Fondamentaux de l'apprentissage statistique

Sylvain Arlot

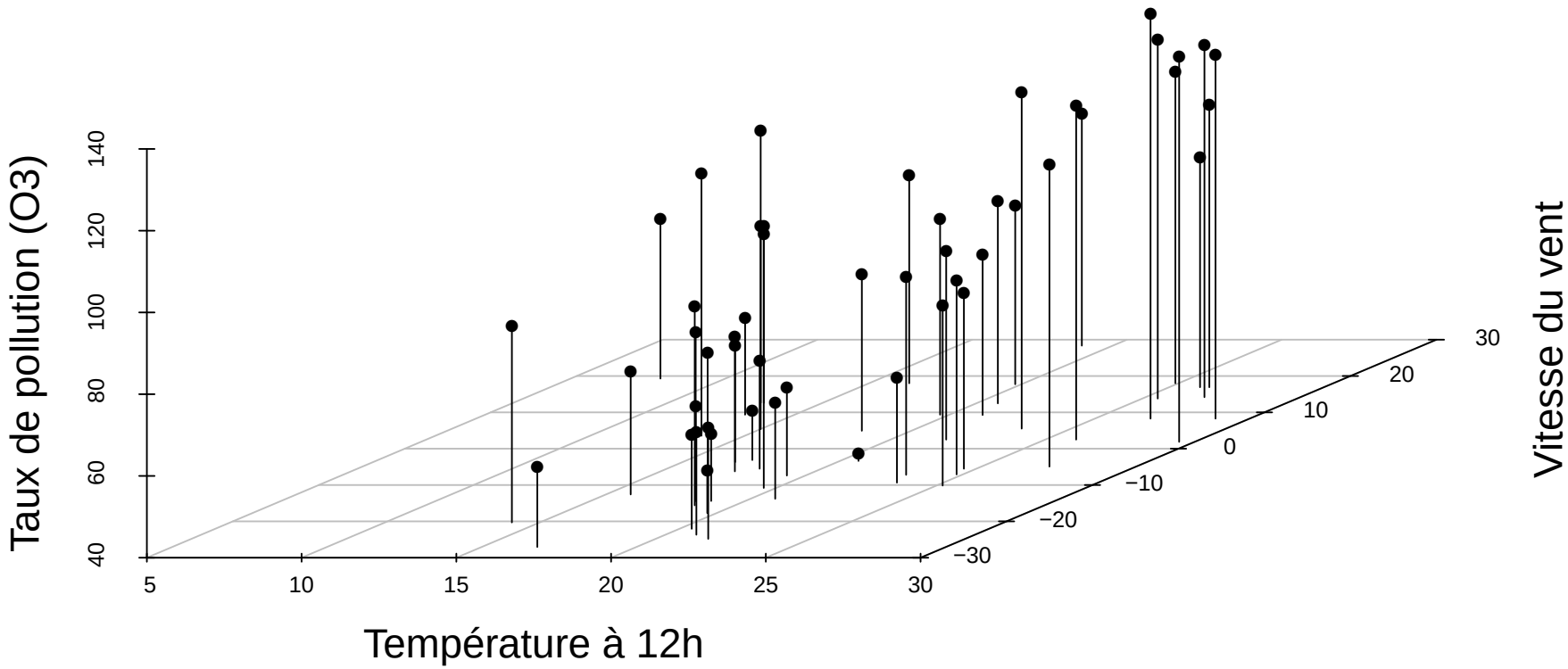
Laboratoire de Mathématiques d'Orsay, Université Paris-Sud

JES 2016, Fréjus
3 Octobre 2016

Plan

1. Prévision
2. Régression et classification
3. Minimisation du risque empirique
4. Moyennes locales
5. On n'a rien sans rien
6. Conclusion : enjeux de l'apprentissage

Exemple : taux de pollution (régression)



Nouvelle observation (T12, Vx) $\Rightarrow$ taux de pollution ?

Figure : [Cornillon and Matzner-Løber, 2011]

Reconnaissance de chiffres manuscrits (MNIST)

 $\Rightarrow$?

http://yann.lecun.com/exdb/mnist/

Reconnaissance d'objets

Drapeau américain :      ...

Papillon :      ...

Ours en peluche :      ...

⇒ Drapeau américain ? Papillon ?
Ours en peluche ? ...

http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Nombreux autres exemples

- Bioinformatique :
  - données de séquençage $\Rightarrow$ diagnostic et pronostic (cancer, ...)
  - médecine personnalisée
  - ...
- Classification de texte :
  - Détection de spams
  - Publicité en ligne
  - Classification automatique de documents
- Reconnaissance d'actions humaines dans des vidéos
- Reconnaissance de parole
- Évaluation des risques-clients (prêt bancaire)
- ...

Formalisation : problème de prévision

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

Formalisation : problème de prévision

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$
  ($\mathcal{F}$: ensemble des prédicteurs)
  Nouvelle observation $X_{n+1} \Rightarrow f(X_{n+1})$ « prévoit » Y_{n+1}

Formalisation : problème de prévision

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$
  ($\mathcal{F}$: ensemble des prédicteurs)
  Nouvelle observation $X_{n+1} \Rightarrow f(X_{n+1})$ « prévoit » Y_{n+1}

- Mesure de qualité : fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty[$
  Risque (erreur de prévision) : $\mathcal{R}_P(f) = \mathbb{E}\Big[c\big(f(X), Y\big)\Big]$

Formalisation : problème de prévision

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$
  ($\mathcal{F}$: ensemble des prédicteurs)
  Nouvelle observation $X_{n+1} \Rightarrow f(X_{n+1})$ « prévoit » Y_{n+1}

- Mesure de qualité : fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty[$
  Risque (erreur de prévision) : $\mathcal{R}_P(f) = \mathbb{E}\Big[c\big(f(X), Y\big)\Big]$

- En résumé : avec D_n uniquement, on cherche un prédicteur $f \in \mathcal{F}$ tel que $\mathcal{R}_P(f)$ est minimal.

Formalisation : problème de régression

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt (Régression : $\mathcal{Y} = \mathbb{R}$)
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$
  ($\mathcal{F}$: ensemble des prédicteurs)
  Nouvelle observation $X_{n+1} \Rightarrow f(X_{n+1})$ « prévoit » Y_{n+1}

- Mesure de qualité : fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty[$
  Risque (erreur de prévision) : $\mathcal{R}_P(f) = \mathbb{E}\Big[c\big(f(X), Y\big)\Big]$
  Ex. en régression : $c(y, y') = (y - y')^2$

- En résumé : avec D_n uniquement, on cherche un prédicteur $f \in \mathcal{F}$ tel que $\mathcal{R}_P(f)$ est minimal.

Formalisation : problème de classification binaire supervisée

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt (Classif. binaire : $\mathcal{Y} = \{0, 1\}$)
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$
  ($\mathcal{F}$: ensemble des prédicteurs)
  Nouvelle observation $X_{n+1} \Rightarrow f(X_{n+1})$ « prévoit » Y_{n+1}

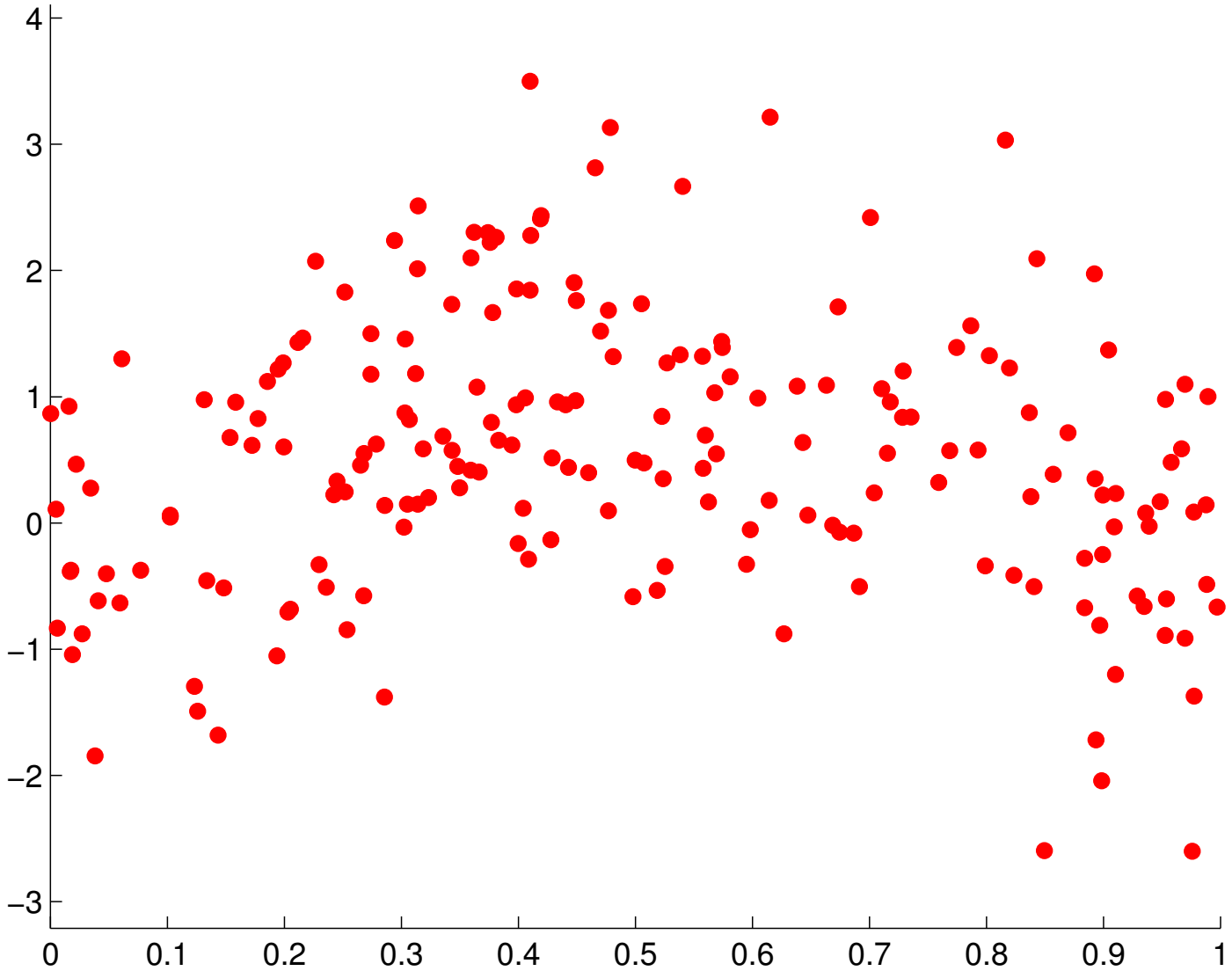
- Mesure de qualité : fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty[$
  Risque (erreur de prévision) : $\mathcal{R}_P(f) = \mathbb{E}\Big[c(f(X), Y)\Big]$
  Ex. en classification : $c(y, y') = \mathbb{1}_{y \neq y'}$

- En résumé : avec D_n uniquement, on cherche un prédicteur $f \in \mathcal{F}$ tel que $\mathcal{R}_P(f)$ est minimal.

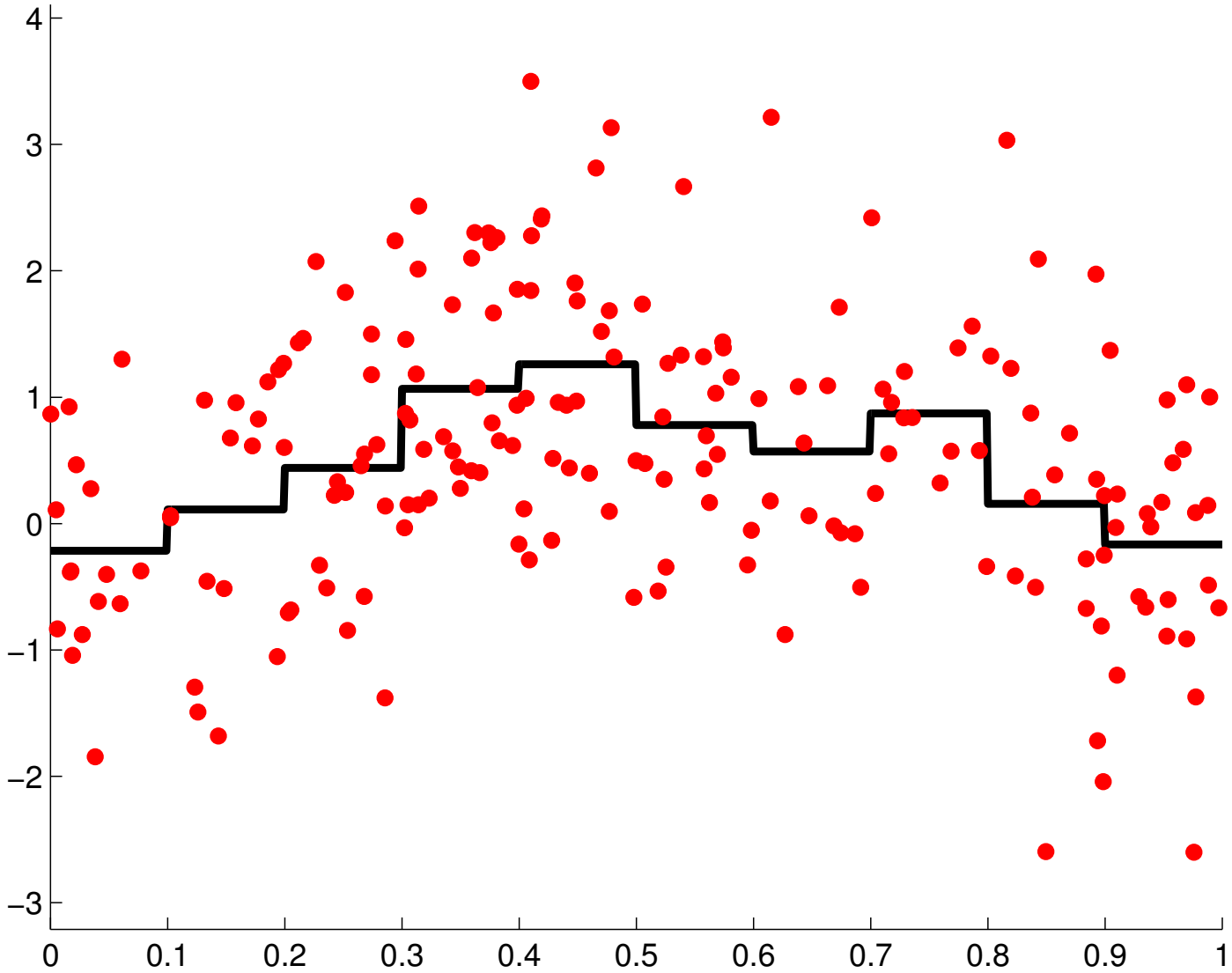
Prévision idéale

- Risque de Bayes : $\mathcal{R}_P^\star := \inf_{f \in \mathcal{F}} \mathcal{R}_P(f)$
- Prédicteur de Bayes : $f^\star \in \operatorname{argmin}_{f \in \mathcal{F}} \{\mathcal{R}_P(f)\}$
- Excès de risque : $\ell(f^\star, f) = \mathcal{R}_P(f) - \mathcal{R}_P^\star \geqslant 0$

Et avec les données uniquement ?



