliquer sur des données qui n'ont pas participé à son estimation ?

GENERALISATION ?

Avantages

- Simplicité d'apprentissage
- Bonne performance générale
- Incrémental

Inconvénients

- Paramétrage difficile
- Impossibilité d'interprétation
- Lenteur
- Sensibilité à la dimentionnalité et aux variables non pertinentes

Les points importants

- Permet de traiter le problèmes avec un grand nombre d'attributs, mais plus le nombre est important, plus le nombre d'exemples doit être grand
- Nécessite un espace mémoire important pour stocker les données
- Le performance de la méthode dépend de la métrique, du nombre de voisins
- Choix de k doit être déterminé par VC. On peut aussi prendre k = nb d'attributs + 1
- il est possible de pondérer l'influence des observations selon leur éloignement dans le voisinnage
- Choix de la métrique qui est susceptible de modifier les résultats (Normalisation)

La méthode des KPPV peur être utilisée pour l'estimation de données manquantes.

- Pour un individus, une valeur manquante est estimée à partir d'autres individus ayant les mêmes caractéristiques

« Même caractéristique » $\leftrightarrow$ « plus proche voisin »

métrique

choix des variables

$\Rightarrow$ Cette méthode est utilisée pour des données manquantes de type MCAR (Missing completely at random) principalement

Exemple : Y1 = age, Y2 = Sexe, Y3 glycémie

La probabilité que l'âge soit NA ne dépend ni du sexe ni des valeurs de la glycénie. Elle est la même pour tous les sujets

- Méthode des k plus proches voisins

$\rightarrow$ Choix d'un entier $k : 1 \geq k \geq n$

$\rightarrow$ Calculer les distances sur les variables renseignées $d(Y_i, Y_j)$

$\rightarrow$ Retenir les k observations pour lesquelles les distances sont les plus petites $Y_{i_k}, \ldots, Y_{j_k}$

$\rightarrow$ Affecter aux valeurs manquantes la moyenne des valeurs des k voisins $Y_{miss} = \frac{1}{k}\left(Y_{i_k} + \ldots + Y_{j_k}\right)$

k plus proches voisins

- choix du k !
- métrique : distance euclidienne ou de Mahalanobis

Imputation par la méthode des plus proches voisins (exemple)

	y1	y2	y3
1	32	*NA*	2.80
2	32	4.9	*NA*
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	32	4.6	2.88
8	32	10.9	2.90
9	32	8.0	3.28
10	30	1.6	3.20

1 *Variables renseignées (y1 et y3) en ôtant le sujet 2*

	y1	y3
1	32	2.80
3	40	4.38
4	10	3.21
5	6	2.73
6	20	2.81
7	32	2.88
8	32	2.90
9	32	3.28
10	30	3.20

$d(Y_i, Y_j)$

	Distances
1-3	8.154532
1-4	22.003820
1-5	26.000094
1-6	12.000004
1-7	0.080000
1-8	0.100000
1-9	0.480000
1-10	2.039608

2 *Calcul de la distance (sur les variables renseignées)*

3 *Estimation de la valeur manquante en fonction de nombre de plus proches voisin k*

	y1	y2	**y3**
1	**32**	***NA***	**2.80**
2	32	4.9	*NA*
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	**32**	**4.6**	**2.88**
8	**32**	**10.9**	**2.90**
9	**32**	**8.0**	**3.28**
10	30	1.6	3.20

k = 1: estim = 4.6

k = 2 : estim = (4.6+10.9)/2 = 7.4

k = 3 : estim = (4.6+10.9 + 8)/ 3 = 7.53

$$Y_{miss} = \frac{1}{k}\left(Y_{i_k} + \ldots + Y_{j_k}\right)$$

La problématique

Le modèle exacte (théorique) de répartition entre malade et non malade est inconnue. Dans notre cas, on approxime cette répartition avec un modèle (classifier = KNN) et sa famille (k = 1, 2,….) .

On dispose d'un nombre fini de données. A partir de ces données, il faut en déduire une relation (dans notre cas une répartition) sur un nombre infini de données inconnues mais certaines)

Comment « estimer » la meilleure répartition avec un modèle descriptif (qui n'est pas le modèle exact ?)
C 'est ce l'on appelle la ***performance de généralisation***

Les composantes nécessaires

1. Des données
2. Une famille de modèles (dans notre cas, le classifieur = KNN et les familles équivalentes à k = 1,2….)
3. Une fonction de cout (dans notre cas, calculer le pourcentage d'erreur que l'on minimise)
4. Un algorithme (dit algorithme d'adaptation) qui utilise les données pour optimiser la fonction de cout

Adaptation

Un algorithme d'adaptation utilise les données pour optimiser la fonction de cout en considérant les phases suivantes

1. Partitionnement des données → **Validation croisée**
2. Pour chaque valeur du classifieur knn =1,...10,…., on calcule (pour chaque partition B = 1,… 10) le taux d'erreur puis la moyenne des erreurs

On effectue, pour chaque degrés de complexité, un partitionnement des données par une validation croisée

Première approche de l'apprentissage statistique

Première approche

Application à la méthode des plus proches voisins

A **Pour chaque partition** B = (1,2,...10)

