Chapitre 1 : Introduction à l'apprentissage supervisé

Cours de classification

2022-2023

Déroulement du cours et programme

- Environ 15 heures de cours et 15 heures de TD/TP. Les TP sont sous R.
- 12 séance de 2h30 du mardi 13 septembre au mardi 6 décembre.
- Modalités d'évaluation : 1 TP à rendre + 1 Examen (le 6 décembre)

Programme :

- classifieur de Bayes
- k-plus proches voisins
- modèles de mélange
- régression logistique
- arbres de décision, bagging, forêts aléatoires et boosting
- (support vector machine) si le temps le permet

Références

- The Elements of Statistical Learning, *Data-Mining, Inference and Prediction*, T. Hastie, R. Tibshirani et J. Friedman. Lien
- A Probabilistic Theory of Pattern Recognition, L. Devroye, L. Györfi et G. Lugosi. Lien
- Polycopié de Charlotte Baey
- Vidéos de Kilian Weinberger

Tous les documents seront accessibles au fur et à mesure sur ma page web

Cadre de l'apprentissage supervisé

- On suppose que l'on observe $(X_1, Y_1), \dots (X_n, Y_n)$ n paires de variables aléatoires indépendantes de même loi que $(X, Y) \sim \mathbb{P}$
- On note $\mathcal{D}_n = \{(X_1, Y_1), \dots (X_n, Y_n)\}$ cet échantillon d'apprentissage.
- Les X_i appartiennent à un espace $\mathcal{X}$, on les appelle variables d'entrées, explicatives ou encore covariables. En général $\mathcal{X} \subset \mathbb{R}^d$, $d \geq 1$.
- Les Y_i appartiennent à un espace $\mathcal{Y}$, on les appelle variables de sorties, étiquettes ou encore variables à expliquer. En général $\mathcal{Y} \subset \mathbb{R}$.

But de l'apprentissage supervisé

Le but est de prévoir l'étiquette Y associée à une nouvelle entrée X, où il est sous entendu que (X, Y) est une nouvelle réalisation de $\mathbb{P}$ et (X, Y) est indépendant de $\mathcal{D}_n$.

Exemple 1 : Reconnaissance de chiffres manuscrits.

- X est la caractéristique d'une image (un pixel),
- Y est le chiffre représenté sur l'image.

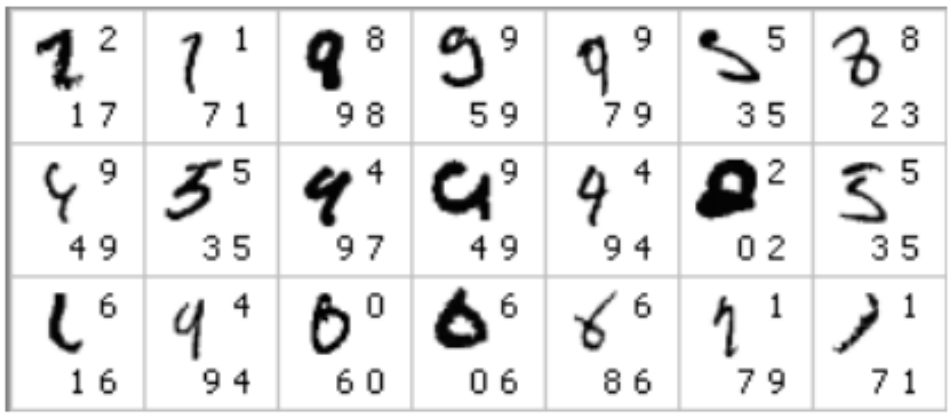


Figure 1: Les données MNIST. Le nombre en haut à droite indique la vraie valeur. Les nombres en bas à droite sont les deux prédictions les plus probables.

But de l'apprentissage supervisé (suite)

Exemple 2 : Détection de spam

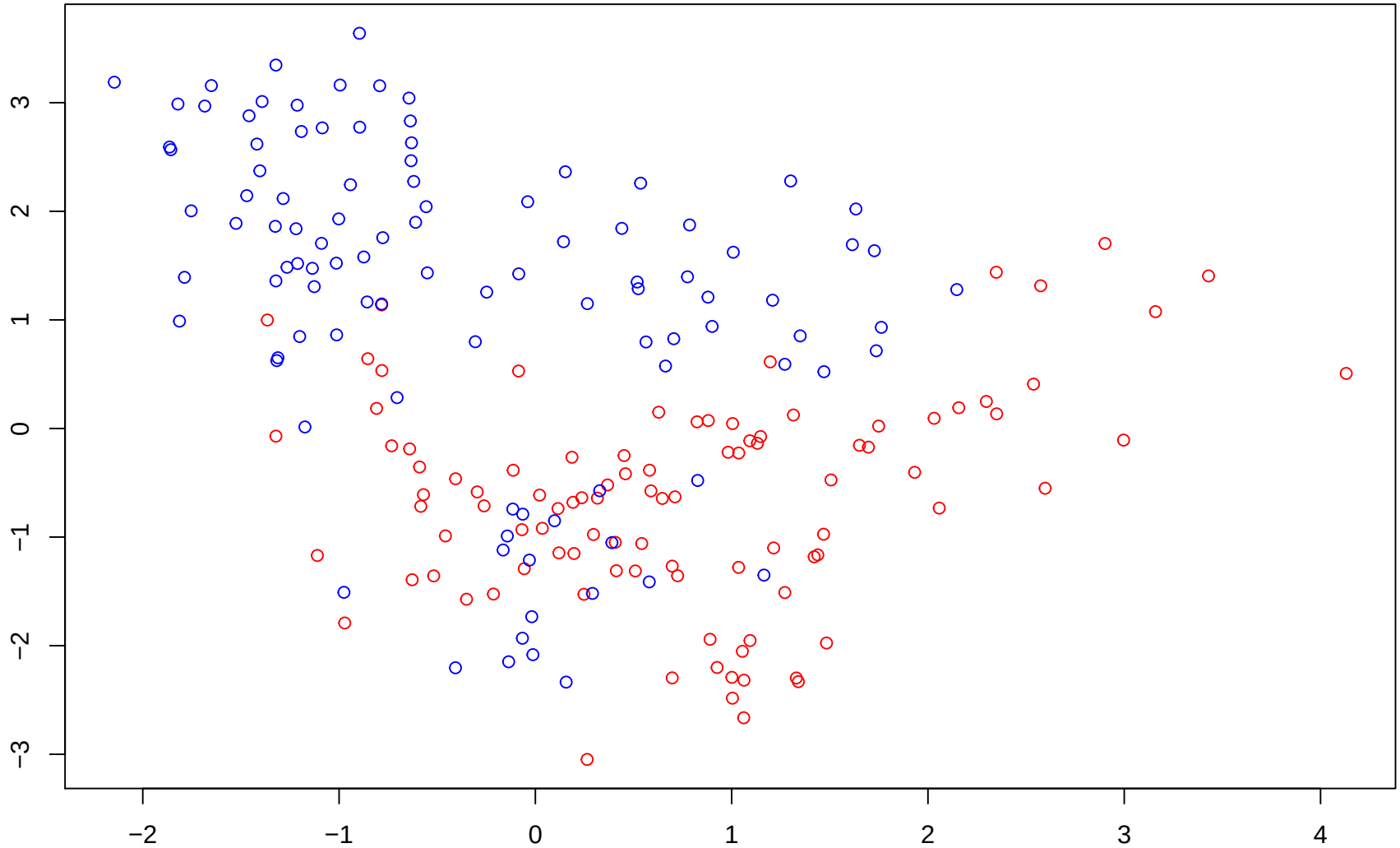
- X est le descriptif du contenu d'un mail (nombre de caractères spéciaux, nombre d'occurrence de certains mots, nombre de majuscules ...).
- Y : est-ce que le mail est un spam ou non ?