Règle d'apprentissage par partition (régression)



Règles d'apprentissage

- Règle d'apprentissage : $\widehat{f} : \bigcup_{n \geqslant 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$
  Entrée : un échantillon D_n (de taille quelconque $n \geqslant 1$)
  Sortie : un prédicteur $\widehat{f}(D_n) : \mathcal{X} \to \mathcal{Y}$

Règles d'apprentissage

- Règle d'apprentissage : $\widehat{f} : \bigcup_{n \geqslant 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$
  Entrée : un échantillon D_n (de taille quelconque $n \geqslant 1$)
  Sortie : un prédicteur $\widehat{f}(D_n) : \mathcal{X} \to \mathcal{Y}$

- Risque conditionnel :

$$\mathcal{R}_P(\widehat{f}(D_n)) = \mathbb{E}\Big[c(\widehat{f}(D_n; X), Y) \,\Big|\, D_n\Big]$$

- Risque moyen :

$$\mathbb{E}\Big[\mathcal{R}_P(\widehat{f}(D_n))\Big] = \mathbb{E}\Big[c(\widehat{f}(D_n; X), Y)\Big]$$

Règles d'apprentissage

- Règle d'apprentissage : $\widehat{f} : \bigcup_{n \geqslant 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$
  Entrée : un échantillon D_n (de taille quelconque $n \geqslant 1$)
  Sortie : un prédicteur $\widehat{f}(D_n) : \mathcal{X} \to \mathcal{Y}$

- Risque conditionnel :

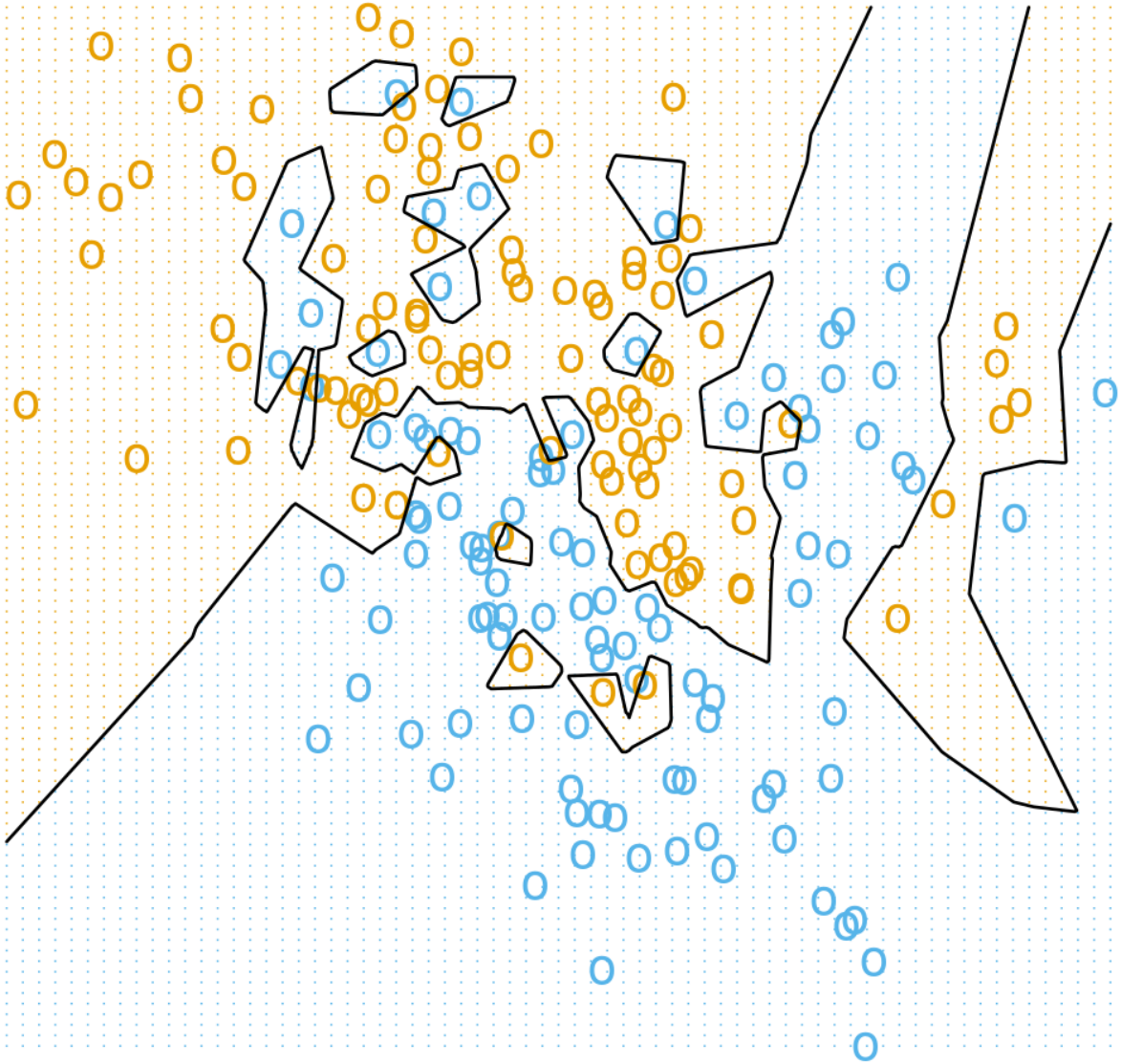
$$\mathcal{R}_P(\widehat{f}(D_n)) = \mathbb{E}\Big[c(\widehat{f}(D_n; X), Y) \,\Big|\, D_n\Big]$$

- Risque moyen :

$$\mathbb{E}\Big[\mathcal{R}_P(\widehat{f}(D_n))\Big] = \mathbb{E}\Big[c(\widehat{f}(D_n; X), Y)\Big]$$

- Exemple : règle par partition

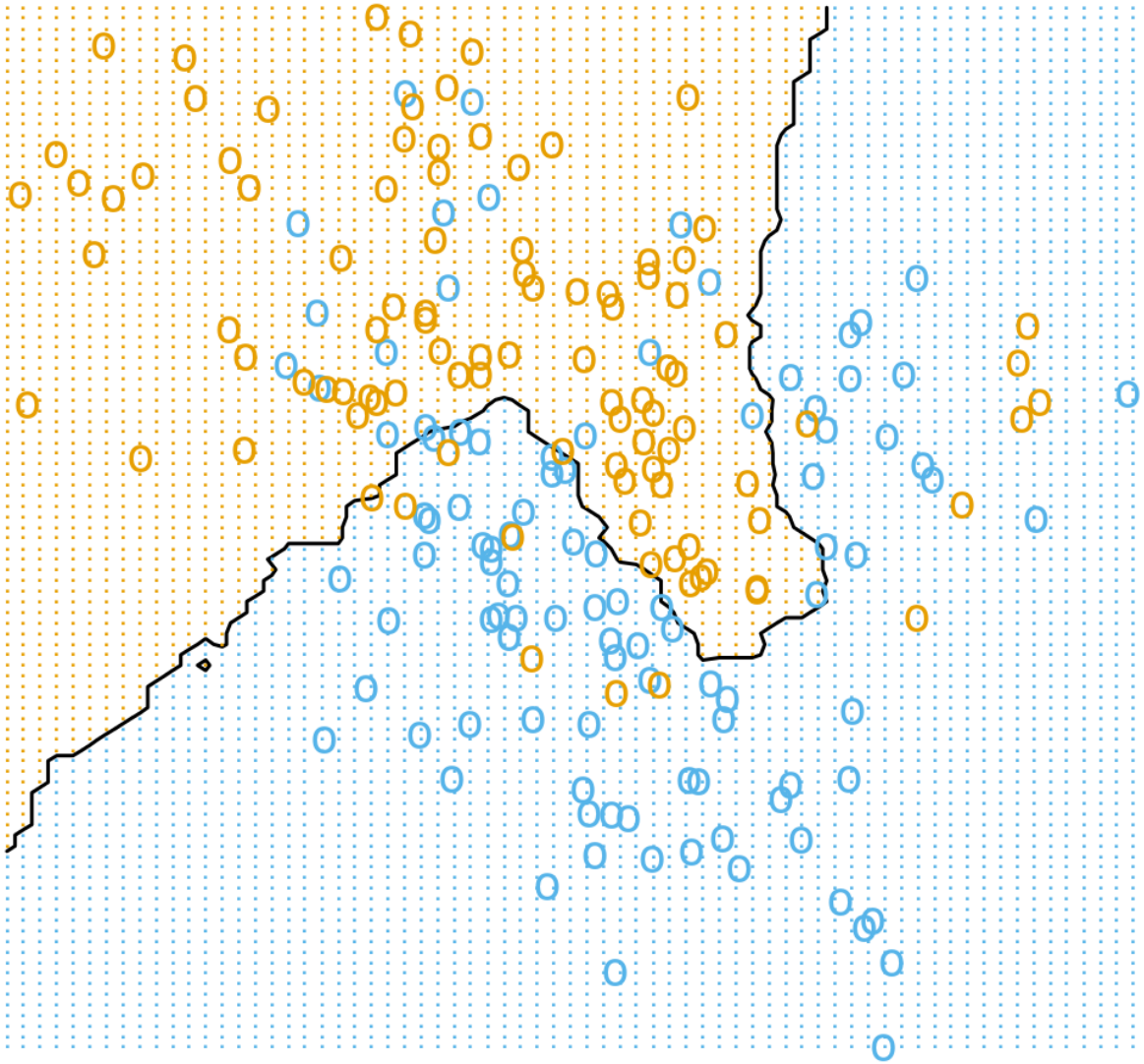
Exemple : plus proche voisin



$x \in \mathcal{X}$
$\rightarrow$ étiquette Y_i
du plus proche
voisin X_i de x
parmi $X_1, \ldots, X_n$

Figure : [Hastie et al., 2009]

Exemple : k plus proches voisins ($k = 15$)



$x \in \mathcal{X}$
$\rightarrow$ vote majoritaire/moyenne parmi les étiquettes des k plus proches voisins de x parmi $X_1, \ldots, X_n$

Figure : [Hastie et al., 2009]

Consistance

- faible (pour P) :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}(D_n)\big)\Big] \xrightarrow[n\to\infty]{} \mathcal{R}^\star$$

Consistance

- faible (pour P) :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}(D_n)\big)\Big] \xrightarrow[n\to\infty]{} \mathcal{R}^\star$$

- forte (pour P) :

$$\mathcal{R}_P\big(\widehat{f}(D_n)\big) \xrightarrow[n\to\infty]{p.s.} \mathcal{R}^\star$$

Consistance

- faible (pour P) :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}(D_n)\big)\Big] \xrightarrow[n\to\infty]{} \mathcal{R}^\star$$

- forte (pour P) :

$$\mathcal{R}_P\big(\widehat{f}(D_n)\big) \xrightarrow[n\to\infty]{p.s.} \mathcal{R}^\star$$

- consistance universelle (faible ou forte) : pour tout P.

Consistance

- faible (pour P) :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}(D_n)\big)\Big] \xrightarrow[n\to\infty]{} \mathcal{R}^\star$$

- forte (pour P) :

$$\mathcal{R}_P\big(\widehat{f}(D_n)\big) \xrightarrow[n\to\infty]{p.s.} \mathcal{R}^\star$$

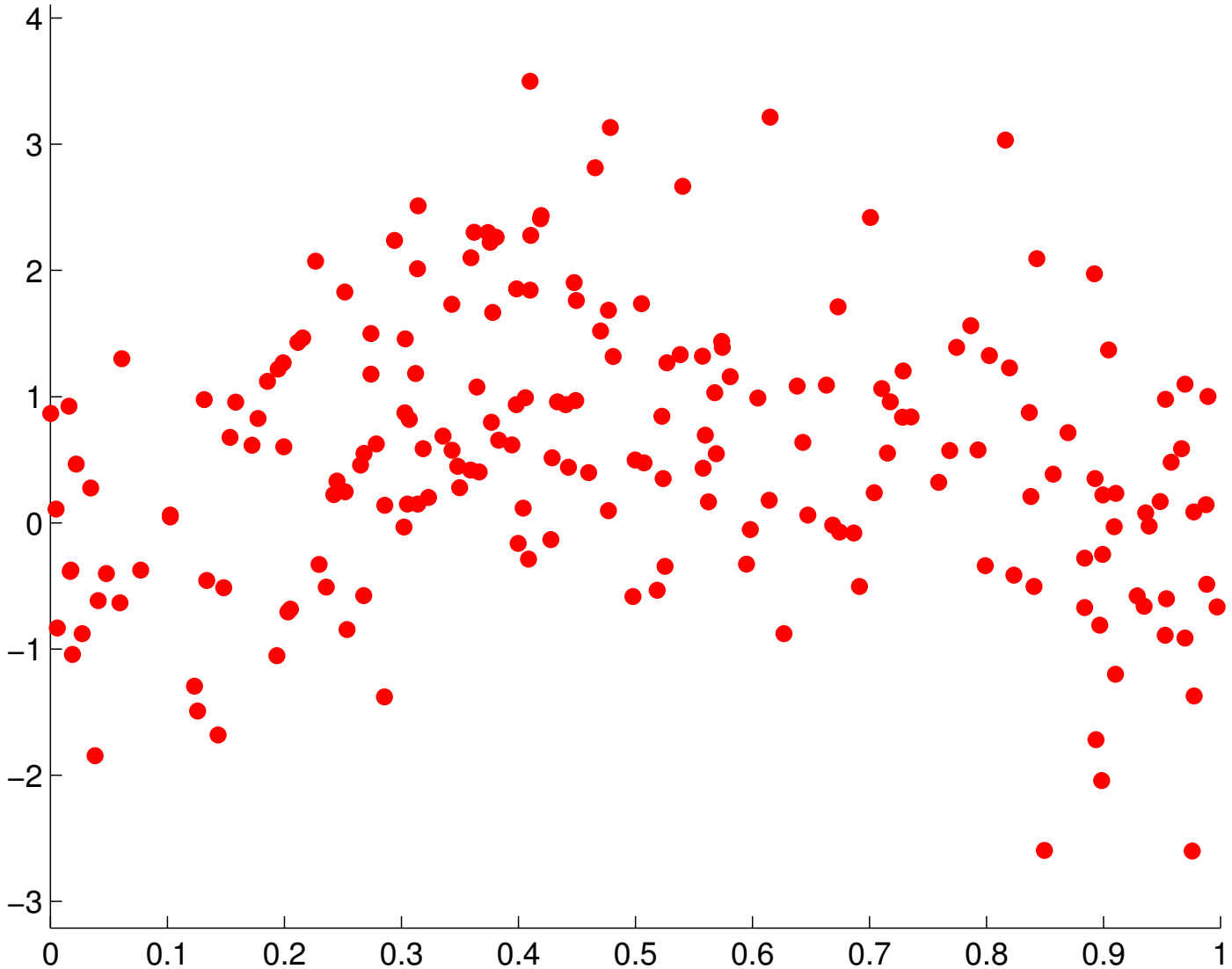
- consistance universelle (faible ou forte) : pour tout P.

- Attention ! Ne donne pas de vitesse de convergence !

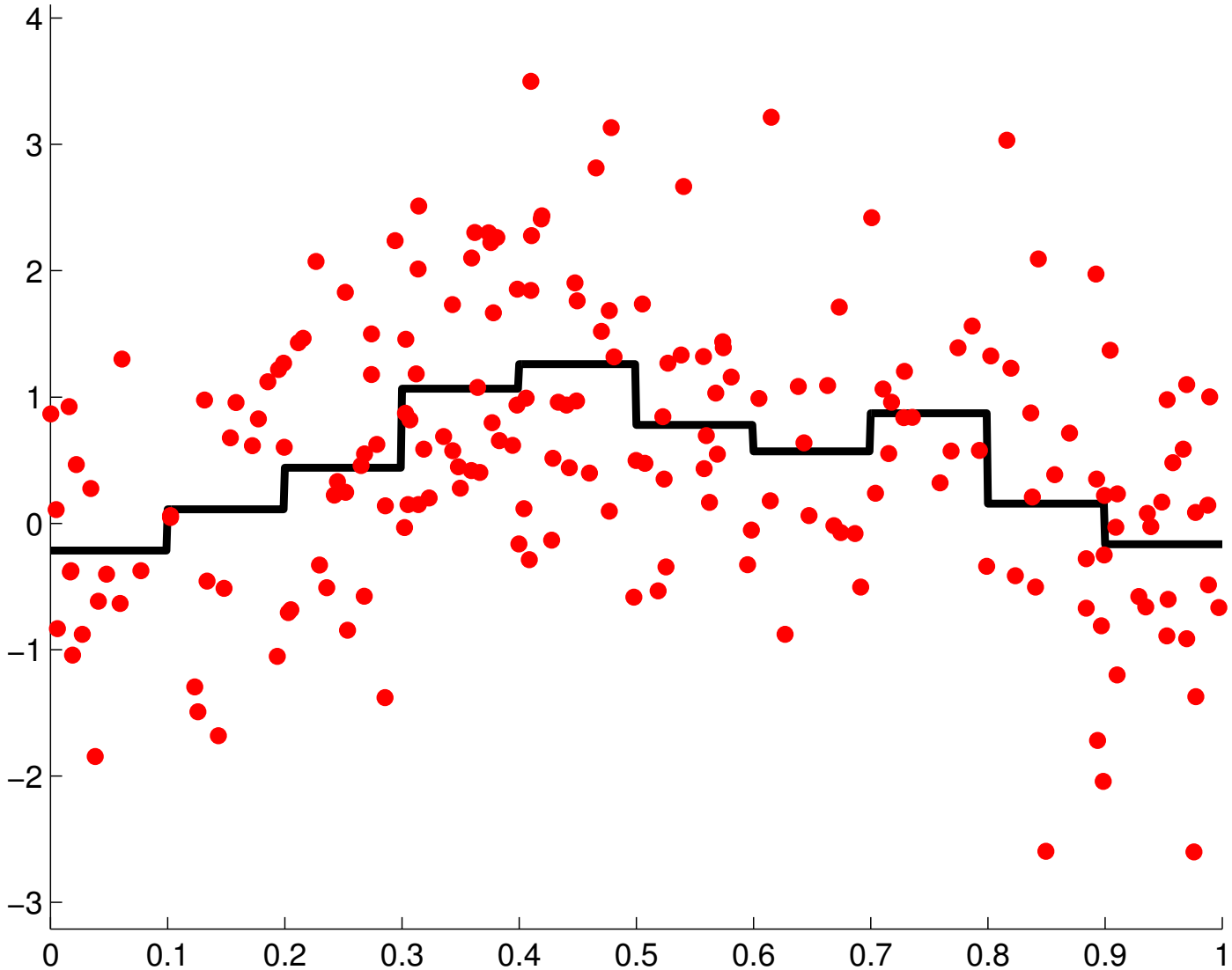
Plan

1. Prévision
2. Régression et classification
3. Minimisation du risque empirique
4. Moyennes locales
5. On n'a rien sans rien
6. Conclusion : enjeux de l'apprentissage

Régression : $Y_i \in \mathcal{Y} = \mathbb{R}$



Règle par partition (régression), $\mathcal{X} = \mathbb{R}$



Arbre de décision ($\Rightarrow$ partition), régression, $\mathcal{X} = \mathbb{R}^2$