Partitionnement

Pour knn = 1 à 40

Classification

B Calcul du taux d'erreur entre les prédictions et les observations

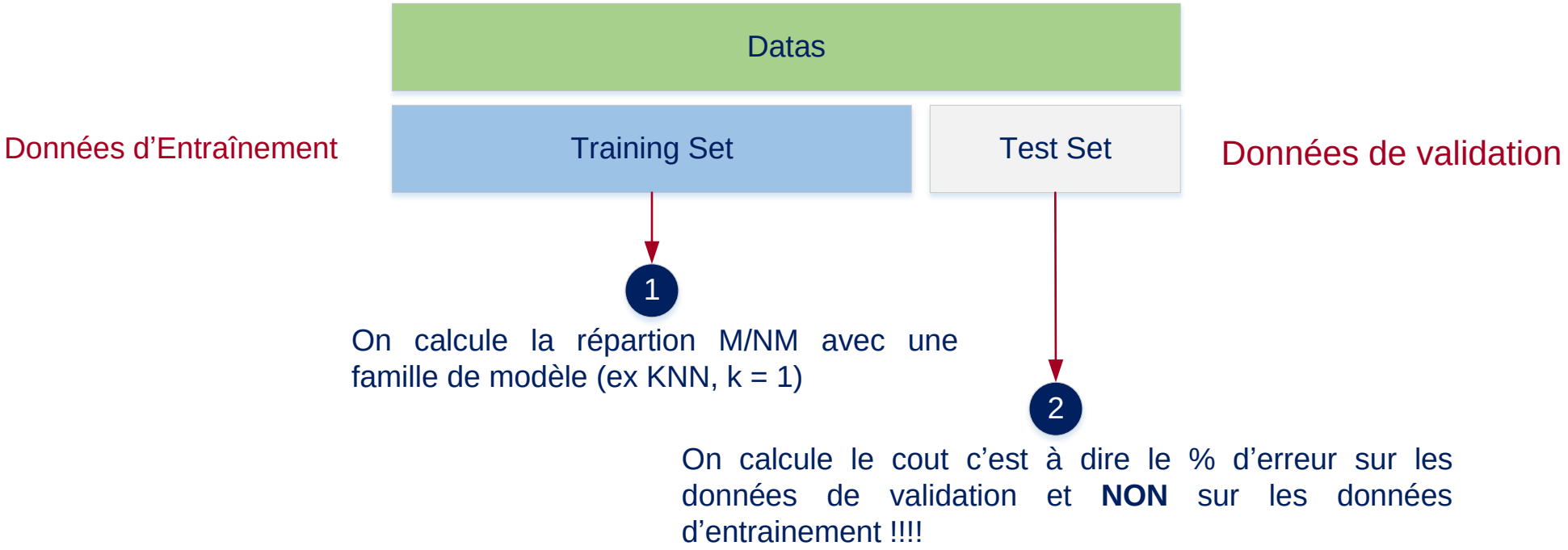
On stocke les résultats

Fin

Fin

C *On calcule le taux d'erreur moyen (à partir de toutes partition pour un degrés de complexité)*

- La validation croisée scinde la base de données disponible de manière à estimer les performances de généralisation du modèle sur des données n'ayant pas servi à l'apprentissage ce qui permet, a priori d'éliminer les solutions surajustées

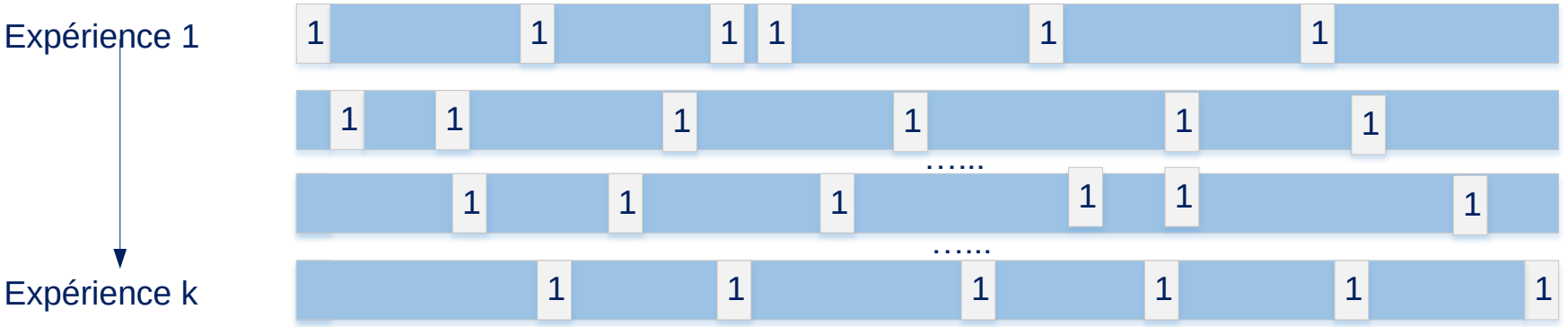


- Les principales méthodes de validation croisée
  - Méthode par échantillonnage (randomisation)
  - Méthode K-Fold (randomisation)
  - Méthode LOOC
  - Boostrap