Une fonction de prédiction est une fonction définie, mesurable de $\mathcal{X}$ dans $\mathcal{Y}$.

Remarques :

- Si Y est quantitative ($Y \in \mathbb{R}$) on parle de régression.
- Si Y est qualitative (Y à valeurs dans un ensemble fini) on parle de classification supervisée. C'est le cadre de ce cours.
- On parle d'apprentissage supervisé car pour chaque X_i de l'échantillon d'apprentissage on dispose de Y_i, l'étiquette. Au contraire, on parlera d'apprentissage non-supervisé lorsque $\mathcal{D}_n$ est simplement constitué des X_i, $\mathcal{D}_n = \{X_1, \ldots, X_n\}$.

Un exemple introductif



- X la coordonnée d'un point, $\mathcal{X} = \mathbb{R}^2$.
- Y la couleur du point, $\mathcal{Y} = \{0, 1\}$.

Méthode de prédiction basée sur les moindres carrés

- On note $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, le vecteur de $\mathbb{R}^n$ des Y et $\mathbf{X}$ la matrice $n \times p$ (ici $p = 2$) des variables d'entrée.
- On prédit Y par $X^T\hat{\beta}$ où $\hat{\beta}$ est obtenu en minimisant le critère des moindres carrés :

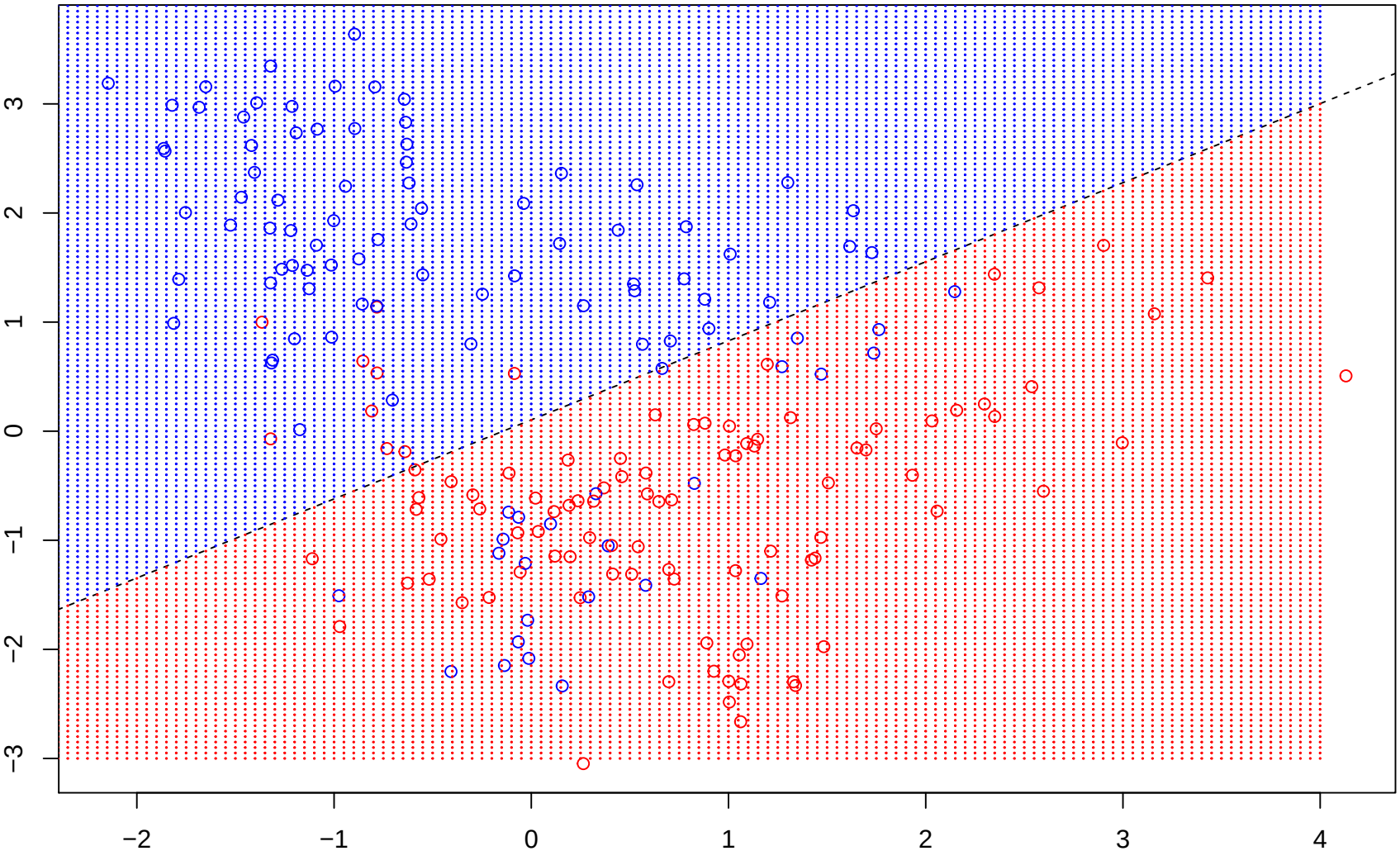
$$\hat{\beta} = \underset{\beta}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

On trouve

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

- La prédiction $X\hat{\beta}$ appartient à $\mathbb{R}$! On décide en fait de prédire Y par $1_{X^T\hat{\beta}>0.5}$
- Plus généralement, on définit le classifieur $g : \mathcal{X} \to [0, 1]$, $g : x \mapsto g(x)$, tel que pour tout $x \in \mathcal{X}$, $g(x) = \mathbb{1}_{x^T\hat{\beta}>0.5}$.

Méthode de prédiction basée sur les moindres carrés



La droite $x^T \hat{\beta} = 0.5$ (en pointillé) représente la frontière de décision. Les points en dessous de cette droite seront classés rouge, ceux au dessus seront classés bleu.