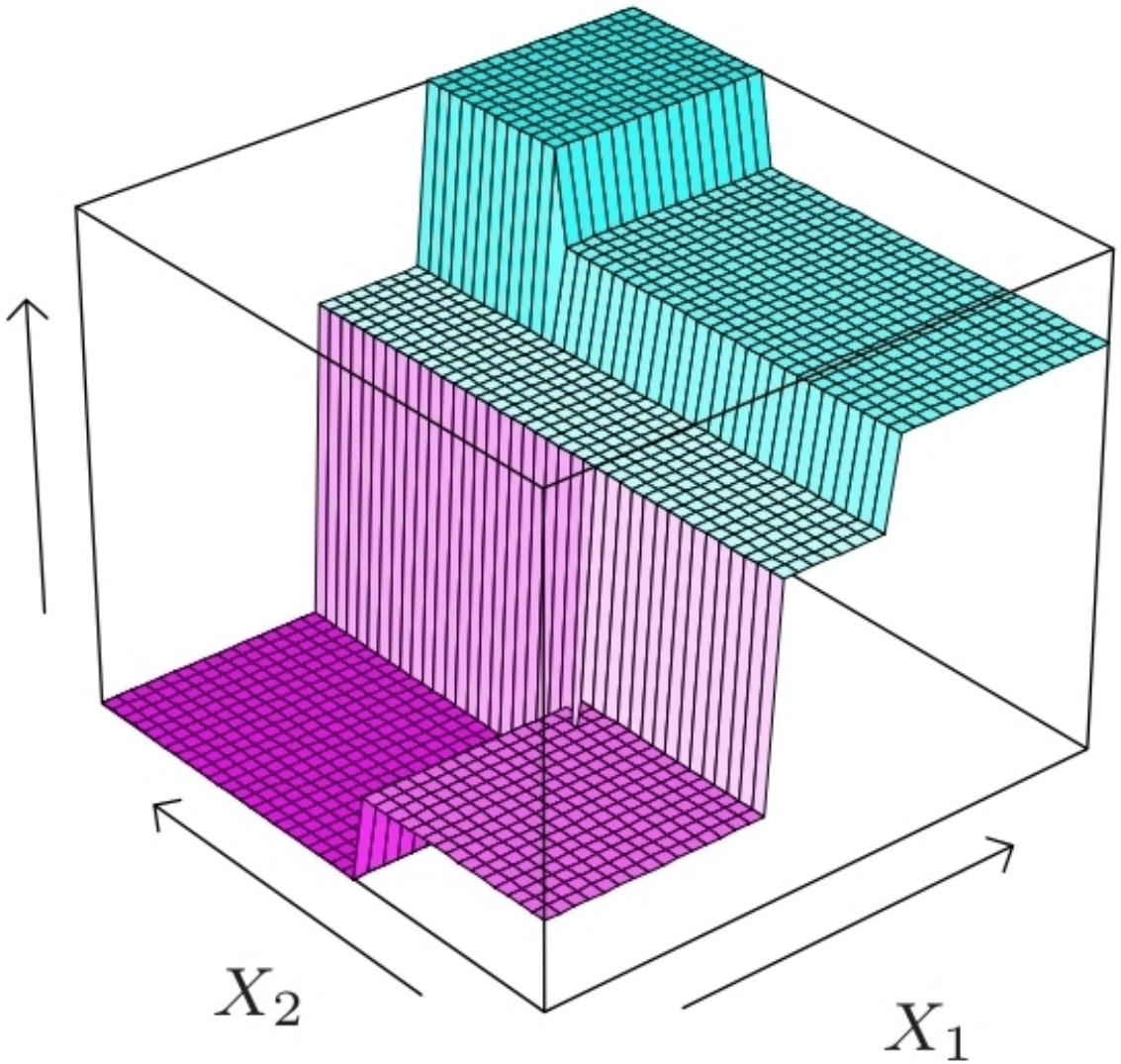
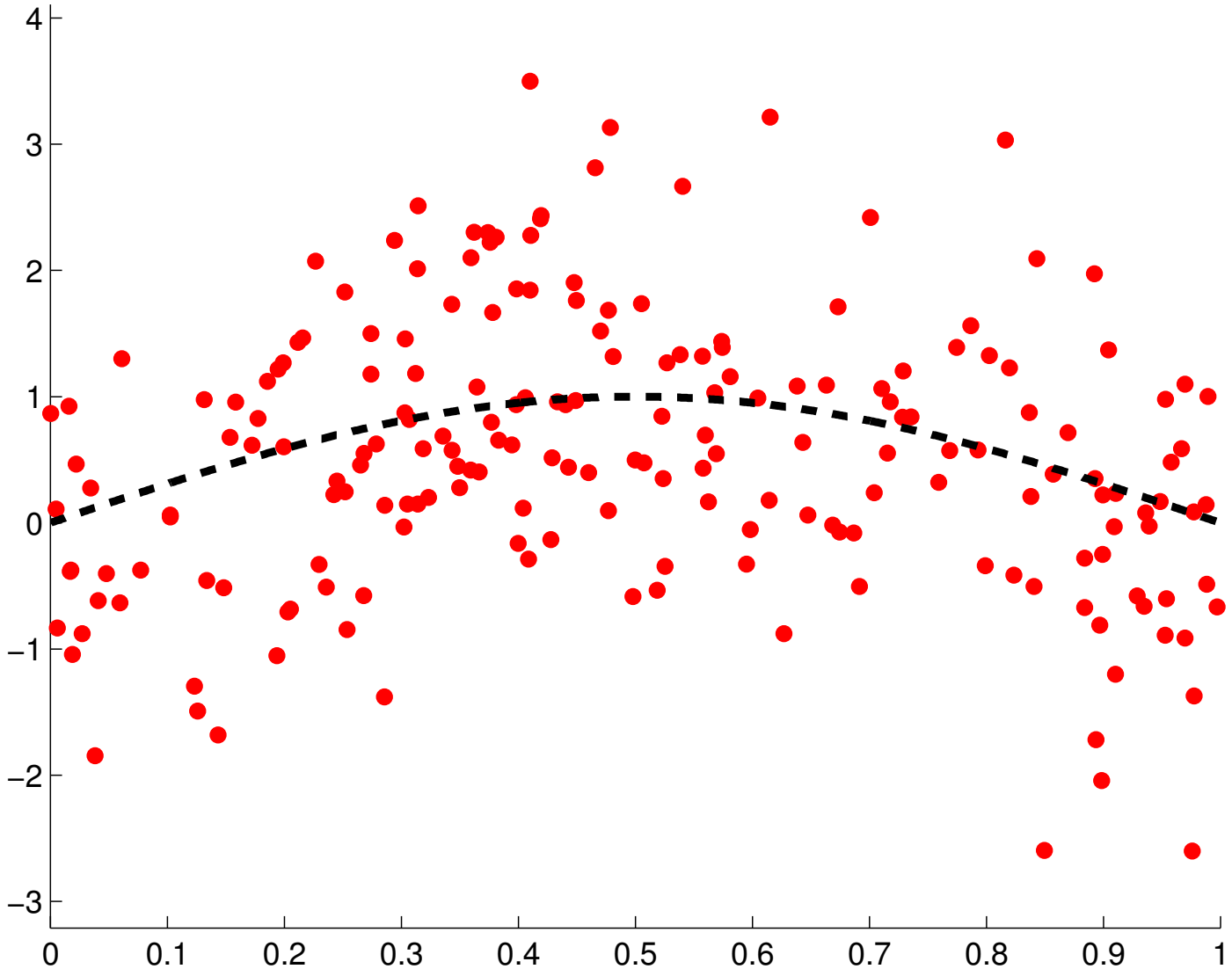
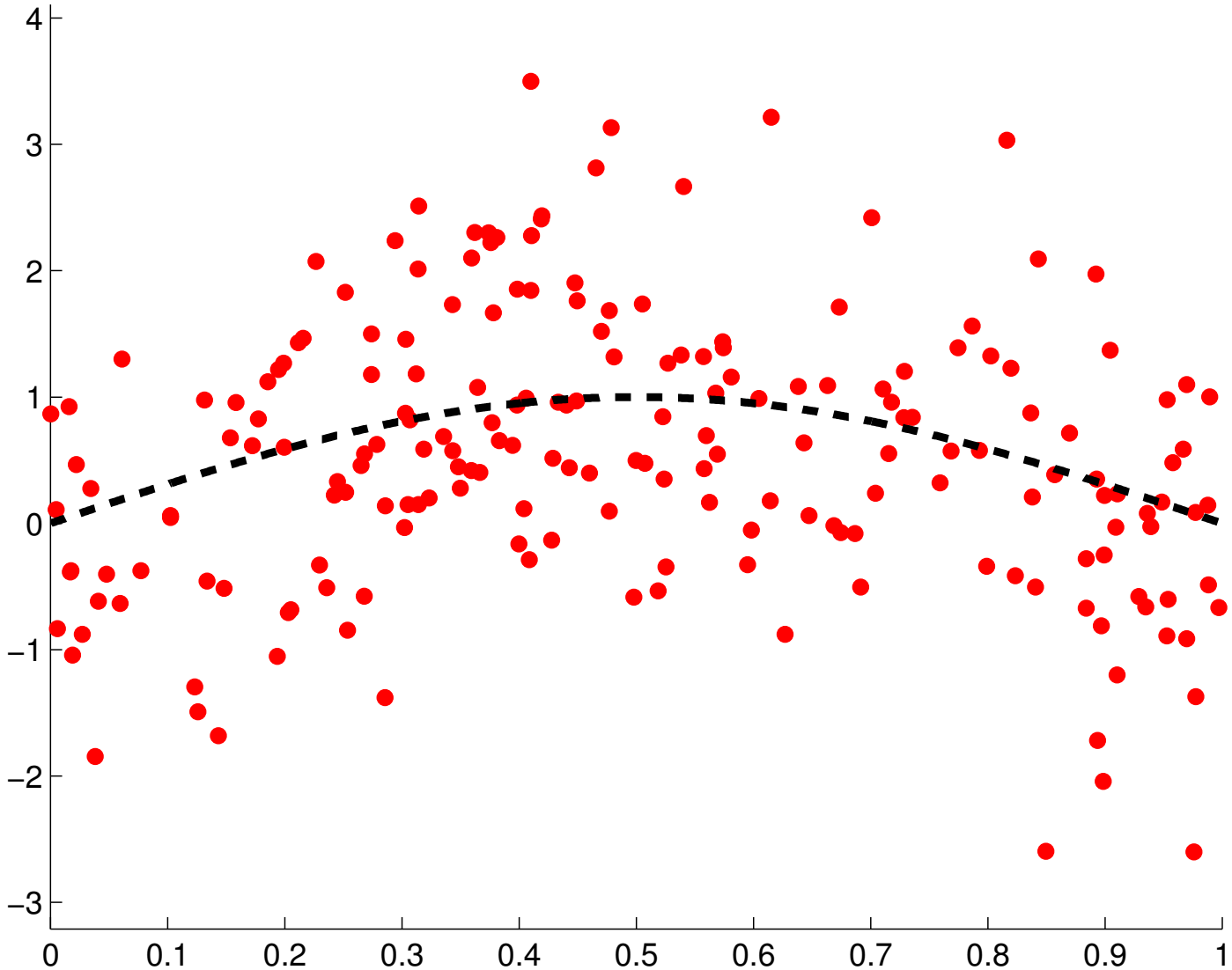


Figure : [Hastie et al., 2009]

Fonction de régression : $\eta(X) := \mathbb{E}[Y \,|\, X]$



Régression : $Y = \eta(X) + \varepsilon$ avec $\mathbb{E}[\varepsilon \,|\, X] = 0$



Règle par partition (régression)

- $\mathcal{A}$: partition de $\mathcal{X}$, finie ou dénombrable.
- $\forall x \in \mathcal{X}$, $\mathcal{A}(x) =$ l'unique élément de $\mathcal{A}$ qui contient x.
- Règle de régression par partition associée à $\mathcal{A}$:

  $\forall n \geqslant 1$, $x \in \mathcal{X}$, $(x_i, y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$\widehat{f}_{\mathcal{A}}^{\,\mathrm{part-reg}}\left((x_i, y_i)_{1 \leqslant i \leqslant n}; x\right) := \frac{\sum_{i=1}^{n} y_i \mathbb{1}_{x_i \in \mathcal{A}(x)}}{\sum_{i=1}^{n} \mathbb{1}_{x_i \in \mathcal{A}(x)}}$$

- Convention : $\frac{0}{0} = 0$

Règle par partition (régression)

- $\mathcal{A}$: partition de $\mathcal{X}$, finie ou dénombrable.
- $\forall x \in \mathcal{X}$, $\mathcal{A}(x) =$ l'unique élément de $\mathcal{A}$ qui contient x.
- Règle de régression par partition associée à $\mathcal{A}$:

  $\forall n \geqslant 1$, $x \in \mathcal{X}$, $(x_i, y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$\widehat{f}_{\mathcal{A}}^{\,\mathrm{part-reg}}\left((x_i, y_i)_{1 \leqslant i \leqslant n}; x\right) := \frac{\sum_{i=1}^{n} y_i \mathbb{1}_{x_i \in \mathcal{A}(x)}}{\sum_{i=1}^{n} \mathbb{1}_{x_i \in \mathcal{A}(x)}}$$

- Convention : $\frac{0}{0} = 0$
- Partition cubique de pas $h > 0$ sur $\mathcal{X} = \mathbb{R}^p$:

$$\mathcal{A}^{\mathrm{cub}}(h) := \left(\prod_{i=1}^{p} \left[hk_i, h(k_i + 1) \right) \right)_{(k_1, \ldots, k_p) \in \mathbb{Z}^p}$$

Coût quadratique, régression idéale (1)

- Coût quadratique (des moindres carrés) : $c(y, y') = (y - y')^2$
- Hypothèse : $\mathbb{E}\left[Y^2\right] < +\infty$
- Risque quadratique de $f \in \mathcal{F}$:

$$\mathcal{R}_P(f) = \mathbb{E}\Big[\big(f(X) - Y\big)^2\Big]$$

Coût quadratique, régression idéale (1)

- Coût quadratique (des moindres carrés) : $c(y, y') = (y - y')^2$
- Hypothèse : $\mathbb{E}\left[Y^2\right] < +\infty$
- Risque quadratique de $f \in \mathcal{F}$:

$$\begin{aligned}\mathcal{R}_P(f) &= \mathbb{E}\Big[\big(f(X) - Y\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X) - \varepsilon\big)^2\Big]\end{aligned}$$

Coût quadratique, régression idéale (1)

- Coût quadratique (des moindres carrés) : $c(y, y') = (y - y')^2$
- Hypothèse : $\mathbb{E}\left[Y^2\right] < +\infty$
- Risque quadratique de $f \in \mathcal{F}$:

$$\begin{aligned}\mathcal{R}_P(f) &= \mathbb{E}\Big[\big(f(X) - Y\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X) - \varepsilon\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big] + \mathbb{E}\left[\varepsilon^2\right] + 2\,\mathbb{E}\Big[\big(f(X) - \eta(X)\big)\varepsilon\Big]\end{aligned}$$

Coût quadratique, régression idéale (1)

- Coût quadratique (des moindres carrés) : $c(y, y') = (y - y')^2$
- Hypothèse : $\mathbb{E}\left[Y^2\right] < +\infty$
- Risque quadratique de $f \in \mathcal{F}$:

$$\begin{aligned}
\mathcal{R}_P(f) &= \mathbb{E}\Big[\big(f(X) - Y\big)^2\Big] \\
&= \mathbb{E}\Big[\big(f(X) - \eta(X) - \varepsilon\big)^2\Big] \\
&= \mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big] + \mathbb{E}\left[\varepsilon^2\right] + 2\underbrace{\mathbb{E}\Big[\big(f(X) - \eta(X)\big)\varepsilon\Big]}_{=0}
\end{aligned}$$

Coût quadratique, régression idéale (1)

- Coût quadratique (des moindres carrés) : $c(y, y') = (y - y')^2$
- Hypothèse : $\mathbb{E}\left[Y^2\right] < +\infty$
- Risque quadratique de $f \in \mathcal{F}$:

$$\begin{aligned}\mathcal{R}_P(f) &= \mathbb{E}\Big[\big(f(X) - Y\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X) - \varepsilon\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big] + \mathbb{E}\left[\varepsilon^2\right]\end{aligned}$$

$\Rightarrow$ minimal si et seulement si $f(X) = \eta(X)$ p.s.

Coût quadratique, régression idéale (1)

- Coût quadratique (des moindres carrés) : $c(y, y') = (y - y')^2$
- Hypothèse : $\mathbb{E}\left[Y^2\right] < +\infty$
- Risque quadratique de $f \in \mathcal{F}$:

$$\begin{aligned}\mathcal{R}_P(f) &= \mathbb{E}\Big[\big(f(X) - Y\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X) - \varepsilon\big)^2\Big] \\ &= \mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big] + \mathbb{E}\left[\varepsilon^2\right]\end{aligned}$$

$\Rightarrow$ minimal si et seulement si $f(X) = \eta(X)$ p.s.