2.1 Partitionnement par simple échantillonnage

- On décompose le jeux de données initiale en k partitions « tirées » au hasard
- Le nombre d'éléments des échantillons d'entrainement et de validation sont constants pour toutes les partitions

Généralement : données d'entrainement = 75 % et données de validation = 25%



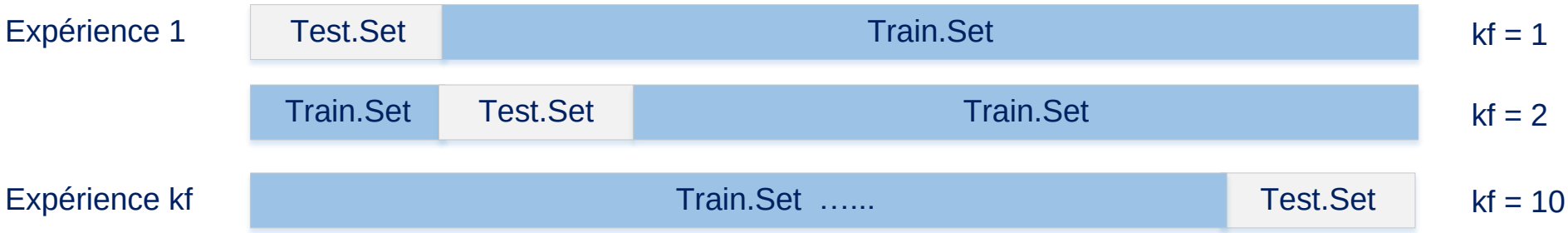
2.2 Partitionnement par LOOCV : Leave-one-out Cross Validation

Le jeux de Validation correspond à une seule valeur (on teste l'ensemble des données de l'échantillon)

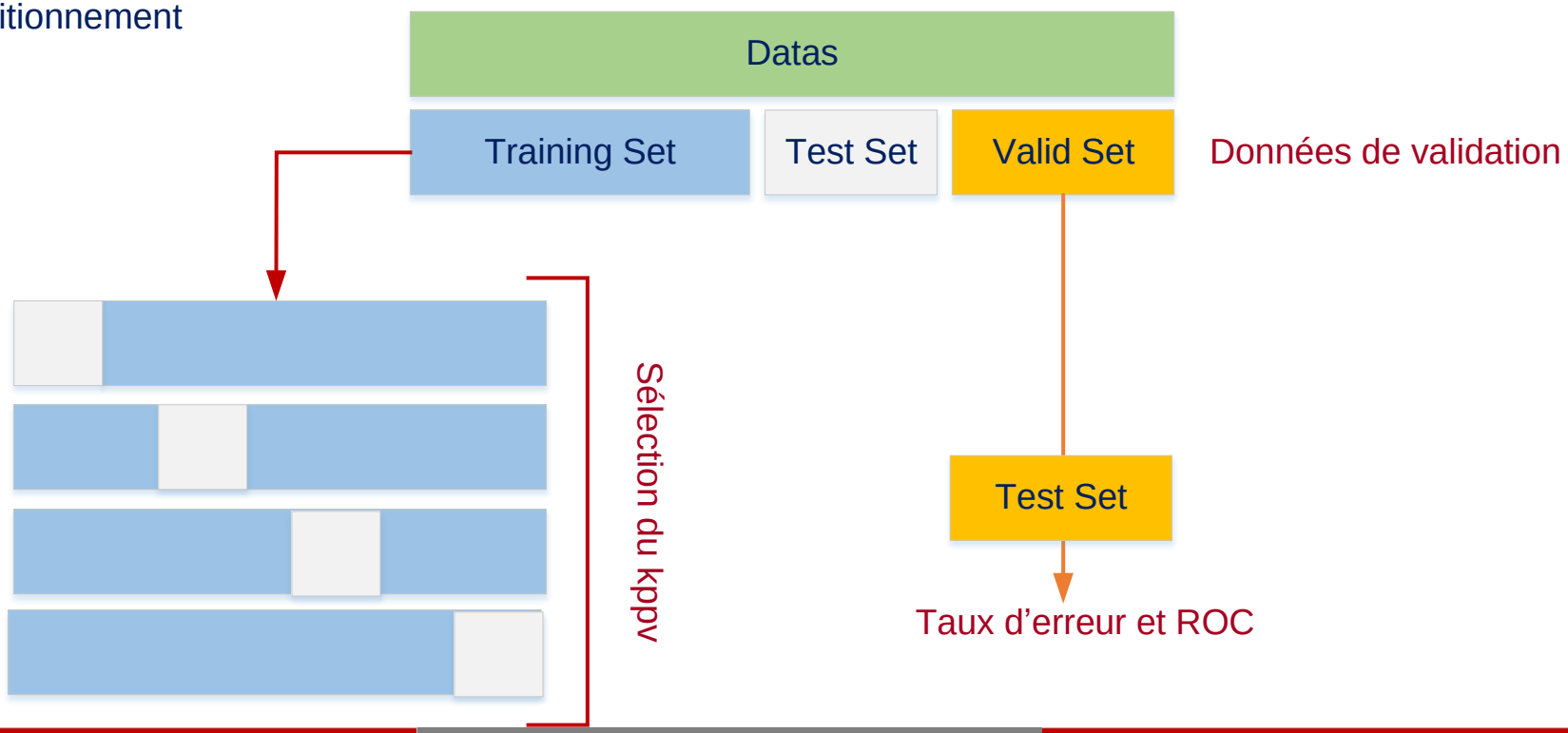
On a donc autant de jeux de validation que de nombre d'individus