Méthode de prédiction des plus proches voisins

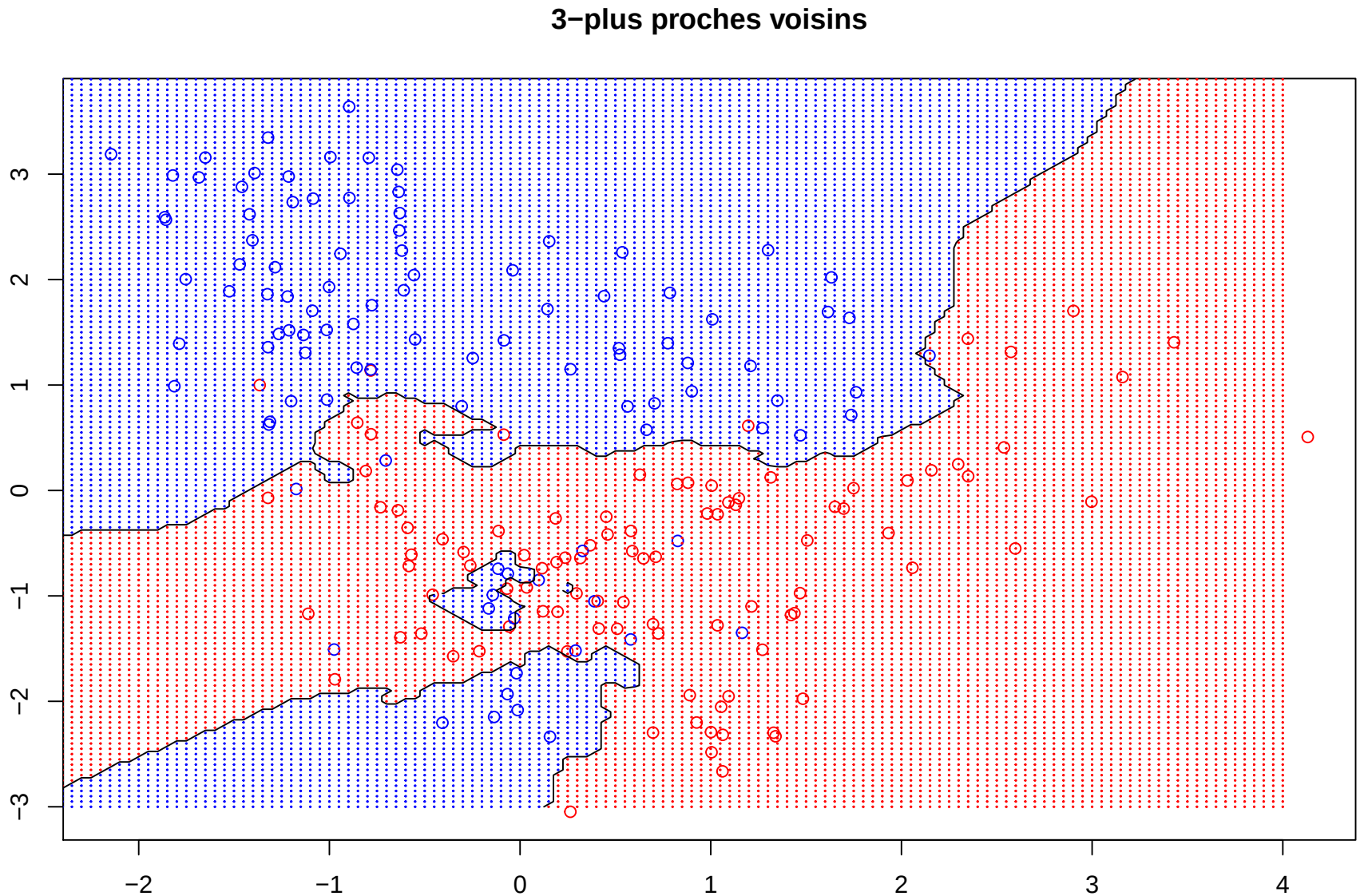
A partir de $x \in \mathcal{X}$ on prédit Y par la méthode des k-plus proches voisins de la façon suivante. On définit tout d'abord :

$$\hat{\eta}_k(x) := \frac{1}{k} \sum_{X_i \in N_k(x)} Y_i,$$

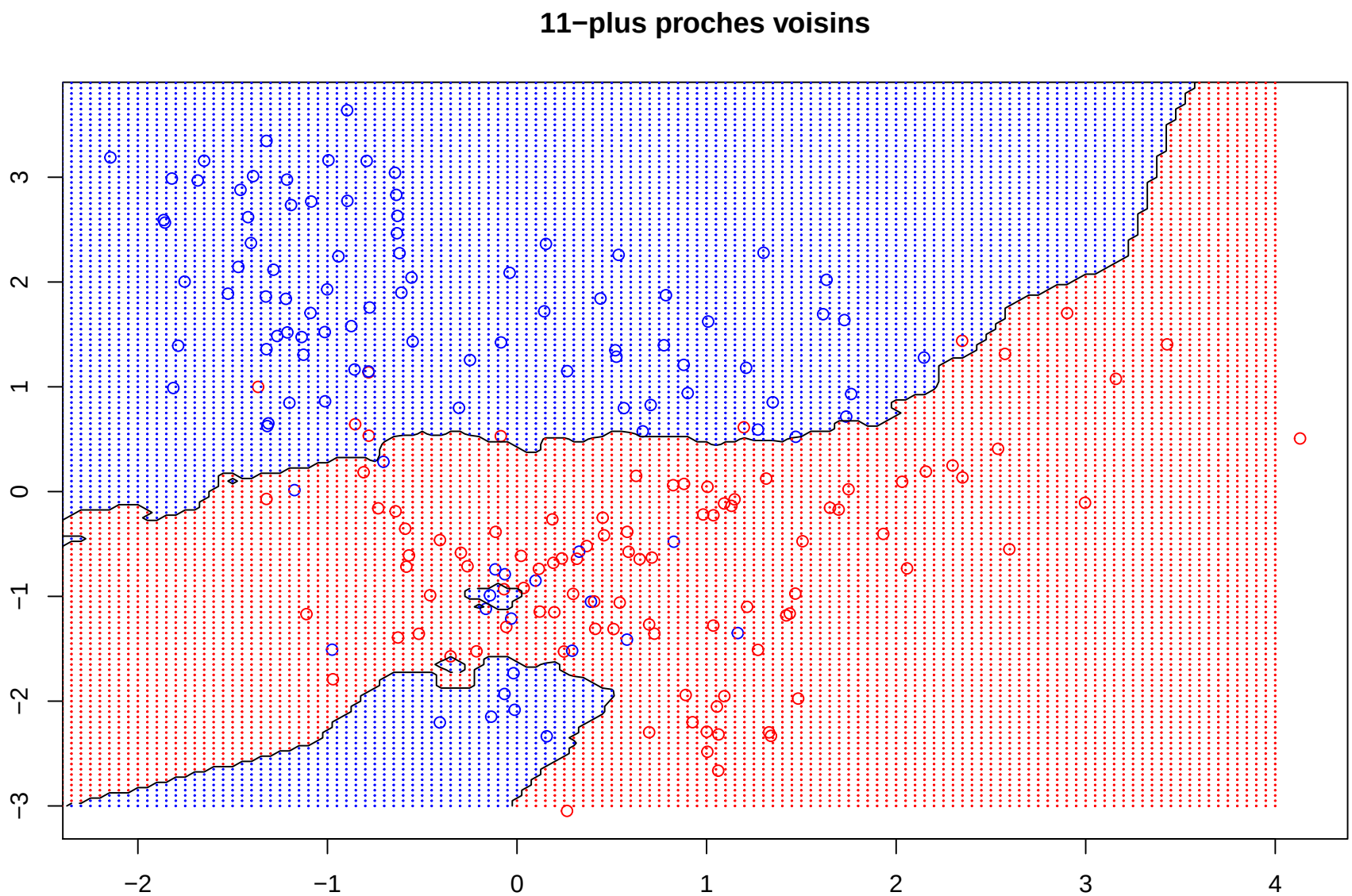
où $N_k(X)$ représente un voisinage de x défini par les k plus proches voisins de x. On utilise généralement la distance euclidienne pour définir le voisinage. Enfin, le classifieur $g : \mathcal{X} \to [0, 1]$ est défini tel que pour tout $x \in \mathcal{X}$, $g(x) = \mathbb{1}_{\hat{\eta}_k(x) > 0.5}$.