$\Rightarrow$ excès de risque $\mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big]$

Coût quadratique, régression idéale (2)

Proposition (2.1)

(i) η *est un prédicteur de Bayes*

(ii) $f \in \mathcal{F}$ *est un prédicteur de Bayes* $\Leftrightarrow f(X) = \eta(X)$ *p.s.*

(iii) *Risque de Bayes* :

$$\mathcal{R}_P^\star = \mathbb{E}\Big[\big(Y - \eta(X)\big)^2\Big] = \mathsf{var}(Y \,|\, X) = \mathbb{E}\big[\varepsilon^2\big]$$

(iv) *Excès de risque de* $f \in \mathcal{F}$:

$$\ell(f^\star, f) = \mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big] = \|f - \eta\|^2_{L^2(P_X)}$$

Coût quadratique, régression idéale (2)

Proposition (2.1)

(i) η *est un prédicteur de Bayes*

(ii) $f \in \mathcal{F}$ *est un prédicteur de Bayes* $\Leftrightarrow f(X) = \eta(X)$ *p.s.*

(iii) *Risque de Bayes* :

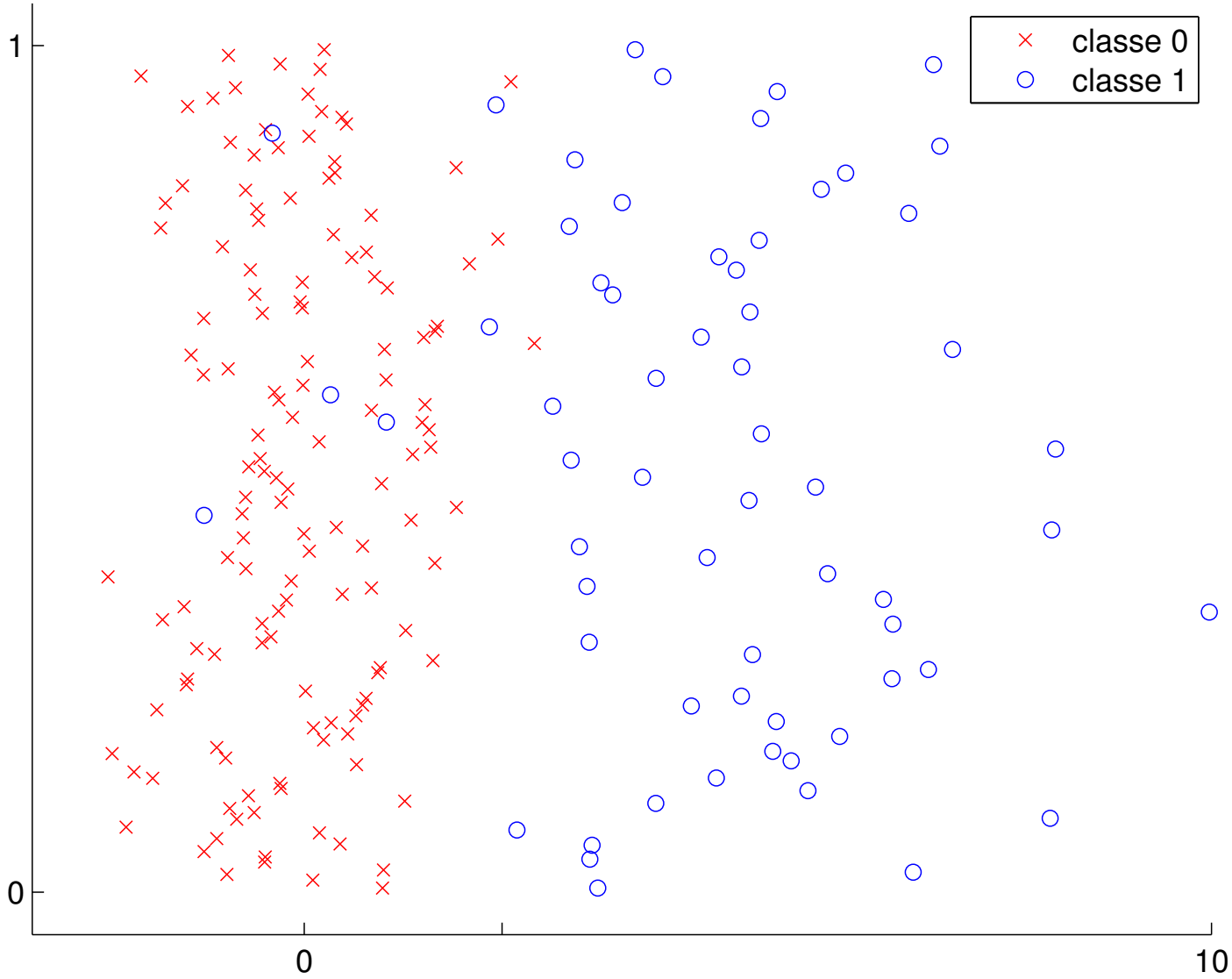
$$\mathcal{R}_P^\star = \mathbb{E}\Big[\big(Y - \eta(X)\big)^2\Big] = \mathsf{var}(Y \,|\, X) = \mathbb{E}\big[\varepsilon^2\big]$$

(iv) *Excès de risque de* $f \in \mathcal{F}$:

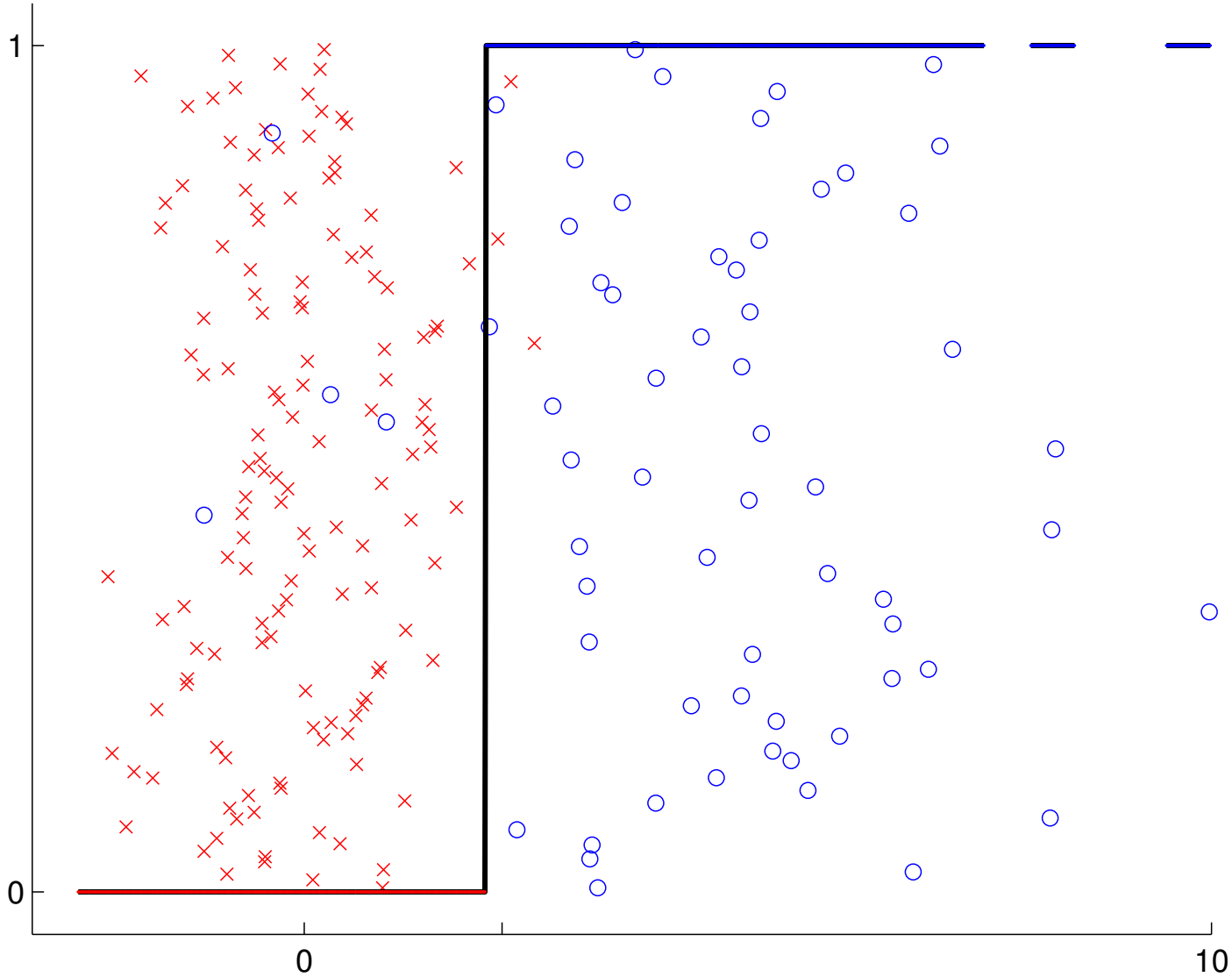
$$\ell(f^\star, f) = \mathbb{E}\Big[\big(f(X) - \eta(X)\big)^2\Big] = \|f - \eta\|^2_{L^2(P_X)}$$

Exercice (2.1–2.2) : avec le coût valeur absolue $c(y, y') = |y - y'|$?

Classification (binaire supervisée) : $Y_i \in \mathcal{Y} = \{0, 1\}$



Règle par partition (classification)



Règle par partition (classification)

- $\mathcal{A}$: partition de $\mathcal{X}$, finie ou dénombrable.
- $\forall x \in \mathcal{X}$, $\mathcal{A}(x) =$ l'unique élément de $\mathcal{A}$ qui contient x.

- Règle de classification par partition associée à $\mathcal{A}$:

  $\forall n \geqslant 1$, $x \in \mathcal{X}$, $(x_i, y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$,

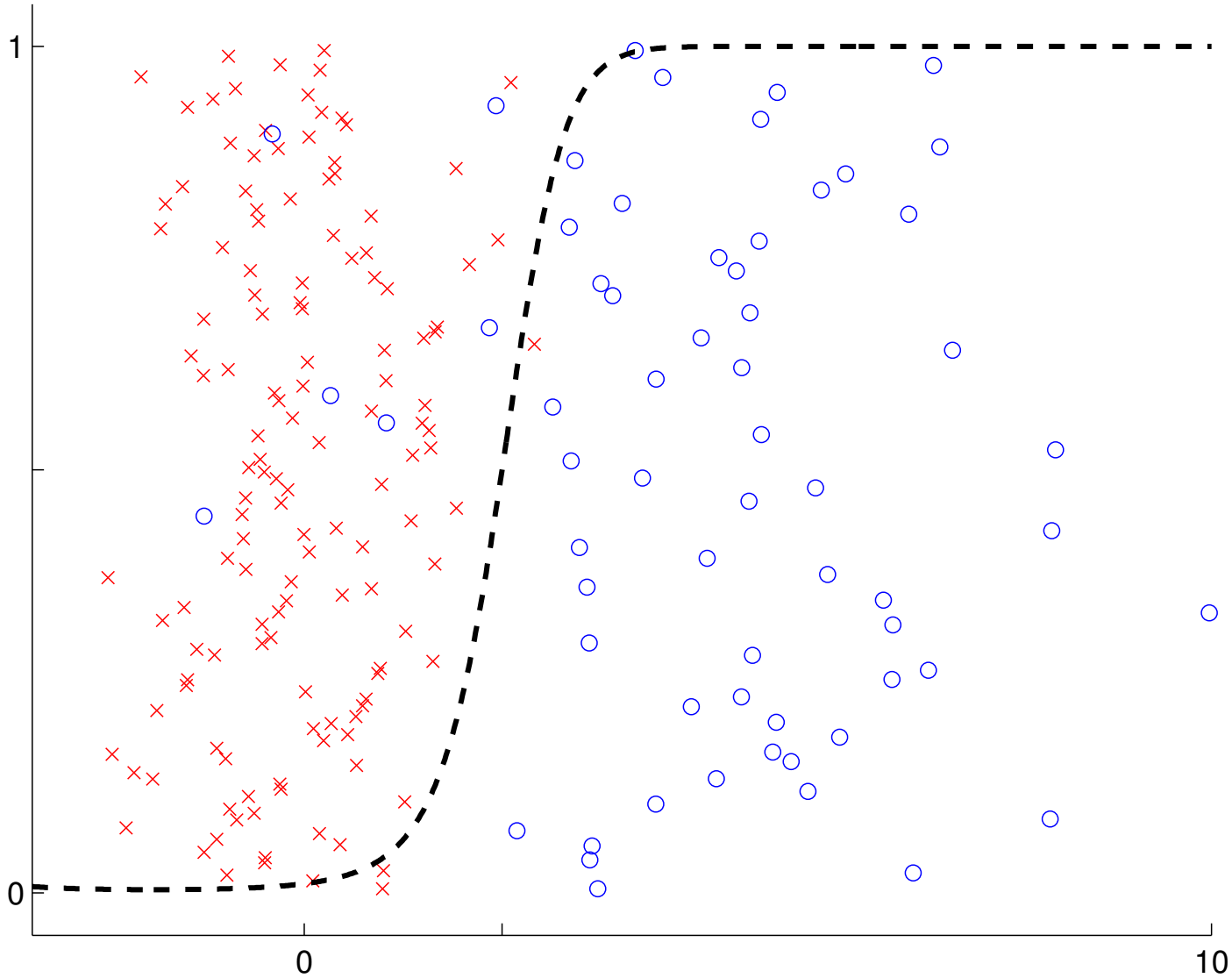
$$\widehat{f}_{\mathcal{A}}^{\,\mathrm{part-class}}\left((x_i, y_i)_{1 \leqslant i \leqslant n}; x\right) := \begin{cases} 1 & \text{si} \quad \mathrm{Card}\left\{i \,/\, y_i = 1 \text{ et } x_i \in \mathcal{A}(x)\right\} \\ & \qquad > \mathrm{Card}\left\{i \,/\, y_i = 0 \text{ et } x_i \in \mathcal{A}(x)\right\} \\ 0 & \text{sinon} \end{cases}$$

  $\Leftrightarrow$ vote majoritaire dans chaque case, avec 0 par défaut

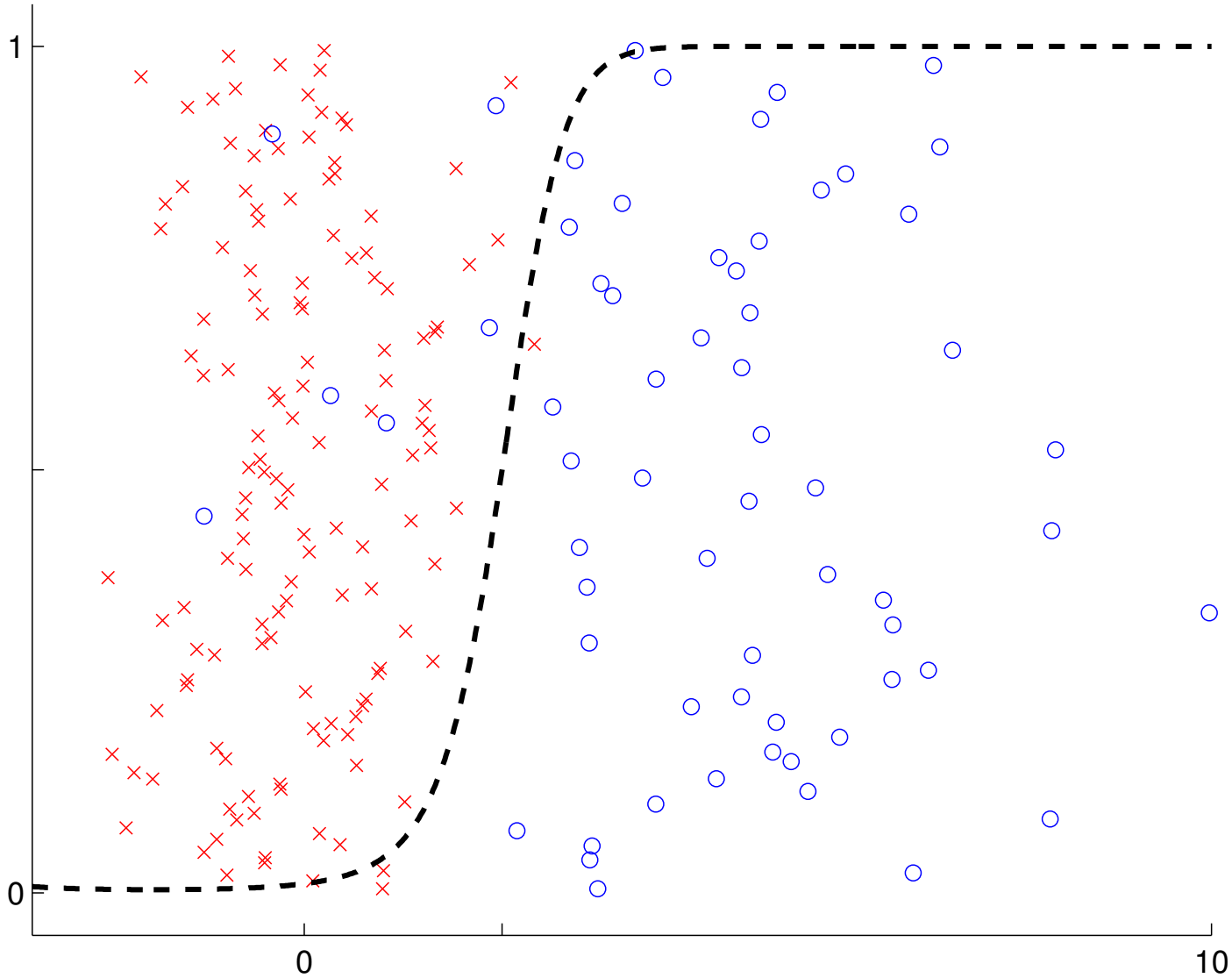
- Partition cubique de pas $h > 0$ sur $\mathcal{X} = \mathbb{R}^p$:

$$\mathcal{A}^{\mathrm{cub}}(h) := \left(\prod_{i=1}^{p} \left[hk_i, h(k_i + 1)\right)\right)_{(k_1, \ldots, k_p) \in \mathbb{Z}^p}$$