2.3 Partitionnement par la méthode des k folds

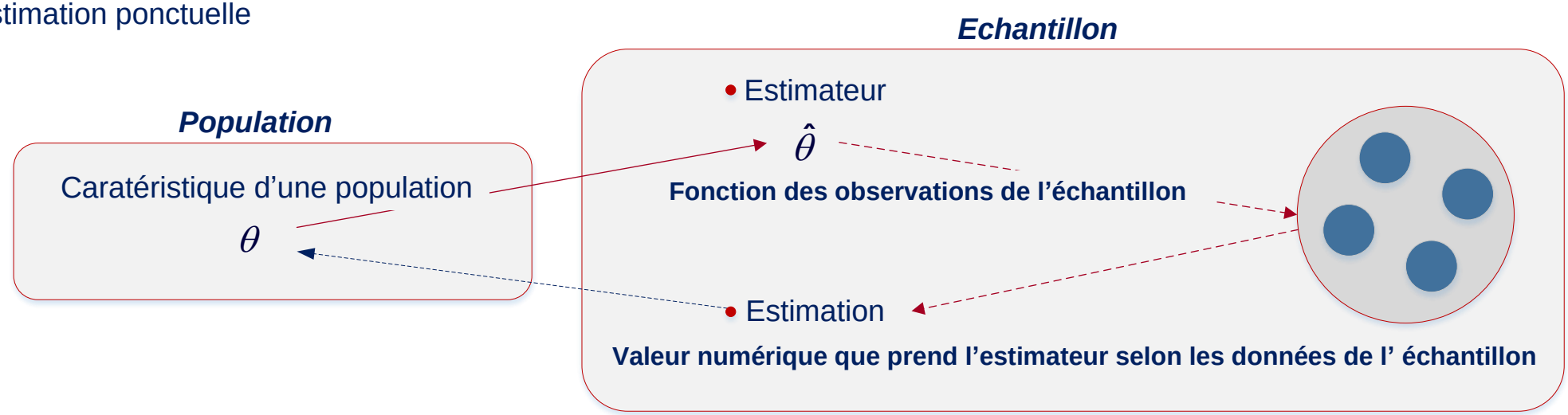
On décompose le jeux de données initiale en *kf* partitions (généralement 5 ou10)



2.4 Double partitionnement



3.1 Estimation ponctuelle



- Lorsqu'une caractéristique d'une population (paramètre θ) est estimée par un seul nombre, déduit des résultats obtenus à partir de l'échantillon, ce nombre est appelé estimation ponctuelle

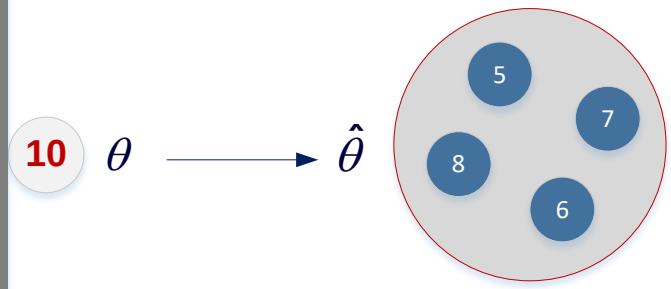
3.2 Définitions variance et biais

- La variance mesure la dispersion de l'estimateur autour de sa moyenne $E(\hat{\theta})$

$$Var(\hat{\theta}) = E\left(\hat{\theta} - E\left(\hat{\theta}\right)\right)^2$$

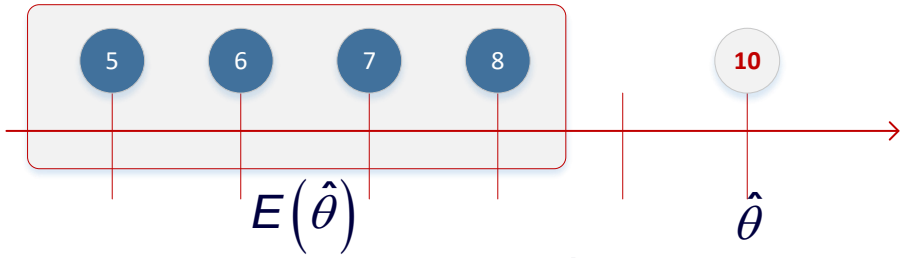
- Le biais mesure la différence entre la moyenne de l'estimateur (obtenue à partir de l'échantillon) et le paramètre de population (caractéristique)

$$B = E(\hat{\theta}) - \theta$$



$\hat{\theta}$	$\left(\hat{\theta} - E\left(\hat{\theta}\right)\right)^2$	
5	2,25	$(5 - 6.5)^2$
6	0,25	
7	0,25	
8	2,25	
6,5	**1,25**	

moyenne $E\left(\hat{\theta}\right)$ $E\left(\hat{\theta} - E\left(\hat{\theta}\right)\right)^2$ variance