Prédiction par les 3 plus proches voisins

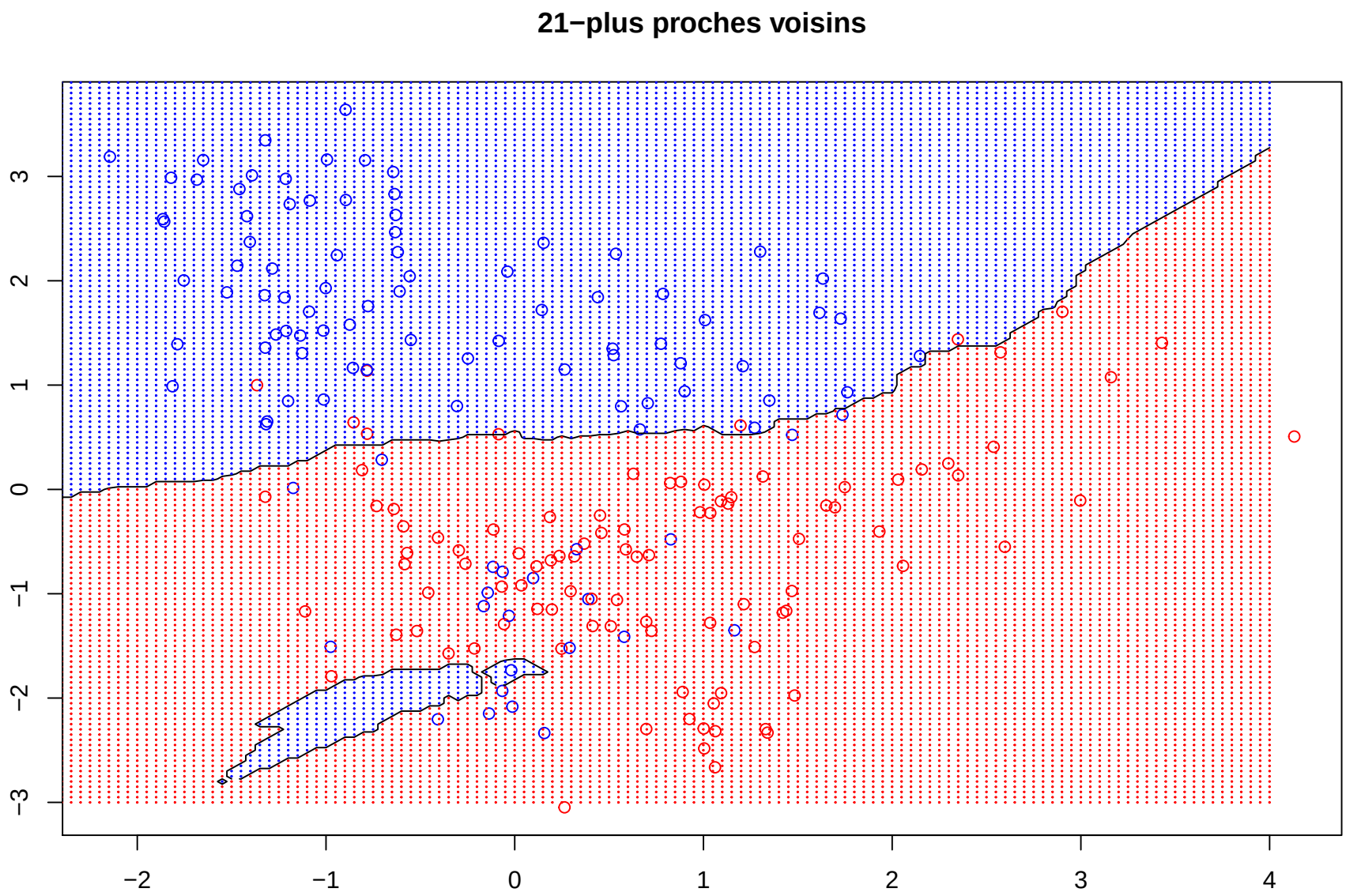
```
## Loading required package: class
```



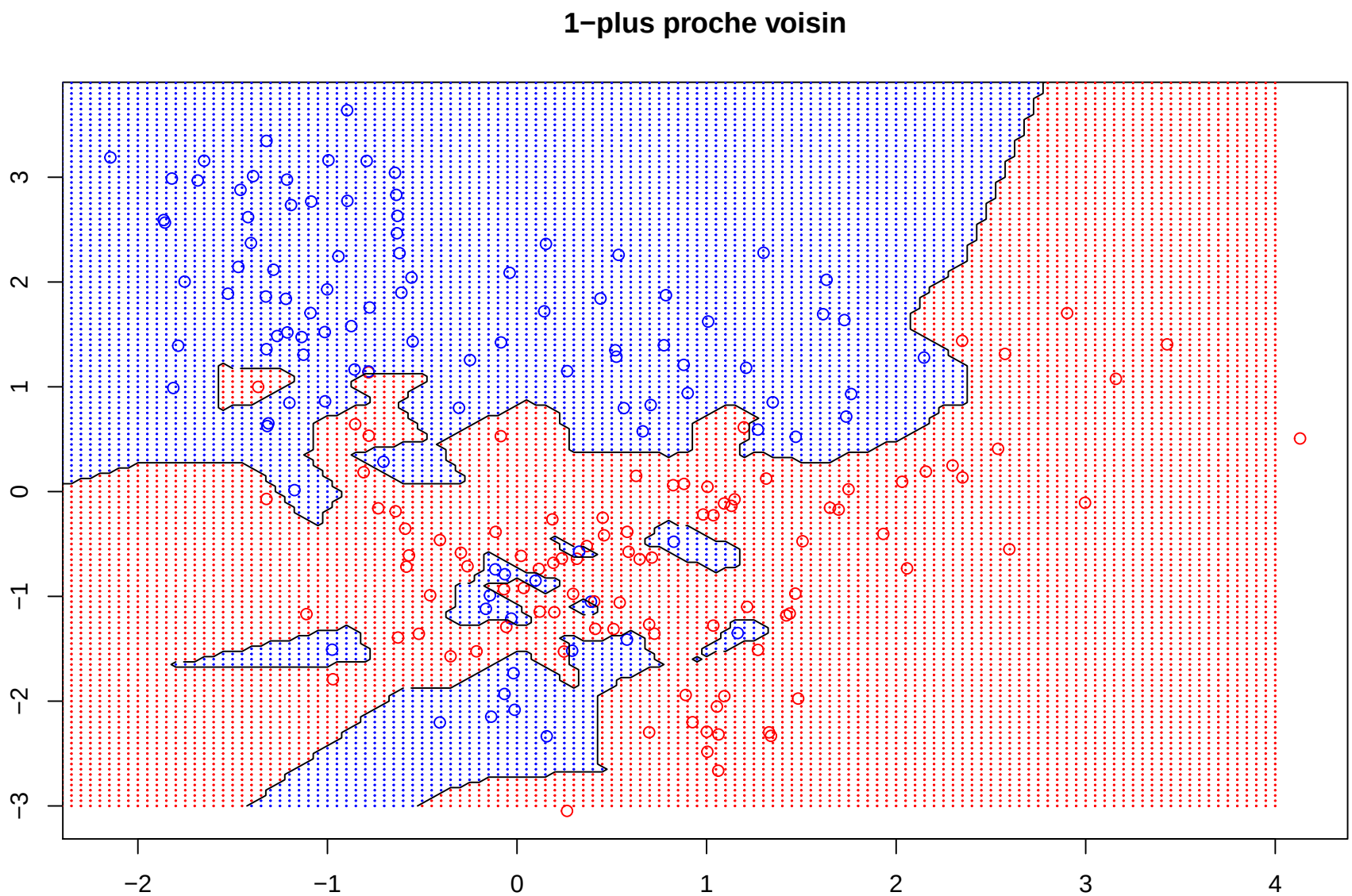
Prédiction par les 11 plus proches voisins