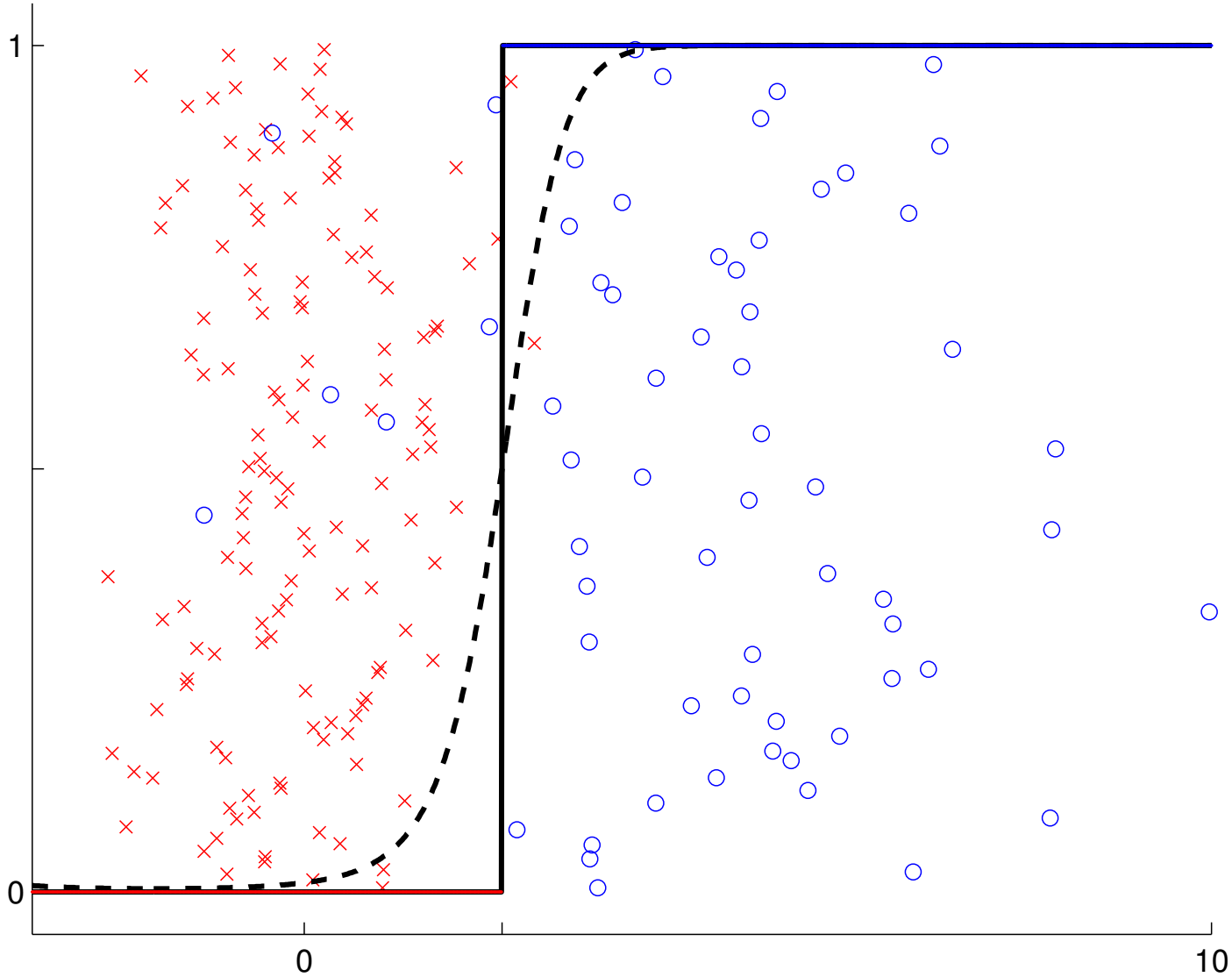
Fonction de régression : $\eta(X) := \mathbb{E}[Y \,|\, X] = \mathbb{P}(Y = 1 \,|\, X)$



Coût 0–1 $c(y, y') = \mathbb{1}_{y \neq y'}$: classification idéale ?



Coût 0–1 $c(y, y') = \mathbb{1}_{y \neq y'}$: classification idéale



Coût 0–1 : classification idéale

Risque 0–1 : $\mathcal{R}_P : f \in \mathcal{F} \mapsto \mathbb{E}[\mathbb{1}_{f(X)\neq Y}] = \mathbb{P}(f(X) \neq Y)$

Proposition (2.2)

(i) $f^\star : x \mapsto \mathbb{1}_{\eta(x)>1/2}$ *est un classifieur de Bayes*

(ii) $f \in \mathcal{F}$ *est un classifieur de Bayes* $\Leftrightarrow f(X) = \mathbb{1}_{\eta(X)>1/2}$ *p.s., sauf éventuellement sur* $\{\eta(X) = 1/2\}$

(iii) *Risque de Bayes* :

$$\mathcal{R}_P^\star = \mathbb{E}\Big[\min\{\eta(X), 1-\eta(X)\}\Big]$$

(iv) *Excès de risque de* $f \in \mathcal{F}$:

$$\ell(f^\star, f) = \mathbb{E}\Big[|2\eta(X) - 1|\mathbb{1}_{f^\star(X)\neq f(X)}\Big]$$

Coût asymétrique

- Coût asymétrique : $c_w : (y, y') \in \{0,1\} \mapsto w_{y'} \mathbb{1}_{y \neq y'}$ avec $w = (w_0, w_1) \in [0, +\infty[^2$ et $w_0 + w_1 > 0$
- Motivations : spams, diagnostic médical, etc.
- Risque associé :

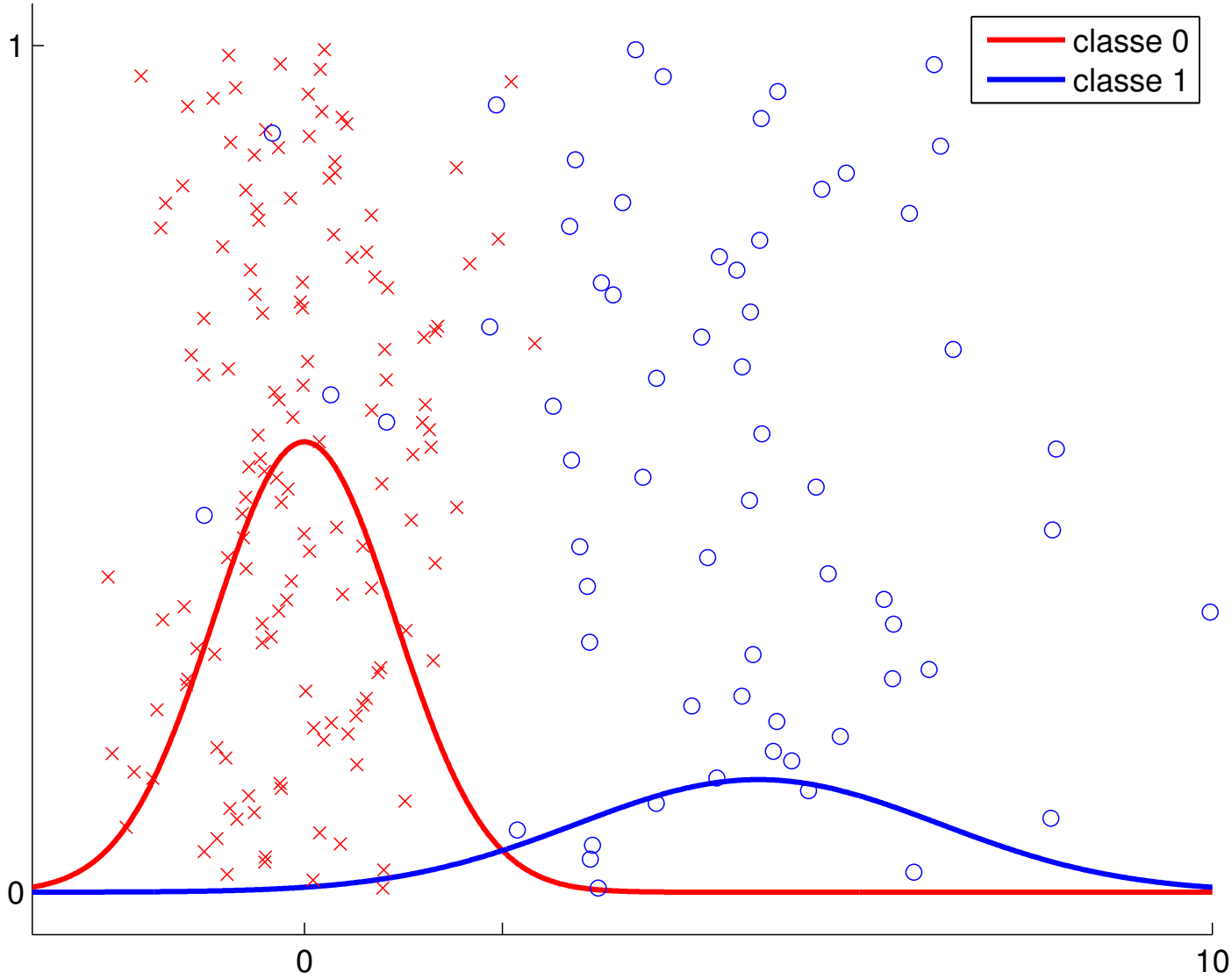
$$\begin{aligned} \mathcal{R}_P^w(f) &= \mathbb{E}[w_Y \mathbb{1}_{f(X) \neq Y}] \\ &= w_1 \mathbb{P}(f(X) = 0 \text{ et } Y = 1) + w_0 \mathbb{P}(f(X) = 1 \text{ et } Y = 0) \,. \end{aligned}$$

- Classifieur de Bayes (Proposition 2.3) :

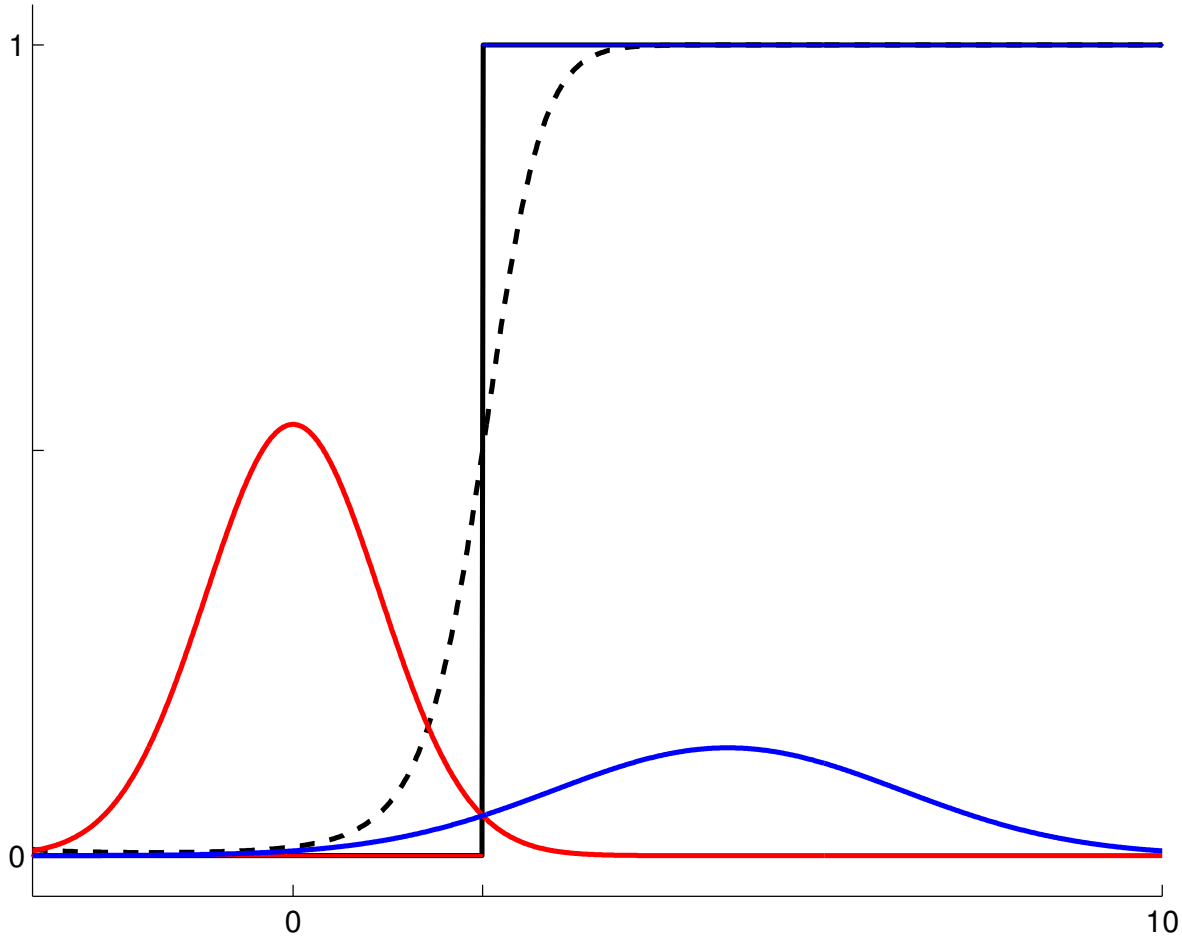
$$f_w^\star : x \mapsto \mathbb{1}_{\eta(x) > \frac{w_0}{w_0 + w_1}}$$

$\Leftrightarrow$ test de rapport de vraisemblance de H_0 : « $x \sim \mathcal{L}(X \mid Y = 0)$ » contre H_1 : « $x \sim \mathcal{L}(X \mid Y = 1)$ ».

Classification binaire et test d'hypothèse (Rq. 2.21)



Classification binaire et test d'hypothèse (Rq. 2.21)



Classifieur de Bayes $f_w^\star$

$\Leftrightarrow$

test de rapport de vraisemblance

$$\mathbb{1}\left\{\frac{g_1(x)}{g_0(x)} \geqslant \frac{\mathbb{P}(Y=0)w_0}{\mathbb{P}(Y=1)w_1}\right\}$$

[Figure : $\mathbb{P}(Y=0)=\mathbb{P}(Y=1)=1/2, \quad w_0=w_1=1$]

Classifieur « plug-in »

- Idée :

$$s^\star(x) = \mathbb{1}_{\eta(x) > \frac{1}{2}}$$

$\Rightarrow$ si $\widehat{\eta}(D_n)$ estime η (régression),

$$\widehat{f}_{\widehat{\eta}}(D_n; x) = \mathbb{1}_{\widehat{\eta}(D_n; x) > \frac{1}{2}}$$

est la règle de classification par plug-in associée à $\widehat{\eta}$

Classifieur « plug-in »

- Idée :

$$s^\star(x) = \mathbb{1}_{\eta(x) > \frac{1}{2}}$$

$\Rightarrow$ si $\widehat{\eta}(D_n)$ estime η (régression),

$$\widehat{f}_{\widehat{\eta}}(D_n; x) = \mathbb{1}_{\widehat{\eta}(D_n; x) > \frac{1}{2}}$$

est la règle de classification par plug-in associée à $\widehat{\eta}$

- Exemples :
  - règles par partitions :

$$\widehat{f}_{\mathcal{A}}^{\mathrm{part-class}}(D_n; x) = \mathbb{1}_{\widehat{f}_{\mathcal{A}}^{\mathrm{part-reg}}(D_n; x) > 1/2}$$

  - k plus proches voisins
  - moyennes locales (Section 2.4)

Borne de risque pour un classifieur « plug-in »

Proposition (2.4)

Pour le coût 0–1 en classification :

$$\ell(f^\star, \widehat{f}_{\widehat{\eta}}) \leqslant 2\mathbb{E}|\widehat{\eta}(X) - \eta(X)| \leqslant 2\sqrt{\mathbb{E}\Big[\big(\widehat{\eta}(X) - \eta(X)\big)^2\Big]}$$

Borne de risque pour un classifieur « plug-in »

Proposition (2.4)

Pour le coût 0–1 en classification :

$$\ell(f^\star, \widehat{f}_{\widehat{\eta}}) \leqslant 2\mathbb{E}|\widehat{\eta}(X) - \eta(X)| \leqslant 2\sqrt{\mathbb{E}\left[(\widehat{\eta}(X) - \eta(X))^2\right]}$$

En particulier :
$\widehat{\eta}$ *faiblement consistante* (*pour* P, *en régression avec le coût quadratique*)
$\Rightarrow \widehat{f}_{\widehat{\eta}}$ *faiblement consistante* (*pour* P, *en classification avec le coût 0–1*).

Coûts convexes : pseudo-classifieurs et Φ-risque

- Pourquoi un coût convexe ?