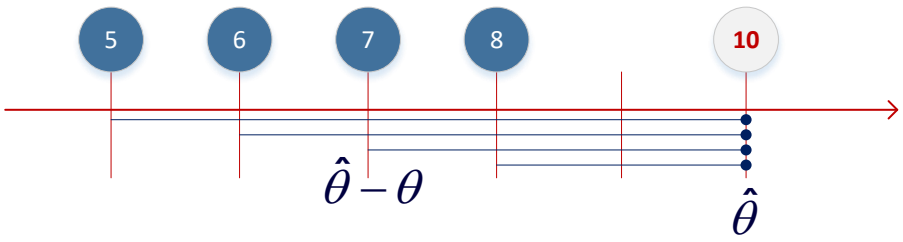


$B = E(\hat{\theta}) - \theta$ → B = 10-6.5 = -3.5

3.3 Définition du risque quadratique

- L'erreur quadratique est une mesure qui caractérise la précision d'un estimateur. Il s'agit d'un critère qui permet la comparaison de différents estimateurs
- L'erreur quadratique (risque quadratique – MSE) d'un estimateur est définie comme l'espérance (moyenne arithmétique) des différences entre les observations et le paramètre (vraie valeur) au carré

$$R(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right]$$



$\hat{\theta}$	$\left(\hat{\theta} - \theta\right)^2$	
5	25	$(5 -10)^2$
6	16	
7	9	
8	4	
6,5	**13,5**	
$E(\hat{\theta})$	$E\left[\left(\hat{\theta} - \theta\right)^2\right]$	

3.4 Relation entre le risque, le biais et la variance

$$R(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] \quad \text{et} \quad B = E(\hat{\theta}) - \theta \Rightarrow \theta = E(\hat{\theta}) - B$$

$$R(\hat{\theta}) = E\left[\left(\hat{\theta} - \left(E(\hat{\theta}) - B\right)\right)^2\right]$$

$$R(\hat{\theta}) = E\left[\hat{\theta}^2 - 2\hat{\theta}\left(E(\hat{\theta}) - B\right) + \left(E(\hat{\theta}) - B\right)^2\right]$$

$$R(\hat{\theta}) = E\left[\hat{\theta}^2 - 2\hat{\theta}E(\hat{\theta}) + 2\hat{\theta}B + E(\hat{\theta})^2 - 2E(\hat{\theta})B + B^2\right]$$

$$R(\hat{\theta}) = E\left[(\hat{\theta}^2 - 2\hat{\theta}E(\hat{\theta}) + E(\hat{\theta})^2) + 2B(\hat{\theta} - E(\hat{\theta})) + B^2\right]$$

$$R(\hat{\theta}) = E\left[(\hat{\theta}^2 - 2\hat{\theta}E(\hat{\theta}) + E(\hat{\theta})^2)\right] + 2BE\left(\hat{\theta} - E(\hat{\theta})\right) + B^2$$

$$R(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 + B^2$$

$$E\left(\hat{\theta} - E(\hat{\theta})\right) = E(\hat{\theta}) - E(E(\hat{\theta})) = E(\hat{\theta}) - E(\hat{\theta}) = 0$$

$$R(\hat{\theta}) = Var(\hat{\theta}) + B^2 \longrightarrow R(\hat{\theta}) = 1.25 + (-3.5)^2 = 13.5$$

3.5 Risque et modèle statistique

Cette partie est fortement simplifiée : github.com/wikistat

- Échantillon d'apprentissage $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$

  X_1 : descripteur ; Y_n : annotation

- Une règle de prévision (**prédicteur)** est une fonction mesurable f qui associe Y à X tel que $f : X \to Y$

  Cette règle (prédiction) est réalisée par un algorithme

- Il est essentiel de mesurer la qualité de prédiction de f

  Pour y parvenir on introduit une nouvelle fonction dite fonction de perte l (à deux paramètres)

  $\Rightarrow$ En régression on définit la perte $l : Y \to Y$ tel que $l(y, y') = |y - y'|^p$

  y' = f(x) valeur prédite

  y : annotation (vraie valeur)

  $\Rightarrow$ En classification (binaire) $l : Y \to Y$ tel que $l(y, y') = 1_{y \neq y'} = \begin{cases} 0 : y = y' \\ 1 : y \neq y \end{cases}$

  On définit le risque empirique associé à Dn $\hat{R}(f, D_n) = \frac{1}{N} \sum_{i=1}^{N} l(Y_i, f(X_i))$