Prédiction par les 21 plus proches voisins



Prédiction par les 1 plus proche voisin



Calcul du taux de mauvaise classification sur données simulées

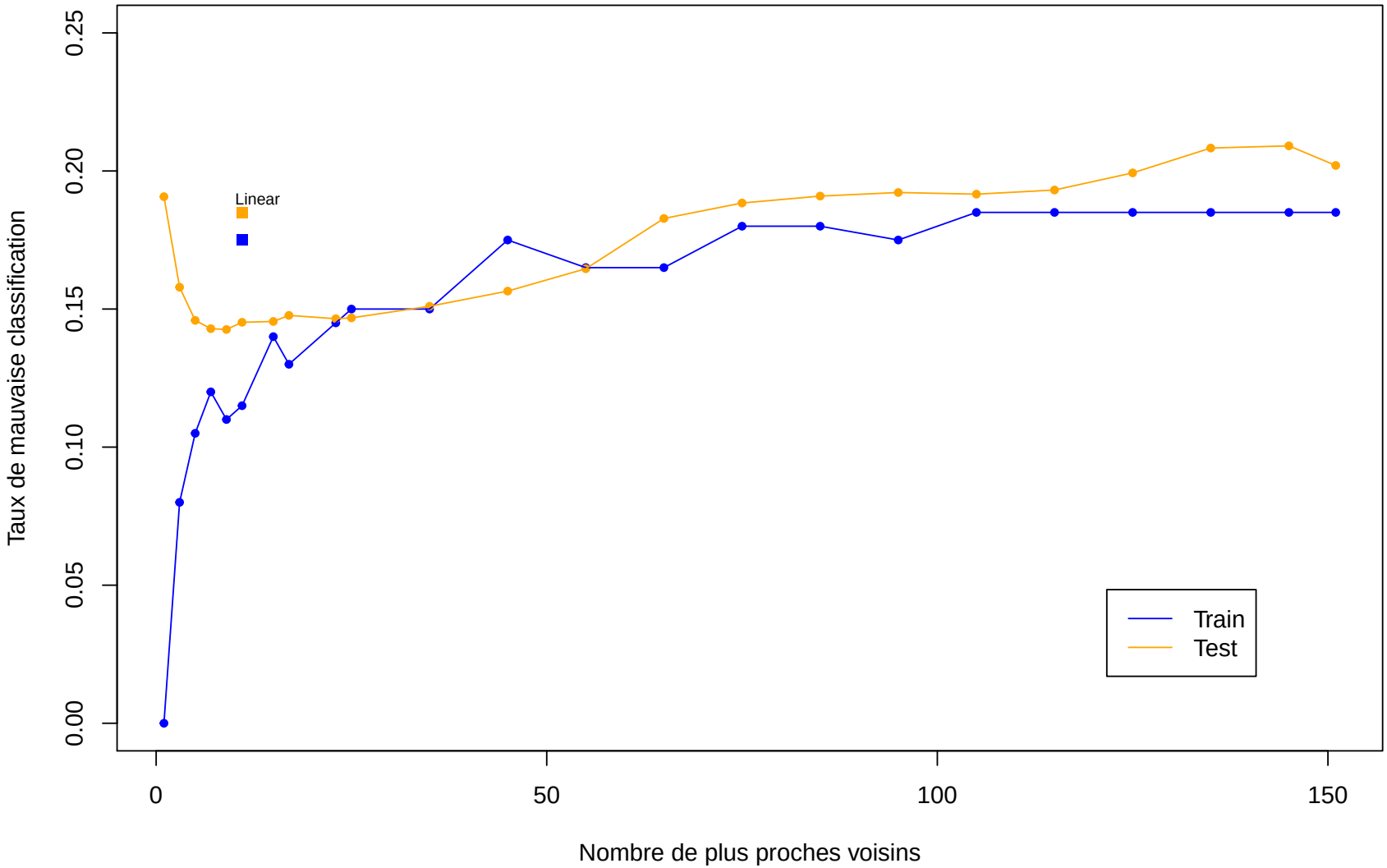
- On définit plusieurs classifieurs : la méthode par moindre carrés, les k-plus proches voisins pour $k = 1, \ldots, 151$.
- On simule un échantillon d'apprentissage de taille $n = 200$.
- On simule un échantillon test de taille 10 000.
- On calcule le taux de mauvaise classification sur l'échantillon d'apprentissage et sur l'échantillon de test. Sur l'échantillon test : $\mathcal{D}_n^{\text{test}} = \{(X_1^{\text{test}}, Y_1^{\text{test}}), \ldots, (X_N^{\text{test}}, Y_N^{\text{test}})\}$, on calcule :

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{Y_i^{\text{test}} \neq g(X_i^{\text{test}})}$$

où g représente un classifieur.

- On compare les différentes méthodes.

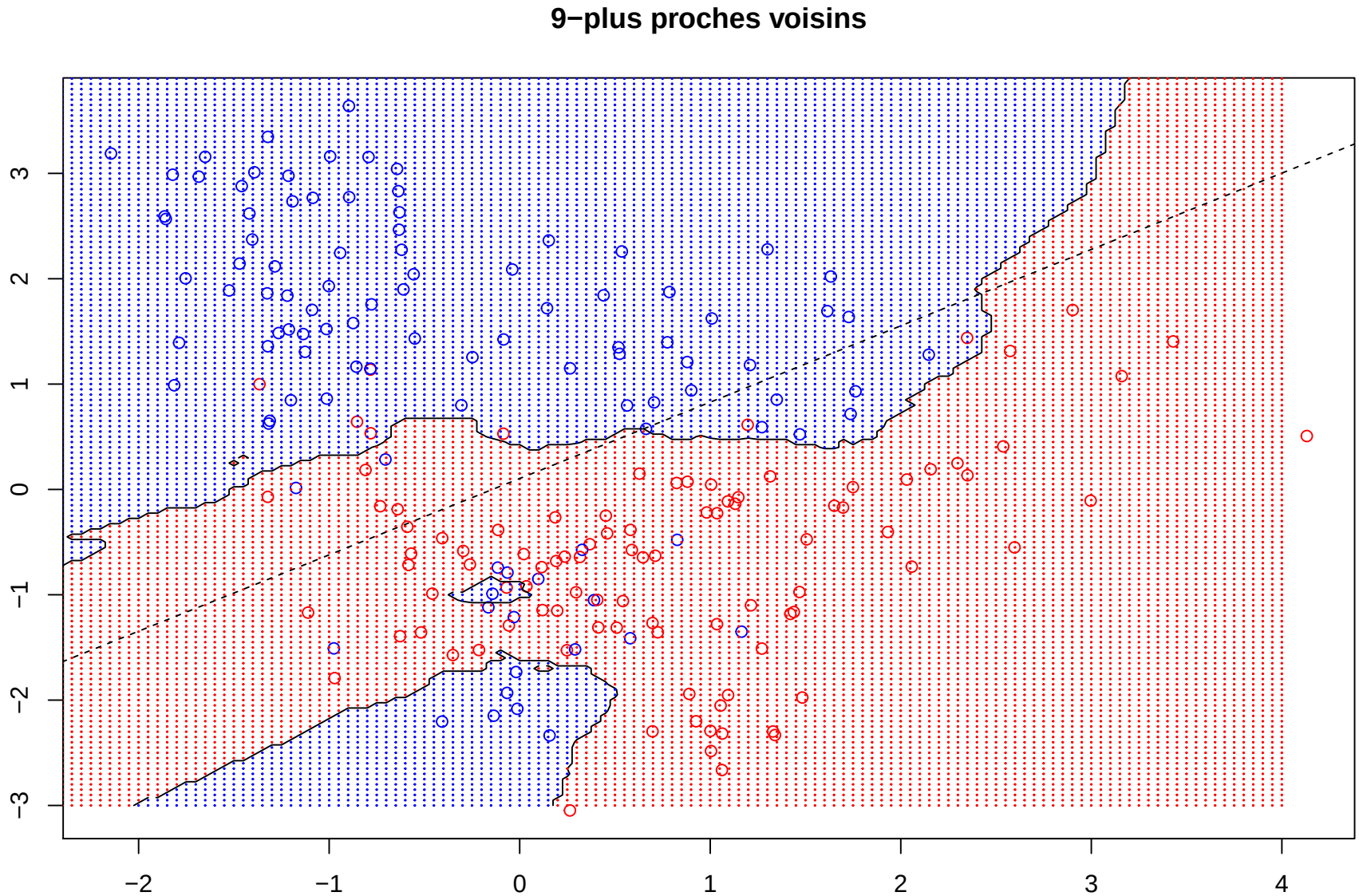
Calcul du taux de mauvaise classification sur données simulées