Coûts convexes : pseudo-classifieurs et Φ-risque

- Pourquoi un coût convexe ?
- Convention (ici seulement) : $Y_i \in \mathcal{Y} = \{-1, 1\}$

Coûts convexes : pseudo-classifieurs et Φ-risque

- Pourquoi un coût convexe ?
- Convention (ici seulement) : $Y_i \in \mathcal{Y} = \{-1, 1\}$
- « Pseudo-classifieur » $g \in \overline{\mathcal{F}}$: fonction $\mathcal{X} \to \overline{\mathbb{R}}$

$\Rightarrow$ classifieur associé :

$$\text{signe}(g) : x \in \mathcal{X} \mapsto \begin{cases} 1 & \text{si } g(x) > 0 \\ -1 & \text{sinon} \end{cases}$$

Coûts convexes : pseudo-classifieurs et Φ-risque

- Pourquoi un coût convexe ?
- Convention (ici seulement) : $Y_i \in \mathcal{Y} = \{-1, 1\}$
- « Pseudo-classifieur » $g \in \overline{\mathcal{F}}$: fonction $\mathcal{X} \to \overline{\mathbb{R}}$

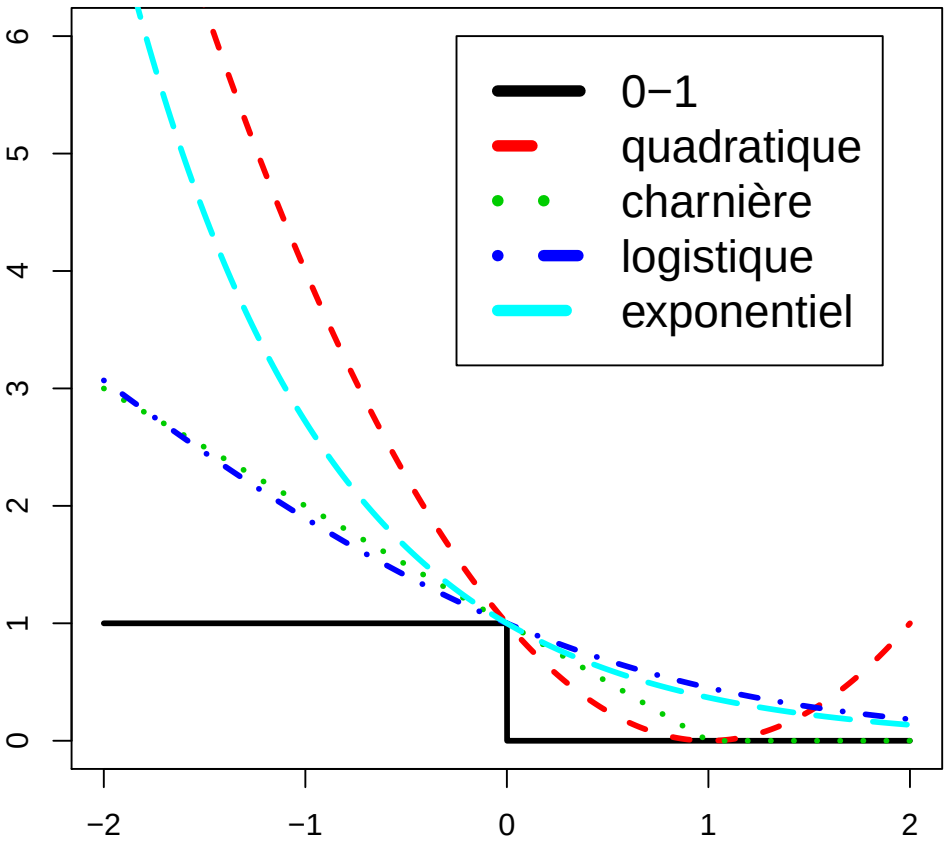
$\Rightarrow$ classifieur associé :

$$\text{signe}(g) : x \in \mathcal{X} \mapsto \begin{cases} 1 & \text{si } g(x) > 0 \\ -1 & \text{sinon} \end{cases}$$

- $\Phi : \overline{\mathbb{R}} \to [0, +\infty] \Rightarrow$ Φ-risque de $g \in \overline{\mathcal{F}}$:

$$\mathcal{R}_P^{\Phi}(g) := \mathbb{E}\Big[\Phi\big(Yg(X)\big)\Big]$$

Coûts convexes : Exemples de fonctions Φ



- 0–1 : $\Phi_{0-1}(u) = \mathbb{1}_{u \leqslant 0}$
- quadratique : $\Phi(u) = (1-u)^2$
- charnière (hinge) : $\Phi(u) = \max\{1-u, 0\}$ $\Rightarrow$ SVM
- logistique : $\Phi(u) = \ln_2(1+\mathrm{e}^{-u})$ $\Rightarrow$ régression logistique
- exponentiel : $\Phi(u) = \mathrm{e}^{-u}$ $\Rightarrow$ boosting, AdaBoost

$$\mathcal{R}_P^{\Phi}(g) := \mathbb{E}\Big[\Phi\big(Yg(X)\big)\Big]$$

Lien entre Φ-risque et risque 0–1

Définition (Définition 2.3 & Proposition 2.6)

Φ est calibrée pour la classification si pour toute loi P, $g_\Phi^\star \in \operatorname{argmin}_{g \in \overline{\mathcal{F}}} \mathcal{R}_P^\Phi(g) \Rightarrow \operatorname{signe}(g_\Phi^\star)$ classifieur de Bayes pour le risque 0–1.

Lien entre Φ-risque et risque 0–1

Définition (Définition 2.3 & Proposition 2.6)

Φ est calibrée pour la classification si pour toute loi P, $g_\Phi^\star \in \operatorname{argmin}_{g \in \overline{\mathcal{F}}} \mathcal{R}_P^\Phi(g) \Rightarrow \text{signe}(g_\Phi^\star)$ classifieur de Bayes pour le risque 0–1.

Proposition (2.7)

On suppose Φ convexe. Alors :

Φ calibrée pour la classification $\Leftrightarrow$ Φ dérivable en 0 et $\Phi'(0) < 0$.

Exemples :

- quadratique
- charnière
- logistique
- exponentiel

Lien entre Φ-risque et risque 0–1

Définition (Définition 2.3 & Proposition 2.6)

Φ est calibrée pour la classification si pour toute loi P, $g_\Phi^\star \in \operatorname{argmin}_{g \in \overline{\mathcal{F}}} \mathcal{R}_P^\Phi(g) \Rightarrow \operatorname{signe}(g_\Phi^\star)$ classifieur de Bayes pour le risque 0–1.

Proposition (2.7)

On suppose Φ convexe. Alors :
Φ calibrée pour la classification $\Leftrightarrow$ Φ dérivable en 0 et $\Phi'(0) < 0$.

Théorème (2.1)

Si Φ est convexe et calibrée pour la classification, il existe une fonction $\Psi : [-1, 1] \to [0, +\infty[$ telle que :

$$\forall P, \forall g \in \overline{\mathcal{F}}, \qquad \Psi\left(\mathcal{R}_P^{0-1}(\operatorname{signe}(g)) - \mathcal{R}_P^{0-1\star}\right) \leqslant \mathcal{R}_P^\Phi(g) - \mathcal{R}_P^{\Phi\star}.$$

Lien entre Φ-risque et risque 0–1

Théorème (2.1)

Si Φ est convexe et calibrée pour la classification, il existe une fonction $\Psi : [-1, 1] \to [0, +\infty[$ telle que :

$$\forall P, \forall g \in \overline{\mathcal{F}}, \qquad \Psi\left(\mathcal{R}_P^{0-1}(\text{signe}(g)) - \mathcal{R}_P^{0-1\star}\right) \leqslant \mathcal{R}_P^{\Phi}(g) - \mathcal{R}_P^{\Phi\star}.$$

Exemples :

- quadratique : $\Psi(\theta) = \theta^2$
- charnière : $\Psi(\theta) = |\theta|$
- logistique : $\Psi(\theta) \geqslant \theta^2/(2 \ln 2)$
- exponentiel : $\Psi(\theta) = 1 - \sqrt{1 - \theta^2} \geqslant \theta^2/2$

Plan

1. Prévision
2. Régression et classification
3. Minimisation du risque empirique
4. Moyennes locales
5. On n'a rien sans rien
6. Conclusion : enjeux de l'apprentissage

Minimisation du risque empirique

- Risque empirique de $f \in \mathcal{F}$:

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} c(f(X_i), Y_i)$$

- Pour tout $f \in \mathcal{F}$:

$$\mathbb{E}\big[\widehat{\mathcal{R}}_n(f)\big] = \mathbb{E}\Big[c(f(X), Y)\Big] = \mathcal{R}_P(f)$$

Minimisation du risque empirique

- Risque empirique de $f \in \mathcal{F}$:

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} c(f(X_i), Y_i)$$

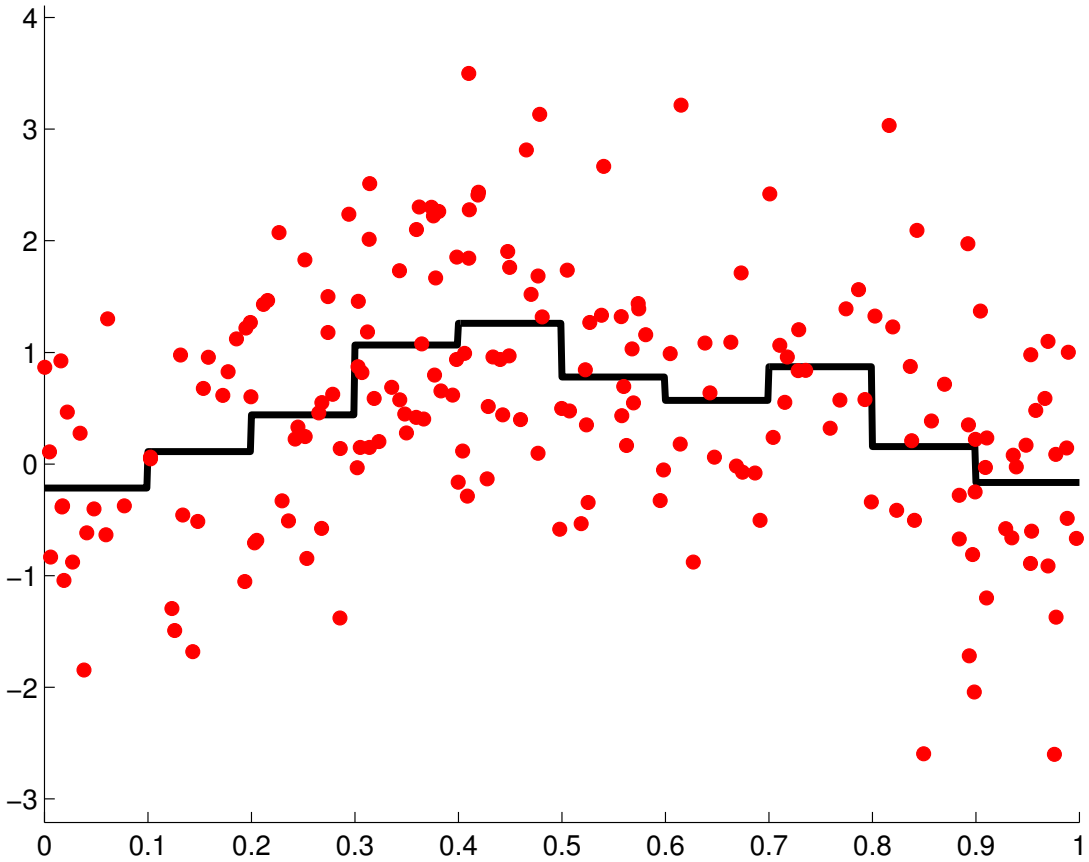
- Pour tout $f \in \mathcal{F}$:

$$\mathbb{E}\big[\widehat{\mathcal{R}}_n(f)\big] = \mathbb{E}\Big[c(f(X), Y)\Big] = \mathcal{R}_P(f)$$

- Minimiseur du risque empirique sur le modèle $S \subset \mathcal{F}$:

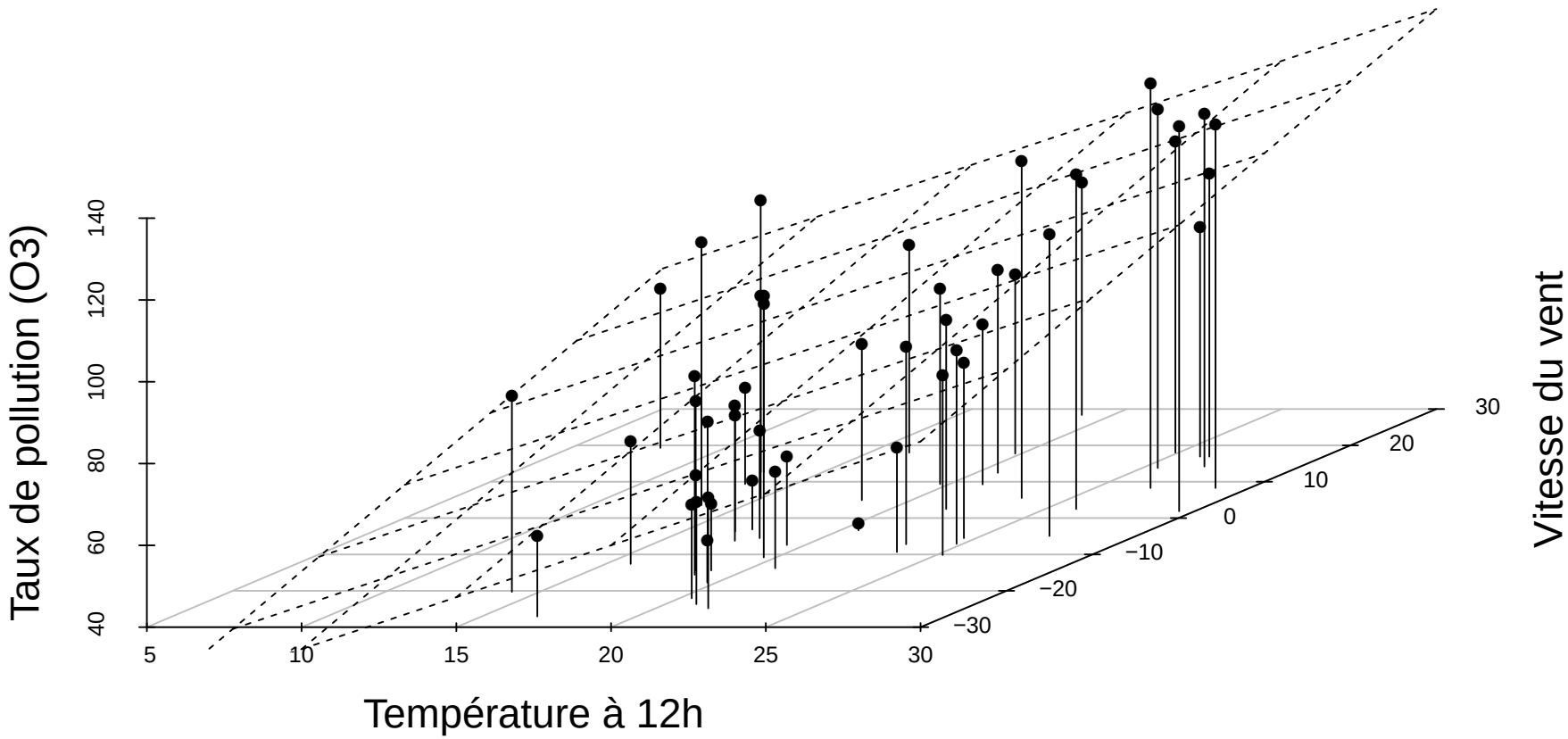
$$\widehat{f}_S \in \operatorname*{argmin}_{f \in S} \{\widehat{\mathcal{R}}_n(f)\}$$

Exemple : partition en régression



$$S_{\text{reg}}^{\text{part}}(\mathcal{A}) = \overline{\text{vect}\{\mathbb{1}_A, A \in \mathcal{A}\}}$$

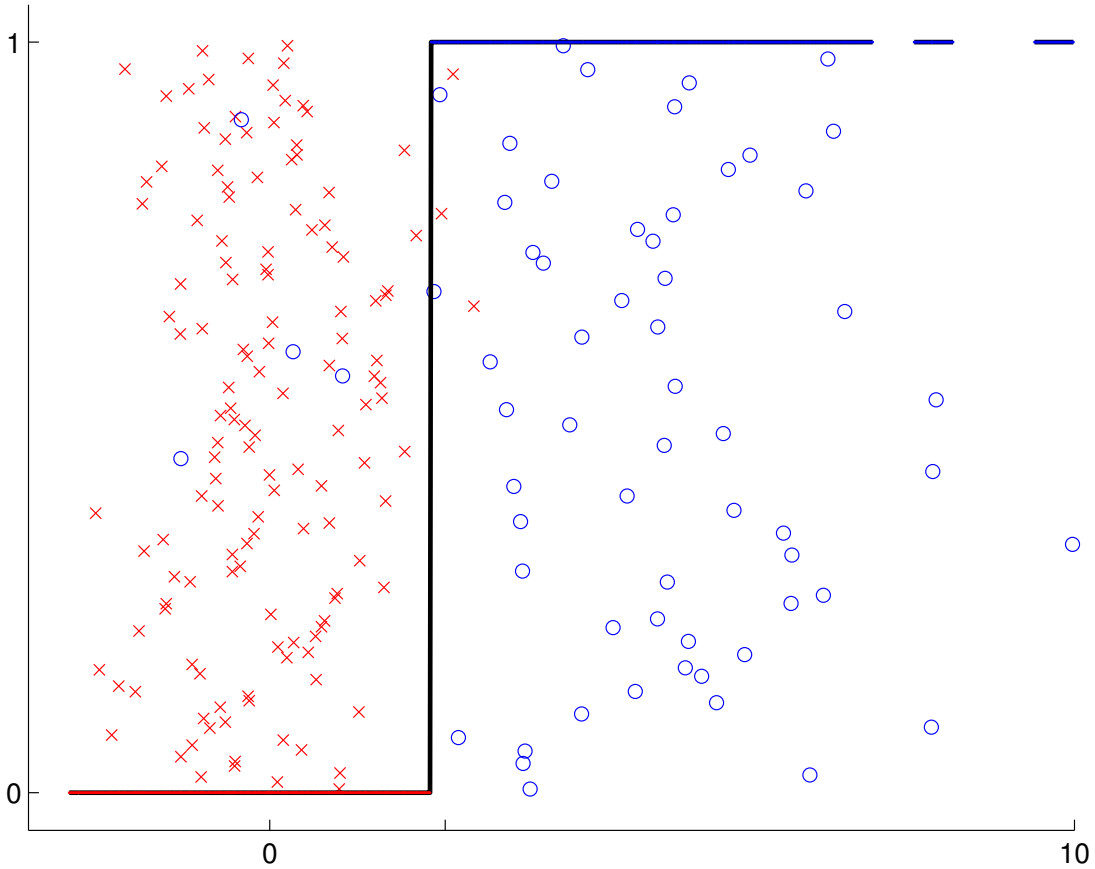
Exemple : régression linéaire



$$S^{\text{lin}} = \left\{ x \in \mathcal{X} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle = \sum_{j=1}^{p} w_j x_j \,/\, \mathbf{w} \in \mathbb{R}^p \right\}$$