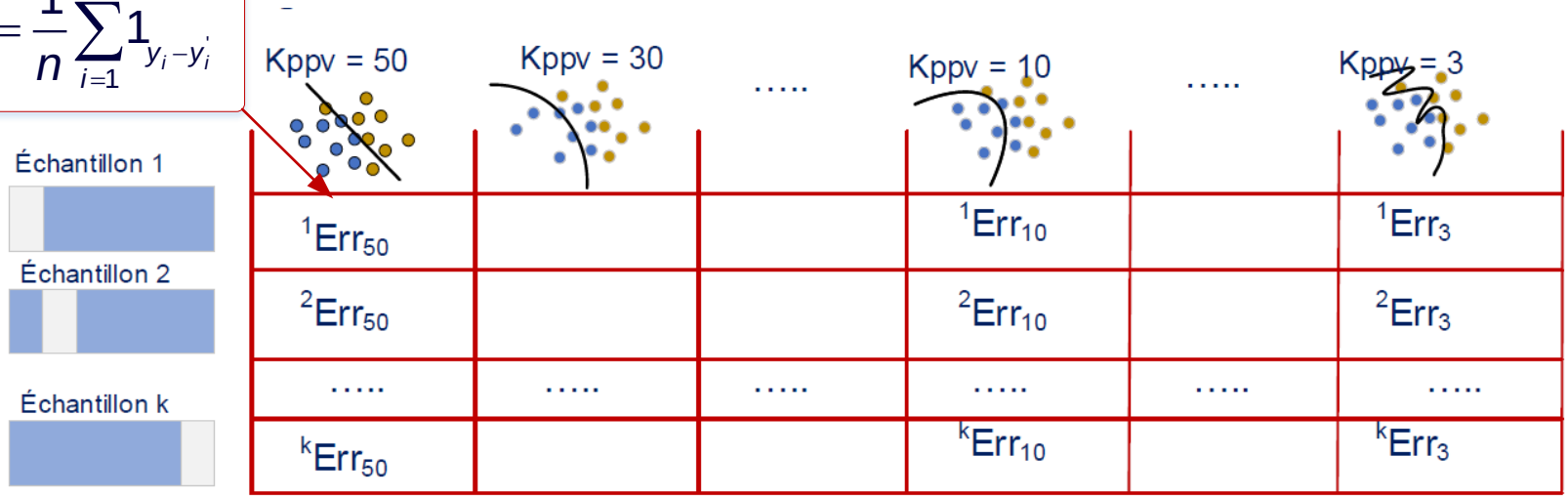
$\hat{R}(f, D_n) = \frac{1}{N} \sum_{i=1}^{N} (y - y')^2$ Risque quadratique en régression……. $= \text{var}(y') + B^2$ avec $B = E(y') - y$

….. que l'on cherche à minimiser

3.6 Compromis biais variance en classification

Risque

$$Err = \frac{1}{n}\sum_{i=1}^{n} 1_{y_i - y_i'}$$



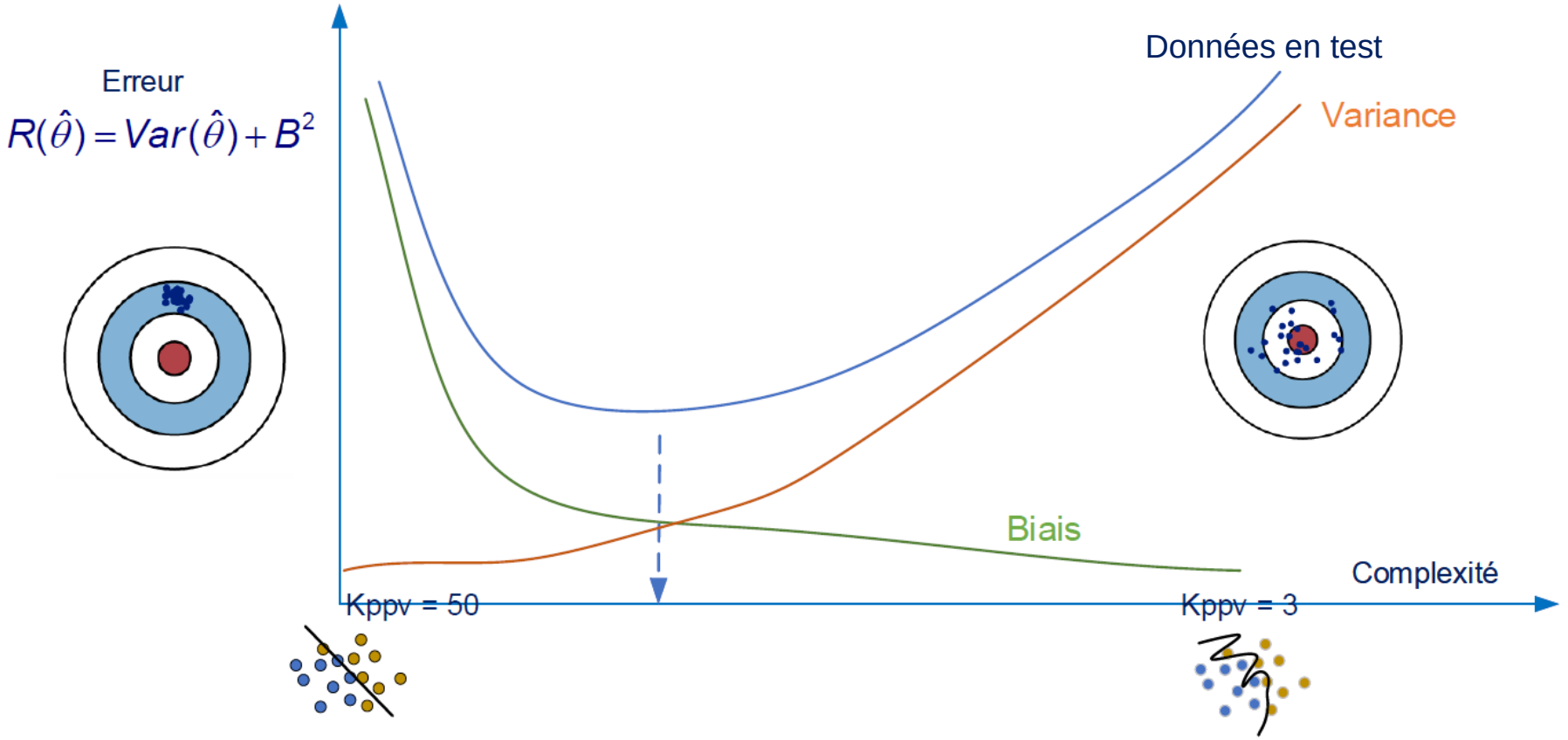
$$E(Err) = E\left(\frac{1}{n}\sum_{i=1}^{n} 1_{y_i - y_i'}\right)$$