Calcul du taux de mauvaise classification sur données simulées

- le modèle linéaire donne un taux de mauvaise classification de 0.175 sur l'échantillon d'apprentissage et de 0.185 sur l'échantillon test.
- les 1 plus proches voisins donne un taux de mauvaise classification de 0 sur l'échantillon d'apprentissage et de 0.1907 sur l'échantillon test.
- le choix optimal de k dans la méthode des plus proches voisins, basé sur le taux de mauvaise classification (sur l'échantillon test), est $k = 9$. Pour cette valeur de k, le taux de mauvaise classification est de 0.1426.

Prédiction par les 9 plus proches voisins et comparaison avec la méthode des moindres carrés



Notions de risque, de prédicteur et classifieur oracles - problème de régression

Rappel : dans un problème de régression, on a $X \in \mathcal{X}$, $Y \in \mathbb{R}$. On considère un prédicteur $f : \mathcal{X} \to \mathbb{R}$ et on suppose généralement que Y est de carré intégrable

$$\mathbb{E}[Y^2] < \infty$$

- On appelle fonction de perte L_2 la fonction ℓ telle que :

$$\forall y, y' \in \mathbb{R}, \ell(y, y') = (y - y')^2$$

- On définit le risque L_2

$$R(f) = \mathbb{E}[\ell(Y, f(X))] = \mathbb{E}[(Y - f(X))^2]$$

Notions de risque, de prédicteur et d'oracle dans le cadre de la régression

Rappels : dans un problème de régression, on a $X \in \mathcal{X}$, $Y \in \mathbb{R}$. On considère un prédicteur $f : \mathcal{X} \to \mathbb{R}$ que l'on suppose appartenir à l'ensemble des fonctions de $\mathcal{X}$ dans $\mathbb{R}$, noté $\mathcal{F}(\mathcal{X}, \mathbb{R})$. On suppose également généralement que Y est de carré intégrable

$$\mathbb{E}[Y^2] < \infty$$

- On appelle fonction de perte L_2 la fonction ℓ telle que :

$$\forall y, y' \in \mathbb{R}, \ell(y, y') = (y - y')^2$$

- On définit le risque L_2

$$R(f) = \mathbb{E}[\ell(Y, f(X))] = \mathbb{E}[(Y - f(X))^2]$$

Remarque : lorsque la perte considérée est la perte L_2 on parle de régression des moindres carrés.

Notions de risque, de prédicteur et d'oracle dans le cadre de la régression

Le prédicteur oracle est défini par

$$f^* \in \underset{f \in \mathcal{F}(\mathcal{X}, \mathbb{R})}{\operatorname{argmin}} R(f)$$

f^* satisfait $R(f^*) \leq R(f)$, $\forall f \in \mathcal{F}(\mathcal{X}, \mathbb{R})$.

Proposition : On a

$$f^*(x) = \mathbb{E}[Y \mid X = x], \forall x \in \mathcal{X}$$

et $\forall f \in \mathcal{F}(\mathcal{X}, \mathbb{R})$,

$$0 \leq R(f) - R(f^*) = \mathbb{E}[(f(X) - f^*(X))^2].$$

Notions de risque, de classifieur et d'oracle dans le cadre de la classification

Dans un problème de classification, on a $X \in \mathcal{X}$, $Y \in \{0,1\}$. On considère un prédicteur $g : \mathcal{X} \to \{0,1\}$ que l'on suppose appartenir à l'ensemble des classifieur noté $\mathcal{G}$.

- On appelle fonction de perte $0-1$ ou perte de mauvaise classification la fonction ℓ telle que :

$$\forall y, y' \in \{0,1\}, \ell(y,y') = \mathbb{1}_{y \neq y'}$$

- Le risque associé à un classifieur g est

$$R(g) = \mathbb{E}[\ell(Y, g(X))] = \mathbb{E}[\mathbb{1}_{Y \neq g(X)}] = \mathbb{P}[Y \neq g(X)]$$

Remarque : $Y \mid X \sim \mathcal{B}(\eta(X))$ avec

$$\eta(X) = \mathbb{E}[Y \mid X] = \mathbb{P}[Y = 1 \mid X]$$

Notions de risque, de classifieur et d'oracle dans le cadre de la classification

Le classifieur oracle est défini par

$$g^* \in \underset{g \in \mathcal{G}}{\operatorname{argmin}} R(g)$$

On l'appelle classifieur de Bayes et g^* satisfait $R(g^*) \leq R(g)$, $\forall g \in \mathcal{G}$.