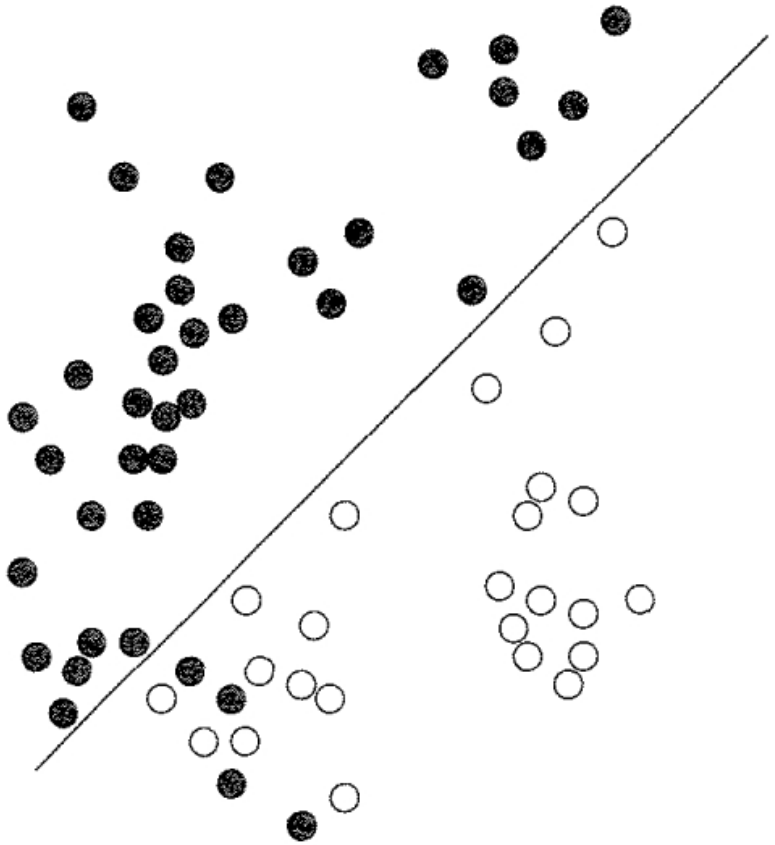
Figure : [Cornillon and Matzner-Løber, 2011]

Exemple : partition en classification



$$S^{\text{part}}_{\text{class}}(\mathcal{A}) = S^{\text{part}}_{\text{reg}}(\mathcal{A}) \cap \mathcal{F}(\mathcal{X}, \{0,1\})$$

Exemple : discrimination linéaire



$$S_{\text{class}}^{\text{lin}} = \{x \in \mathcal{X} \mapsto \mathbb{1}_{\langle \mathbf{w}, \mathbf{x} \rangle \geqslant 0} \,/\, \mathbf{w} \in \mathbb{R}^p\}$$

Figure : [Devroye et al., 1996]

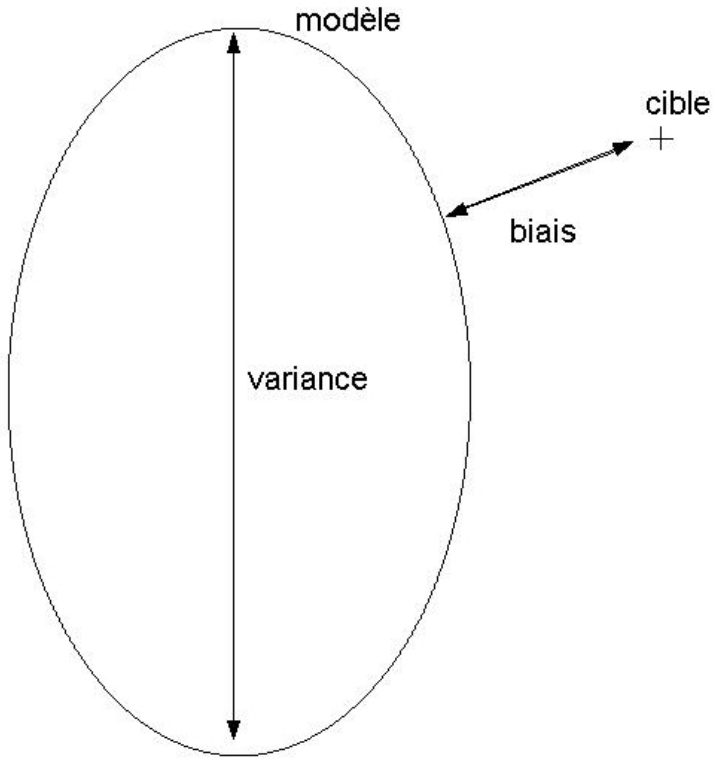
Erreurs d'approximation et d'estimation

Erreur d'approximation (ou biais)

$$\ell(f^\star, S) := \inf_{f \in S} \ell(f^\star, f)$$

Erreur d'estimation (ou variance)

$$\mathcal{R}_P(\widehat{f}_S) - \inf_{f \in S} \mathcal{R}_P(f)$$
$$= \ell(f^\star, \widehat{f}_S) - \ell(f^\star, S) \geqslant 0$$



$$\ell(f^\star, \widehat{f}_S) = \underbrace{\ell(f^\star, S)}_{\text{erreur d'approximation}} + \underbrace{\ell(f^\star, \widehat{f}_S) - \ell(f^\star, S)}_{\text{erreur d'estimation}}$$

Majoration générale de l'erreur d'estimation

Pour tout $g \in S$, on a :

$$\begin{aligned}
&\mathcal{R}_P(\widehat{f}_S) - \mathcal{R}_P(g) \\
&= \underbrace{\mathcal{R}_P(\widehat{f}_S) - \widehat{\mathcal{R}}_n(\widehat{f}_S)}_{\leqslant \sup_{f\in S}|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)|} + \underbrace{\widehat{\mathcal{R}}_n(\widehat{f}_S) - \widehat{\mathcal{R}}_n(g)}_{\leqslant 0} + \underbrace{\widehat{\mathcal{R}}_n(g) - \mathcal{R}_P(g)}_{\leqslant \sup_{f\in S}|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)|} \\
&\leqslant 2 \sup_{f\in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right| .
\end{aligned}$$

Majoration générale de l'erreur d'estimation

Pour tout $g \in S$, on a :

$$\begin{aligned}
&\mathcal{R}_P(\widehat{f}_S) - \mathcal{R}_P(g) \\
&= \underbrace{\mathcal{R}_P(\widehat{f}_S) - \widehat{\mathcal{R}}_n(\widehat{f}_S)}_{\leqslant \sup_{f\in S}|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)|} + \underbrace{\widehat{\mathcal{R}}_n(\widehat{f}_S) - \widehat{\mathcal{R}}_n(g)}_{\leqslant 0} + \underbrace{\widehat{\mathcal{R}}_n(g) - \mathcal{R}_P(g)}_{\leqslant \sup_{f\in S}|\mathcal{R}_P(f)-\widehat{\mathcal{R}}_n(f)|} \\
&\leqslant 2 \sup_{f\in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right| .
\end{aligned}$$

En prenant le sup sur $g \in S$, on obtient :

Proposition (2.8)

Si $\widehat{f}_S$ minimise le risque empirique sur S :

$$\ell(f^\star, \widehat{f}_S) - \ell(f^\star, S) \leqslant 2 \sup_{f\in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right| .$$

Majoration générale de l'erreur d'estimation

Proposition (2.8)

Si $\widehat{f}_S$ minimise le risque empirique sur S :

$$\ell(f^\star, \widehat{f}_S) - \ell(f^\star, S) \leqslant 2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right| .$$

Proposition (2.9)

Si $\widehat{f}_S$ minimise le risque empirique sur S :

$$\mathbb{E}\left[\ell(f^\star, \widehat{f}_S) - \ell(f^\star, S)\right] \leqslant \mathbb{E}\left[\sup_{f \in S} \left\{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right\} \right] .$$

Erreur d'estimation et minimisation approchée

Proposition (2.8)

Si $\widehat{f}_{S,\rho}$ minimise le risque empirique sur S à ρ près :

$$\ell(f^\star, \widehat{f}_{S,\rho}) - \ell(f^\star, S) \leqslant 2 \sup_{f \in S} \left| \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right| + \rho \,.$$

Proposition (2.9)

Si $\widehat{f}_{S,\rho}$ minimise le risque empirique sur S à ρ près :

$$\mathbb{E}\left[\ell(f^\star, \widehat{f}_{S,\rho}) - \ell(f^\star, S)\right] \leqslant \mathbb{E}\left[\sup_{f \in S} \left\{ \mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f) \right\} \right] + \rho \,.$$

Coût borné, modèle fini

Proposition (2.11)

Si $\widehat{f}_S$ minimise le risque empirique sur S et si

$$c\big(f(X), Y\big) \in [a, b] \qquad \textit{presque sûrement,}$$

alors,

$$\mathbb{E}\big[\ell(f^\star, \widehat{f}_S)\big] \leqslant \ell(f^\star, S) + (b-a)\sqrt{\frac{\ln(\operatorname{Card} S)}{2n}}\,.$$

Coût borné, modèle fini

Proposition (2.11)

Si $\widehat{f}_S$ minimise le risque empirique sur S et si

$$c\big(f(X), Y\big) \in [a, b] \qquad \textit{presque sûrement,}$$

alors,

$$\mathbb{E}\big[\ell(f^\star, \widehat{f}_S)\big] \leqslant \ell(f^\star, S) + (b-a)\sqrt{\frac{\ln(\operatorname{Card} S)}{2n}}\,.$$

Application : règle par partition finie en classification 0–1 :

$$\mathbb{E}\big[\ell(f^\star, \widehat{f}_{\mathcal{A}}^{\mathrm{part-class}})\big] \leqslant \ell(f^\star, S_{\mathrm{class}}^{\mathrm{part}}(\mathcal{A})) + \sqrt{\frac{\ln(2)}{2}}\sqrt{\frac{\operatorname{Card}(\mathcal{A})}{n}}$$

Classification 0–1, modèle quelconque

- Tout se passe comme si S était fini de cardinal :

$$\text{Card}\left\{ (f(X_i))_{1 \leqslant i \leqslant n} \,/\, f \in S \right\} =: \exp(H_S(X_1, \ldots, X_n))$$

Classification 0–1, modèle quelconque

- Tout se passe comme si S était fini de cardinal :

$$\text{Card}\left\{ \left(f(X_i)\right)_{1\leqslant i\leqslant n} / \, f \in S \right\} =: \exp\left(H_S(X_1, \ldots, X_n)\right)$$

- On a toujours :

$$\sup_{x_1,\ldots,x_n \in \mathcal{X}} \text{Card}\left\{ \left(f(x_i)\right)_{1\leqslant i\leqslant n} / \, f \in S \right\} \leqslant 2^n$$

Classification 0–1, modèle quelconque

- Tout se passe comme si S était fini de cardinal :

$$\text{Card}\left\{ (f(X_i))_{1 \leqslant i \leqslant n} \,/\, f \in S \right\} =: \exp(H_S(X_1, \ldots, X_n))$$

- On a toujours :

$$\sup_{x_1, \ldots, x_n \in \mathcal{X}} \text{Card}\left\{ (f(x_i))_{1 \leqslant i \leqslant n} \,/\, f \in S \right\} \leqslant 2^n$$

- Classe de Vapnik-Chervonenkis de dimension $V(S)$:
  - égalité pour $n = 1, \ldots, V(S)$ ($\exists x_1, \ldots, x_n$ « pulvérisés » par S)
  - inégalité stricte pour $n > V(S)$
- Exemple : discrimination linéaire dans $\mathbb{R}^d$: $V(S) = d$

Classification 0–1, modèle quelconque

- Tout se passe comme si S était fini de cardinal :

$$\mathrm{Card}\left\{ \left(f(X_i)\right)_{1\leqslant i\leqslant n} / \, f \in S \right\} =: \exp\left(H_S(X_1,\ldots,X_n)\right)$$

- On a toujours :

$$\sup_{x_1,\ldots,x_n\in\mathcal{X}} \mathrm{Card}\left\{ \left(f(x_i)\right)_{1\leqslant i\leqslant n} / \, f \in S \right\} \leqslant 2^n$$

- Classe de Vapnik-Chervonenkis de dimension $V(S)$:
  - égalité pour $n = 1,\ldots,V(S)$ ($\exists x_1,\ldots,x_n$ « pulvérisés » par S)
  - inégalité stricte pour $n > V(S)$
- Exemple : discrimination linéaire dans $\mathbb{R}^d$: $V(S) = d$
- Lemme de Sauer :
  $n > 2V(S) \Rightarrow H_S(x_1,\ldots,x_n) \leqslant V(S)\ln(\mathrm{e}n/V(S))$

Classification 0–1, modèle quelconque : récapitulatif

$$
\begin{aligned}
&\mathbb{E}\big[\ell(f^\star, \widehat{f}_S)\big] \\
&\leqslant \ell(f^\star, S) + \mathbb{E}\left[\sup_{f \in S}\{\mathcal{R}_P(f) - \widehat{\mathcal{R}}_n(f)\}\right] && \text{(Prop. 2.9)} \\
&\leqslant \ell(f^\star, S) + 2\mathbb{E}\left[\sup_{f \in S}\left\{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i c(f(X_i), Y_i)\right\}\right] && \text{(symétrisation)} \\
&\leqslant \ell(f^\star, S) + \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\Big[\sqrt{H_S(X_1, \ldots, X_n)}\Big] && \text{(entropie combinatoire)} \\
&\leqslant \ell(f^\star, S) + \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{\sup_{x_1, \ldots, x_n \in \mathcal{X}} H_S(x_1, \ldots, x_n)} \\
&\leqslant \ell(f^\star, S) + 2\sqrt{2}\sqrt{\frac{V(S)}{n}\ln\left(\frac{\mathrm{e}n}{V(S)}\right)}\,. && \text{(Vapnik-Chervonenkis)}
\end{aligned}
$$

Classification zéro-erreur

Proposition (2.12)

Hypothèses :

- *classification* 0–1
- $\eta(X) \in \{0,1\}$ *p.s.* (*classification zéro-erreur*),
- $\widehat{f}_S$ *minimise le risque empirique sur* S *fini avec* $f^\star \in S$

Alors :

$$\mathbb{E}\Big[\ell\big(f^\star, \widehat{f}_S(D_n)\big)\Big] \leqslant \frac{1 + \ln\big(\mathrm{Card}(S)\big)}{n} .$$

Classification zéro-erreur

Proposition (2.12)

Hypothèses :

- *classification* 0–1
- $\eta(X) \in \{0,1\}$ *p.s.* (*classification zéro-erreur*),
- $\widehat{f}_S$ *minimise le risque empirique sur* S *fini avec* $f^\star \in S$

Alors :

$$\mathbb{E}\Big[\ell\big(f^\star, \widehat{f}_S(D_n)\big)\Big] \leqslant \frac{1 + \ln\big(\mathrm{Card}(S)\big)}{n} .$$

Application : règle par partition finie en classification 0–1 :

$$\text{si } f^\star \in S^{\mathrm{part}}_{\mathrm{class}}(\mathcal{A}), \qquad \mathbb{E}\big[\ell\big(f^\star, \widehat{f}^{\,\mathrm{part-class}}_{\mathcal{A}}\big)\big] \leqslant \frac{1 + \ln(2)\,\mathrm{Card}(\mathcal{A})}{n} .$$

Classification zéro-erreur

Proposition (2.12)

Hypothèses :

- *classification* 0–1
- $\eta(X) \in \{0,1\}$ *p.s.* (*classification zéro-erreur*),
- $\widehat{f}_S$ *minimise le risque empirique sur* S *fini avec* $f^\star \in S$

Alors :