n = nombre d'observations par échantillon

k = nombre d'échantillons

$$E(Err) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_k}\sum_{i=1}^{n} 1_{y_i - y_i'} \xrightarrow{N = nk} \frac{1}{N}\sum_{i=1}^{N} 1_{y_i - y_i'}$$

On obtient une estimation du risque quadratique pour des valeurs de k différentes



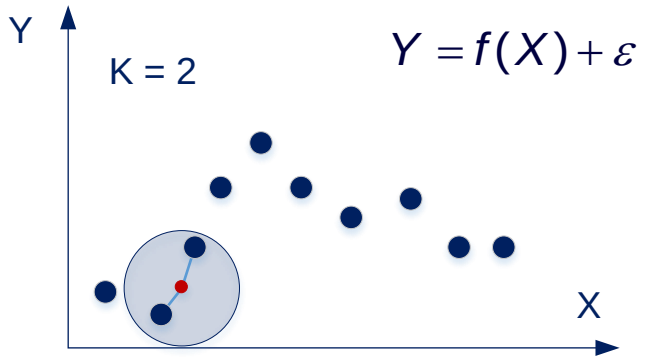
Variance: faible sensibilité du modèle aux données

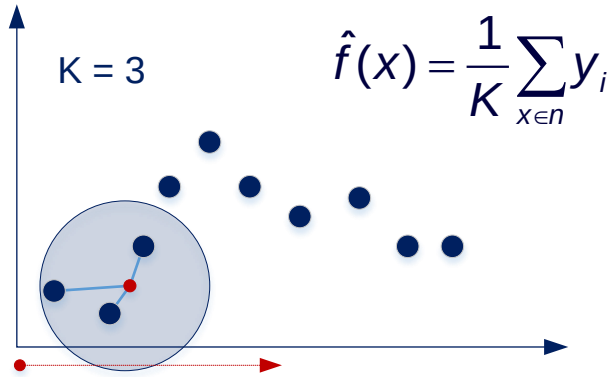
Biais : erreur par rapport modèle importante

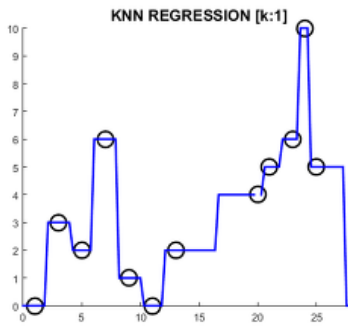
- Variance: forte sensibilité du modèle aux données
- Biais : erreur par rapport au modèle faible

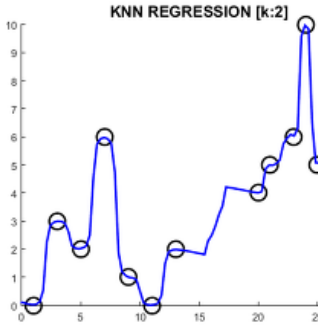
3.7 Compromis biais variance en régression

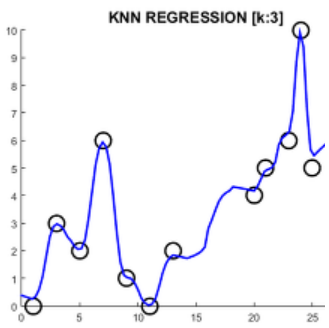
3.7.1 K plus proches voisins en régression