Proposition : Soit

$$g^*(x) = \mathbb{1}_{\eta(x) \geq 1/2},$$

on a alors $\forall g \in \mathcal{G}$,

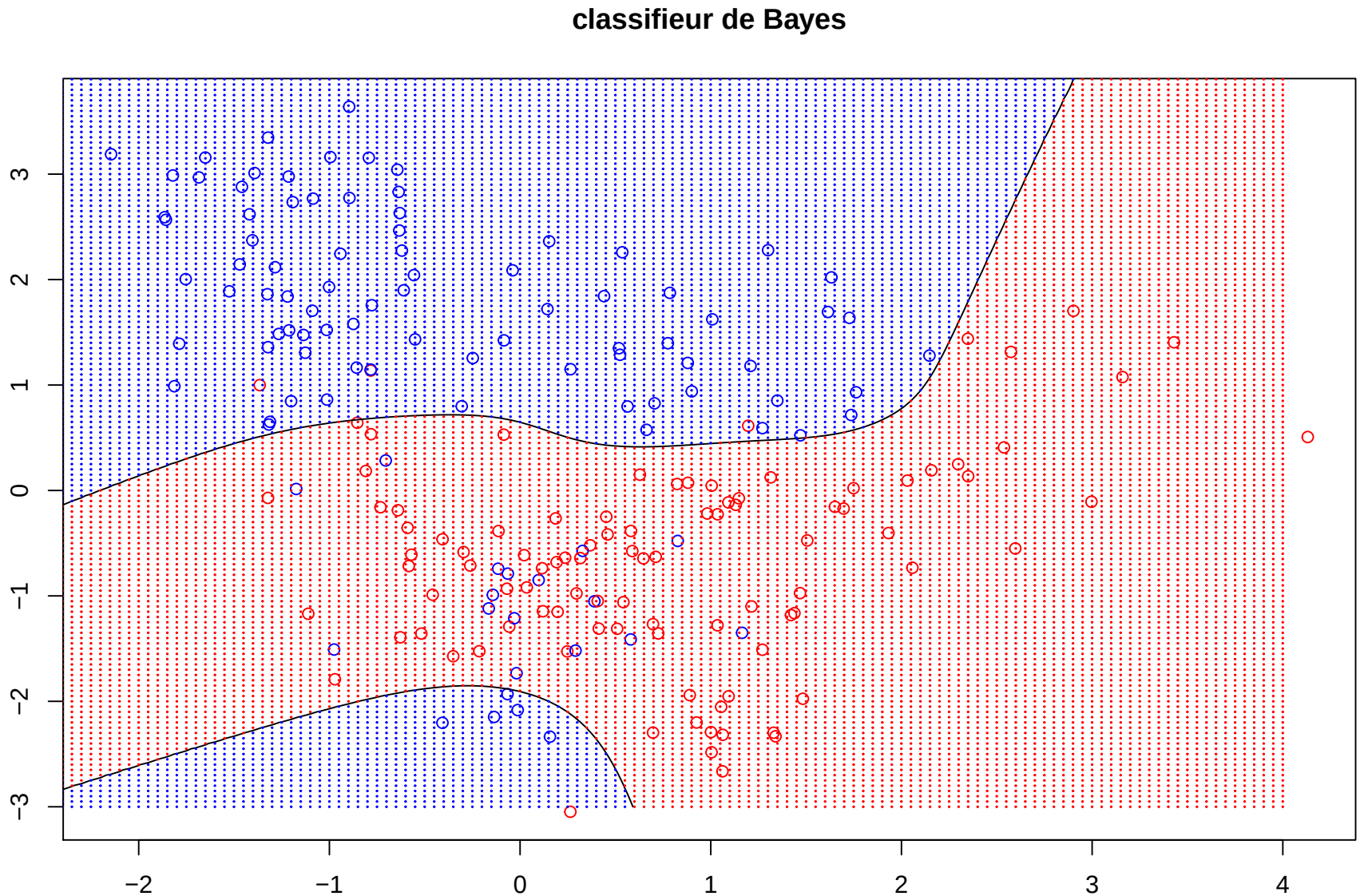
(1)
$$R(g^*) = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}] \leq R(g)$$

(2)
$$0 \leq R(g) - R(g^*) = \mathbb{E}[|2\eta(X) - 1| \mathbb{1}_{g^*(X) \neq g(X)}].$$

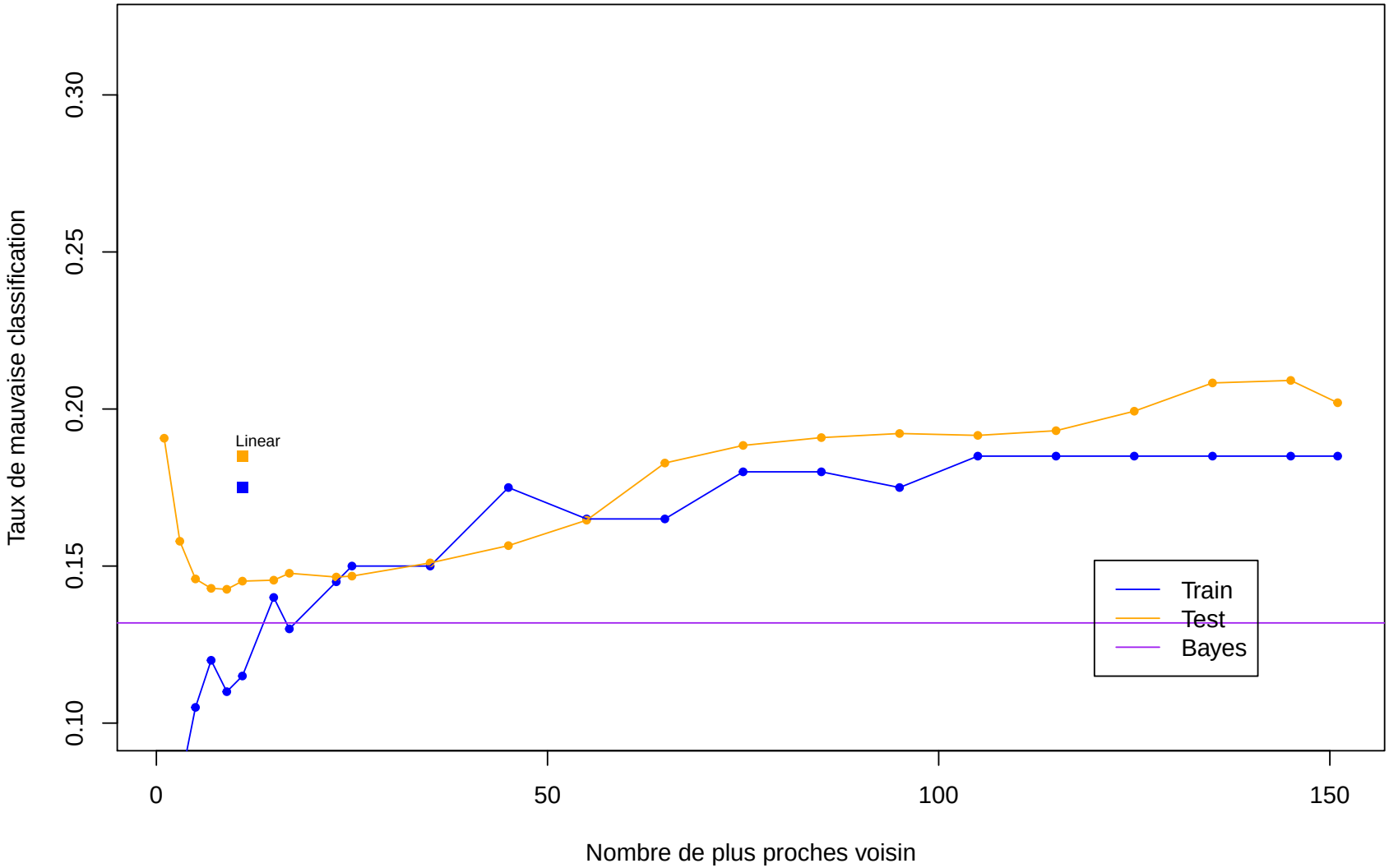
Preuve en cours.

Retour aux données simulées - classifieur de Bayes

```
## Loading required package: mvtnorm
```



Calcul du risque de mauvaise classification avec le classifieur de Bayes



Le taux de mauvaise classification sur l'échantillon test pour le classifieur de Bayes est égal à 0.1319.

Retour sur le classifieur de Bayes

En pratique, on ne connait pas $\eta(X) = \mathbb{P}[Y = 1 \mid X]$ et on ne peut donc pas calculer le classifieur de Bayes. Mais on peut estimer η.

- On suppose que l'on connait $\tilde{\eta} : \mathcal{X} \to [0, 1]$. ($\tilde{\eta}$ peut être vu comme une approximation de η) et on considère $\tilde{g}(x) = \mathbb{1}_{\tilde{\eta}(x) \geq 1/2}$.
- On a alors :

$$0 \leq R(\tilde{g}) - R(g^*) \leq 2\sqrt{\mathbb{E}\left[(\tilde{\eta}(X) - \eta(X))^2\right]}$$

Preuve en cours.

Consistance d'un algorithme de classification

Un algorithme de classification $\hat{g} = \{\hat{g}_n, n \geq 1\}$ est une suite de fonctions $\hat{g}_n : \mathcal{X} \times \mathcal{Y} \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, $\hat{g}_n : \mathcal{D}_n \mapsto \hat{g}_n(\mathcal{D}_n)$. C'est donc une fonction que l'on applique à des données et qui renvoie une règle de classification.

Remarques :

- comme $\mathcal{D}_n$ est aléatoire, $\hat{g}_n(\mathcal{D}_n)(x)$ est également aléatoire quelque soit $x \in \mathcal{X}$.
- on a

$$R(\hat{g}_n(\mathcal{D}_n)) := \mathbb{E}[\ell(Y, \hat{g}_n(\mathcal{D}_n)(X)) \mid \mathcal{D}_n] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, \hat{g}_n(\mathcal{D}_n)(x)) d\mathbb{P}(x, y)$$

qui est également aléatoire. Nous pouvons donc considérer l'espérance de cette variable aléatoire par rapport à $\mathcal{D}_n \overset{\text{iid}}{\sim} \mathbb{P}$, notée $\mathbb{E}[R(\hat{g}_n(\mathcal{D}_n))]$.

Consistance d'un algorithme de classification

Définition : on dit qu'un algorithme d'apprentissage est consistant par rapport à la loi $\mathbb{P}$ si et seulement si

$$\mathbb{E}[R(\hat{g}_n(\mathcal{D}_n))] \underset{n\to\infty}{\longrightarrow} R(g^*)$$

- On dira qu'un algorithme d'apprentissage est consistant par rapport à une famille de lois $\mathcal{P}$ si et seulement si il est consistant par rapport à tout $\mathbb{P} \in \mathcal{P}$.
- On dira qu'un algorithme d'apprentissage est universellement consistant si et seulement si il est consistant par rapport à toute probabilité $\mathbb{P}$ sur $\mathcal{X} \times \mathcal{Y}$.

Exemple d'algorithme consistant

On considère l'exemple où $\mathcal{X}$ est un ensemble fini, avec $\text{card}(\mathcal{X}) = K$. On étudie alors l'algorithme de minimisation du risque empirique. Soit

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Y_i}$$

Remarques :

- On a $\mathbb{E}[\hat{R}(g)] = \mathbb{P}[g(X) \neq Y] = R(g)$
- Par ailleurs, $\hat{R}(g) \underset{n \to \infty}{\longrightarrow} R(g)$ p.s. d'après la loi des grands nombres.

On considère alors $\hat{g}_n(\mathcal{D}_n)$ défini par $\hat{g}_n(\mathcal{D}_n) \in \underset{g \in \mathcal{G}}{\text{argmin}}\ \hat{R}(g)$. On espère que $\hat{g}_n(\mathcal{D}_n)$ soit une "bonne'' approximation de $g^* \in \underset{g \in \mathcal{G}}{\text{argmin}}\ R(g)$.

Exemple d'algorithme consistant - ingrédients de preuve

On note $\mathcal{G} = \{g : \mathcal{X} \to \{0,1\}\}$ l'ensemble des classifieurs. On a $\text{card}(\mathcal{G}) = 2^K$.

On définit l'excès de risque pour un classifieur $g \in \mathcal{G}$ par :

$$\mathcal{E}(g) = R(g) - R(g^*) \geq 0.$$

On a : $\mathcal{E}(\hat{g}(\mathcal{D}_n)) = R(\hat{g}(\mathcal{D}_n)) - R(g^*)$.

Attention : $\mathcal{E}(\hat{g}(\mathcal{D}_n))$ est une variable aléatoire qui dépend de $\mathcal{D}_n$.

- **Etape 1** : Décomposition de l'excès de risque

$$\mathcal{E}(\hat{g}(\mathcal{D}_n)) = R(\hat{g}(\mathcal{D}_n)) - \hat{R}(\hat{g}(\mathcal{D}_n)) + \hat{R}(\hat{g}(\mathcal{D}_n)) - R(g^*)$$

Attention : $\hat{R}(\hat{g}(\mathcal{D}_n)) = \sum_{i=1}^n \mathbb{1}_{\hat{g}(\mathcal{D}_n)(X_i) \neq Y_i}/n$ et $\mathbb{E}[\hat{R}(\hat{g}(\mathcal{D}_n))] \neq \mathbb{E}[R(\hat{g}(\mathcal{D}_n))]$ car ici (X_i, Y_i) et $\hat{g}(\mathcal{D}_n)$ ne sont pas indépendants !

On a :

$$\mathcal{E}(\hat{g}(\mathcal{D}_n)) \leq R(\hat{g}(\mathcal{D}_n)) - \hat{R}(\hat{g}(\mathcal{D}_n)) + \hat{R}(g^*) - R(g^*) \leq 2 \max_{g \in \mathcal{G}} |\hat{R}(g) - R(g)|$$

Exemple d'algorithme consistant - ingrédients de preuve

- **Etape 2** : Contrôle des déviations de $\max_{g \in \mathcal{G}} |\hat{R}(g) - R(g)|$

De l'étape 1 on a donc

$$\mathbb{E}[\mathcal{E}(\hat{g}(\mathcal{D}_n))] \leq 2\,\mathbb{E}[\max_{g \in \mathcal{G}} |\hat{R}(g) - R(g)|]$$

En utilisant Fubini-Tonelli, on écrit :

$$\mathbb{E}[\max_{g \in \mathcal{G}} |\hat{R}(g) - R(g)|] = \int_0^\infty \mathbb{P}[\max_{g \in \mathcal{G}} |\hat{R}(g) - R(g)| \geq t]\, dt$$

On cherche ensuite à contrôler les déviations du processus empirique $\max_{g \in \mathcal{G}} |\hat{R}(g) - R(g)|$. On utilise pour cela l'inégalité de Hoeffding. Preuve à terminer en cours.

Annexe : code ayant servi à générer les données

```
set.seed(30)
mred<-cbind(rnorm(10,1,1),rnorm(10,0,1))
mblue<-cbind(rnorm(10,0,1),rnorm(10,1,1))

#Generation des observations pour l'echantillon d'apprentissage
ntrain<-200
p<-0.5
Ytrain<-rbinom(ntrain,1,p)
n1train<-sum(Ytrain)
mured<-sample(1:10,size=n1train,replace=TRUE)#generation des moyennes
#pour chaque observation du groupe 1 (rouge)
mublue<-sample(1:10,size=(ntrain-n1train),replace=TRUE)#generation
#des moyennes pour chaque observation du groupe 2 (bleu)

Xtrain_red<-t(apply(mred[mured,],1,function(x)
      c(rnorm(1,x[1],1/sqrt(5)),rnorm(1,x[2],1/sqrt(5)))))
Xtrain_blue<-t(apply(mblue[mublue,],1,function(x)
      c(rnorm(1,x[1],1/sqrt(5)),rnorm(1,x[2],1/sqrt(5)))))
Xtrain=rbind(Xtrain_red,Xtrain_blue)
grouptrain<-c(rep(1,n1train),rep(0,ntrain-n1train))
```