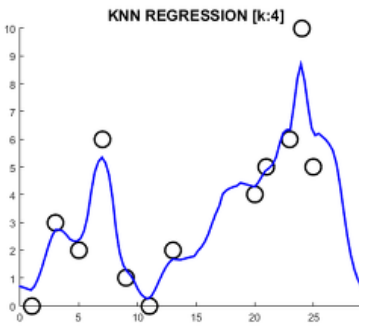


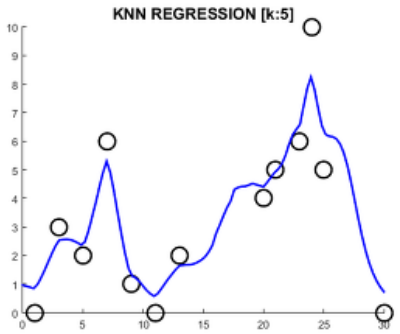




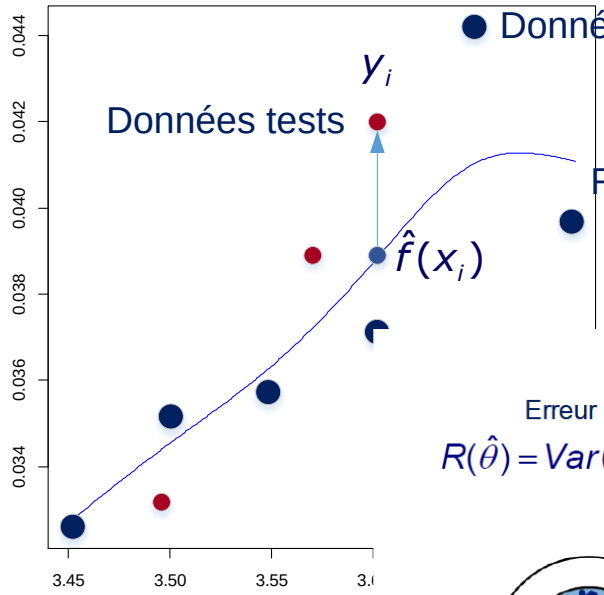








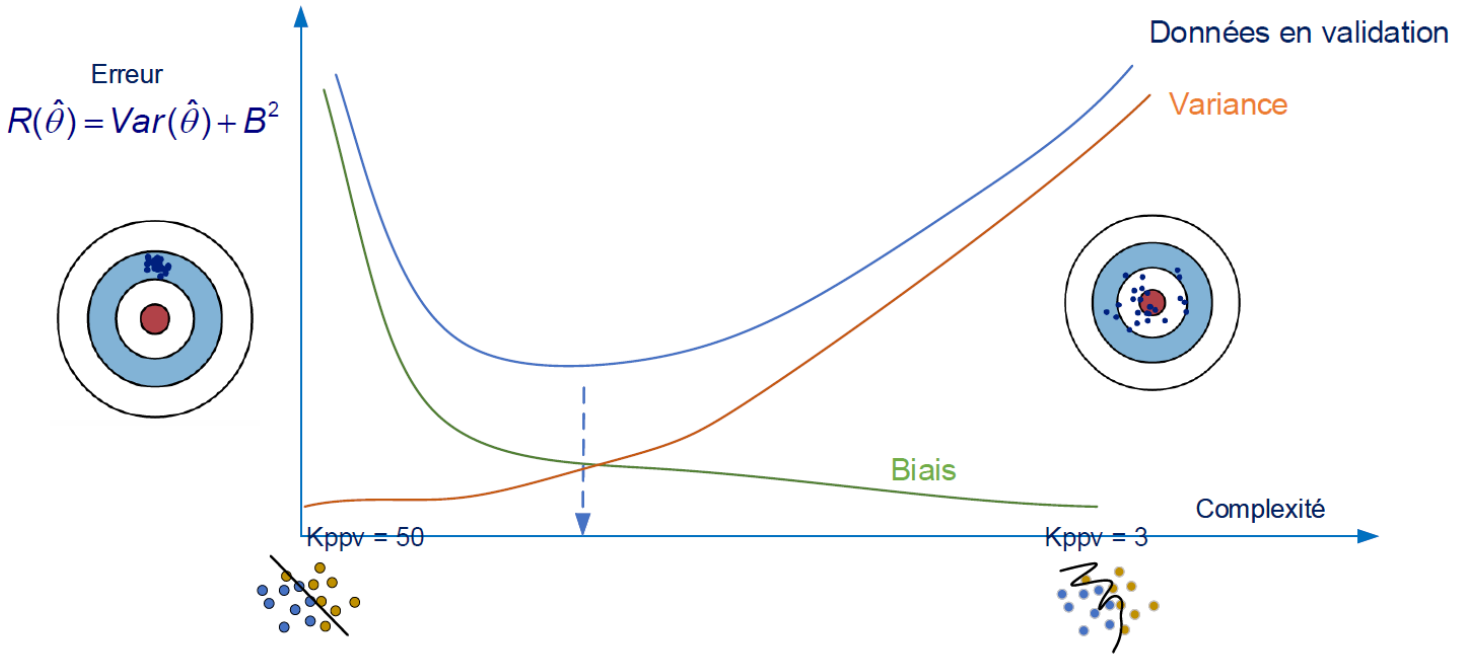
3.7.2 Fonction de risque



$$Err = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Erreur

$$R(\hat{\theta}) = Var(\hat{\theta}) + B^2$$



Variance: faible sensibilité du modèle aux données

Biais : erreur par rapport modèle importante

- Variance: forte sensibilité du modèle aux données
- Biais : erreur par rapport au modèle faible