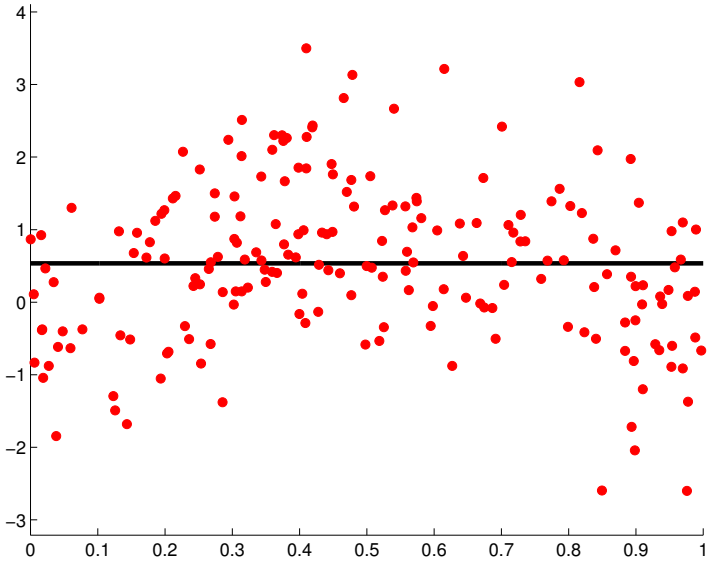
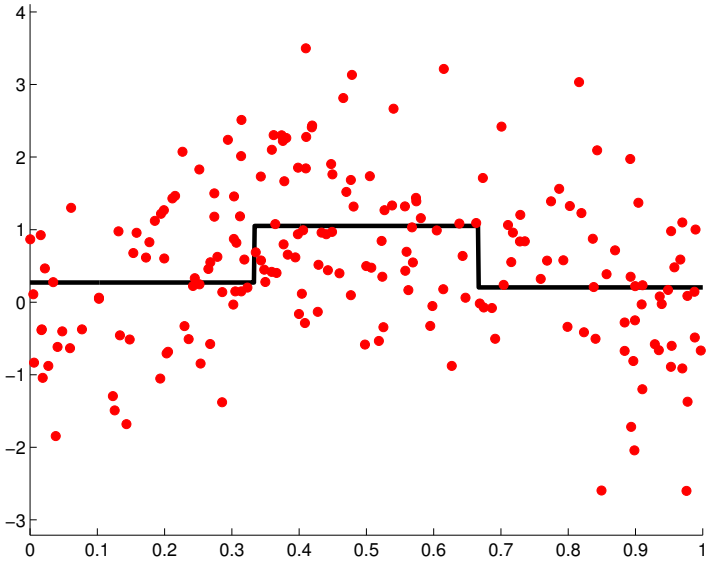
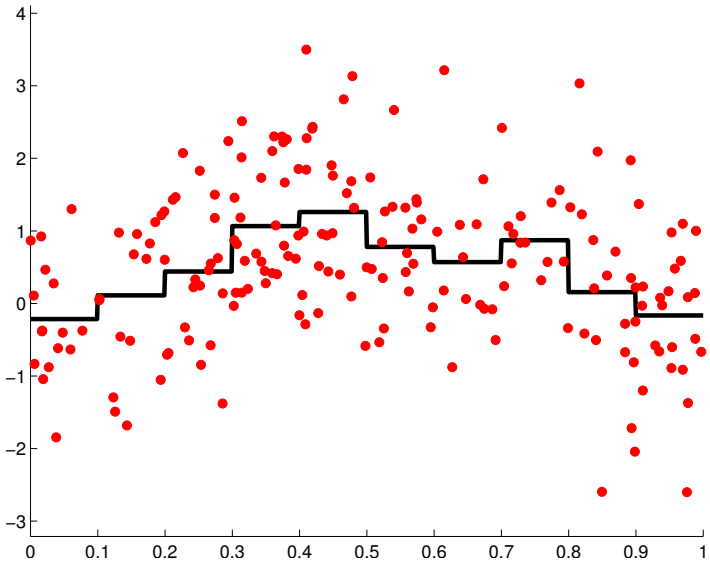
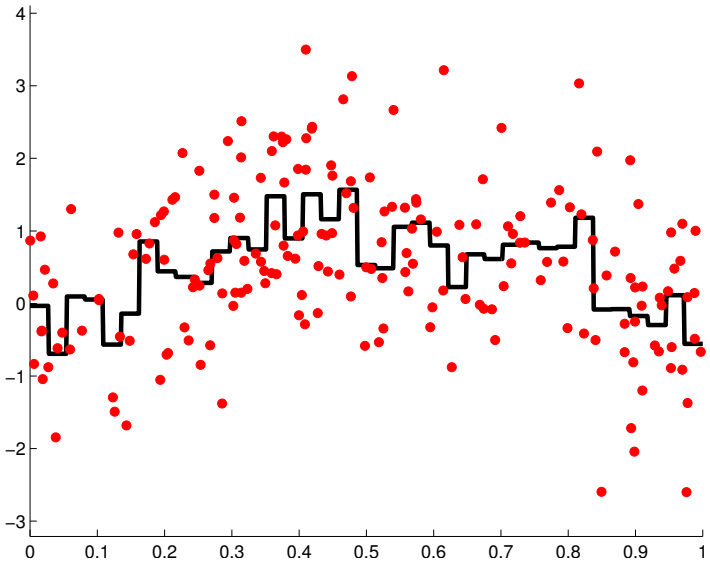
$$\mathbb{E}\Big[\ell\big(f^\star, \widehat{f}_S(D_n)\big)\Big] \leqslant \frac{1 + \ln\big(\mathrm{Card}(S)\big)}{n} .$$

Application : règle par partition finie en classification 0–1 :

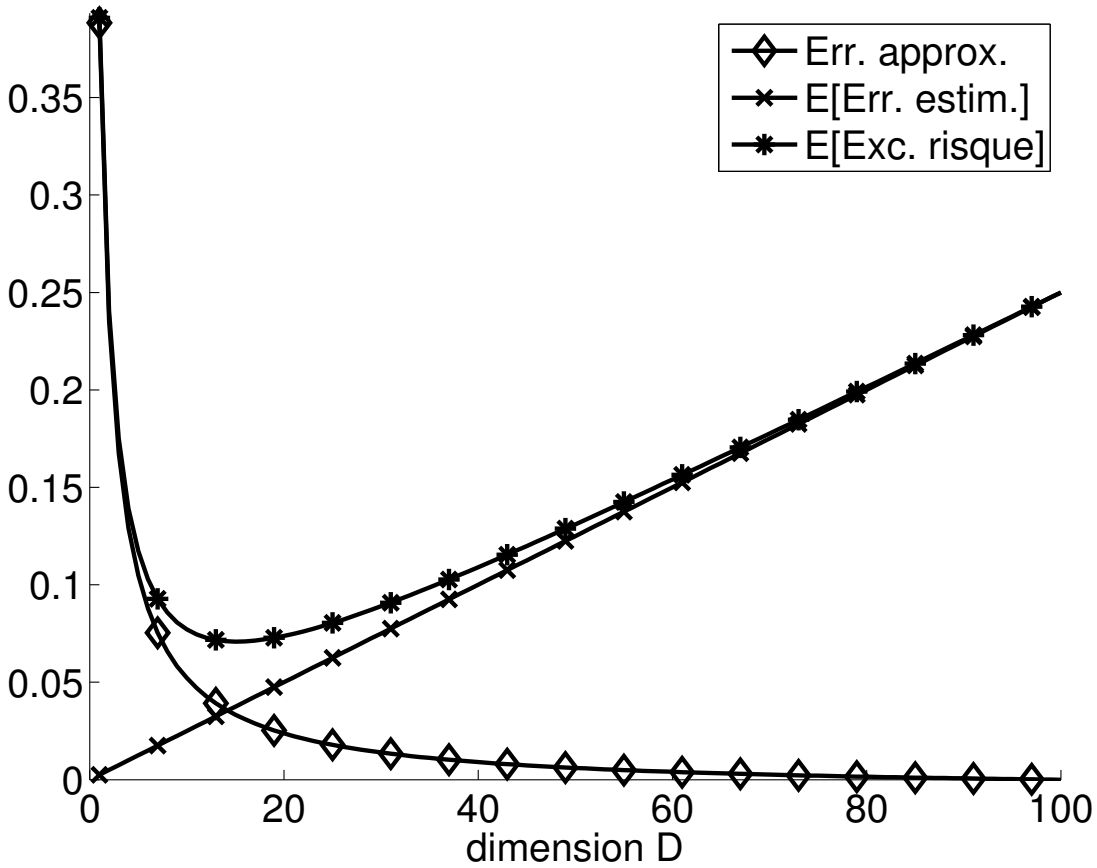
$$\text{si } f^\star \in S_{\text{class}}^{\text{part}}(\mathcal{A}), \qquad \mathbb{E}\big[\ell\big(f^\star, \widehat{f}_{\mathcal{A}}^{\text{part}-\text{class}}\big)\big] \leqslant \frac{1 + \ln(2)\,\mathrm{Card}(\mathcal{A})}{n} .$$

Se généralise à S quelconque (entropie combinatoire, dimension de Vapnik-Chervonenkis, ...).

Sélection de modèles : partitions cubiques

Compromis biais-variance



Sous-apprentissage

Sur-apprentissage

Sélection de modèles

- famille de modèles $(S_m)_{m \in \mathcal{M}} \Rightarrow$ famille des règles $\widehat{f}_m = \widehat{f}_{S_m}$

Sélection de modèles

- famille de modèles $(S_m)_{m \in \mathcal{M}} \Rightarrow$ famille des règles $\widehat{f}_m = \widehat{f}_{S_m}$

$\Rightarrow$ Objectif de prévision : choisir $\widehat{m} = \widehat{m}(D_n)$ tel que $\mathcal{R}(\widehat{f}_{\widehat{m}(D_n)}(D_n))$ est minimal ?

Sélection de modèles

- famille de modèles $(S_m)_{m\in\mathcal{M}} \Rightarrow$ famille des règles $\widehat{f}_m = \widehat{f}_{S_m}$

$\Rightarrow$ Objectif de prévision : choisir $\widehat{m} = \widehat{m}(D_n)$ tel que $\mathcal{R}(\widehat{f}_{\widehat{m}(D_n)}(D_n))$ est minimal ?

- Optimalité asymptotique :

$$\frac{\ell(f^\star, \widehat{f}_{\widehat{m}(D_n)}(D_n))}{\inf_{m\in\mathcal{M}}\left\{\ell(f^\star, \widehat{f}_m(D_n))\right\}} \xrightarrow[n\to+\infty]{\text{p.s.}} 1\,.$$

- Inégalité oracle (en espérance / avec grande probabilité) :

$$\ell(f^\star, \widehat{f}_{\widehat{m}(D_n)}(D_n)) \leqslant C_n \inf_{m\in\mathcal{M}}\left\{\ell(f^\star, \widehat{f}_m(D_n))\right\} + R_n$$

Sélection de modèles

- famille de modèles $(S_m)_{m \in \mathcal{M}} \Rightarrow$ famille des règles $\widehat{f}_m = \widehat{f}_{S_m}$

$\Rightarrow$ Objectif de prévision : choisir $\widehat{m} = \widehat{m}(D_n)$ tel que $\mathcal{R}(\widehat{f}_{\widehat{m}(D_n)}(D_n))$ est minimal ?

- Optimalité asymptotique :

$$\frac{\ell(f^\star, \widehat{f}_{\widehat{m}(D_n)}(D_n))}{\inf_{m \in \mathcal{M}} \left\{ \ell(f^\star, \widehat{f}_m(D_n)) \right\}} \xrightarrow[n \to +\infty]{\text{p.s.}} 1 \,.$$

- Inégalité oracle (en espérance / avec grande probabilité) :

$$\ell(f^\star, \widehat{f}_{\widehat{m}(D_n)}(D_n)) \leqslant C_n \inf_{m \in \mathcal{M}} \left\{ \ell(f^\star, \widehat{f}_m(D_n)) \right\} + R_n$$

- Attention à l'interprétation de $\widehat{m}$: le meilleur modèle peut être faux / le vrai modèle peut être moins bon qu'un modèle plus simple.

Méthodes de sélection de modèle

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\mathrm{crit}(m; D_n)\}$$

Méthodes de sélection de modèle

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\text{crit}(m; D_n)\}$$

- Estimation sans biais du risque (AIC, C_p, validation croisée...) :

$$\mathbb{E}\big[\text{crit}(m; D_n)\big] \approx \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] \qquad \text{(à translation près)}$$

$\Rightarrow$ $\text{crit}(m; D_n) \approx \mathcal{R}\big(\widehat{f}_{\widehat{m}(D_n)}(D_n)\big)$ simultanément pour tous les $m \in \mathcal{M}$ (si $\mathcal{M}$ « petite »)

$\Rightarrow$ inégalité oracle « optimale »

Méthodes de sélection de modèle

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\mathrm{crit}(m; D_n)\}$$

- Estimation sans biais du risque (AIC, C_p, validation croisée...) :

$$\mathbb{E}\big[\mathrm{crit}(m; D_n)\big] \approx \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] \qquad \text{(à translation près)}$$

$\Rightarrow$ $\mathrm{crit}(m; D_n) \approx \mathcal{R}\big(\widehat{f}_{\widehat{m}(D_n)}(D_n)\big)$ simultanément pour tous les $m \in \mathcal{M}$ (si $\mathcal{M}$ « petite »)

$\Rightarrow$ inégalité oracle « optimale »

- Avec une majoration du risque :

$$\mathrm{crit}(m; D_n) \geqslant \mathcal{R}_P\big(\widehat{f}_m(D_n)\big) \qquad \text{(à translation près)}$$

$\Rightarrow$ inégalité oracle si la majoration n'est pas trop large

Plan

1. Prévision
2. Régression et classification
3. Minimisation du risque empirique
4. Moyennes locales
5. On n'a rien sans rien
6. Conclusion : enjeux de l'apprentissage

Moyennes locales

Définition (2.7)

$\forall x, x_1, \ldots, x_n \in \mathcal{X}, \forall i \in \{1, \ldots, n\}, \qquad W_i(x_{1\ldots n}; x) \geqslant 0$

- Régression :

$$\widehat{\eta}\left((x_i, y_i)_{1 \leqslant i \leqslant n}; x\right) := \sum_{i=1}^{n} W_i(x_{1\ldots n}; x) y_i$$

Moyennes locales

Définition (2.7)

$\forall x, x_1, \ldots, x_n \in \mathcal{X}, \forall i \in \{1, \ldots, n\}, \qquad W_i(x_{1\ldots n}; x) \geqslant 0$

- Régression :

$$\widehat{\eta}\left((x_i, y_i)_{1 \leqslant i \leqslant n}; x\right) := \sum_{i=1}^{n} W_i(x_{1\ldots n}; x) y_i$$

- Classification (plug-in) :

$$\widehat{f}\left((x_i, y_i)_{1 \leqslant i \leqslant n}; x\right) := \mathbb{1}_{\widehat{\eta}((x_i, y_i)_{1 \leqslant i \leqslant n}; x) > 1/2}$$

Moyennes locales

Définition (2.7)

$\forall x, x_1, \ldots, x_n \in \mathcal{X}, \forall i \in \{1, \ldots, n\}, \qquad W_i(x_{1\ldots n}; x) \geqslant 0$

- Régression :

$$\widehat{\eta}((x_i, y_i)_{1 \leqslant i \leqslant n}; x) := \sum_{i=1}^{n} W_i(x_{1\ldots n}; x) y_i$$

- Classification (plug-in) :

$$\widehat{f}\,((x_i, y_i)_{1 \leqslant i \leqslant n}; x) := \mathbb{1}_{\widehat{\eta}((x_i, y_i)_{1 \leqslant i \leqslant n}; x) > 1/2}$$

- Hypothèse générale :

$$\forall x, x_1, \ldots, x_n \in \mathcal{X}, \qquad \sum_{i=1}^{n} W_i(x_{1\ldots n}; x) \approx 1$$

Moyennes locales : exemples

- Partitions :

$$W_i^{\mathcal{A}}(x_{1\dots n}; x) := \begin{cases} \frac{\mathbb{1}_{x_i \in \mathcal{A}(x)}}{\sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)}} & \text{si } \sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)} > 0 \\ 0 & \text{sinon.} \end{cases}$$

$$\Rightarrow \qquad \sum_{i=1}^n W_i^{\mathcal{A}}(x_{1\dots n}; x) = \mathbb{1}_{\sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)} > 0} \leqslant 1$$

Moyennes locales : exemples

- Partitions :

$$W_i^{\mathcal{A}}(x_{1\dots n}; x) := \begin{cases} \dfrac{\mathbb{1}_{x_i \in \mathcal{A}(x)}}{\sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)}} & \text{si } \sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)} > 0 \\ 0 \qquad \text{sinon.} \end{cases}$$

$$\Rightarrow \qquad \sum_{i=1}^n W_i^{\mathcal{A}}(x_{1\dots n}; x) = \mathbb{1}_{\sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)} > 0} \leqslant 1$$

- k plus proches voisins :

$$W_i^{k-\mathrm{ppv}}(x_{1\dots n}; x) := \frac{1}{k} \times \mathbb{1}_{\left\{x_i \in \{k\text{-ppv de } x \text{ parmi } x_{1\dots n}\}\right\}}$$

Moyennes locales : exemples

- Partitions :

$$W_i^{\mathcal{A}}(x_{1\dots n}; x) := \begin{cases} \dfrac{\mathbb{1}_{x_i \in \mathcal{A}(x)}}{\sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)}} & \text{si } \sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)} > 0 \\ 0 & \text{sinon.} \end{cases}$$

$$\Rightarrow \qquad \sum_{i=1}^n W_i^{\mathcal{A}}(x_{1\dots n}; x) = \mathbb{1}_{\sum_{j=1}^n \mathbb{1}_{x_j \in \mathcal{A}(x)} > 0} \leqslant 1$$

- k plus proches voisins :

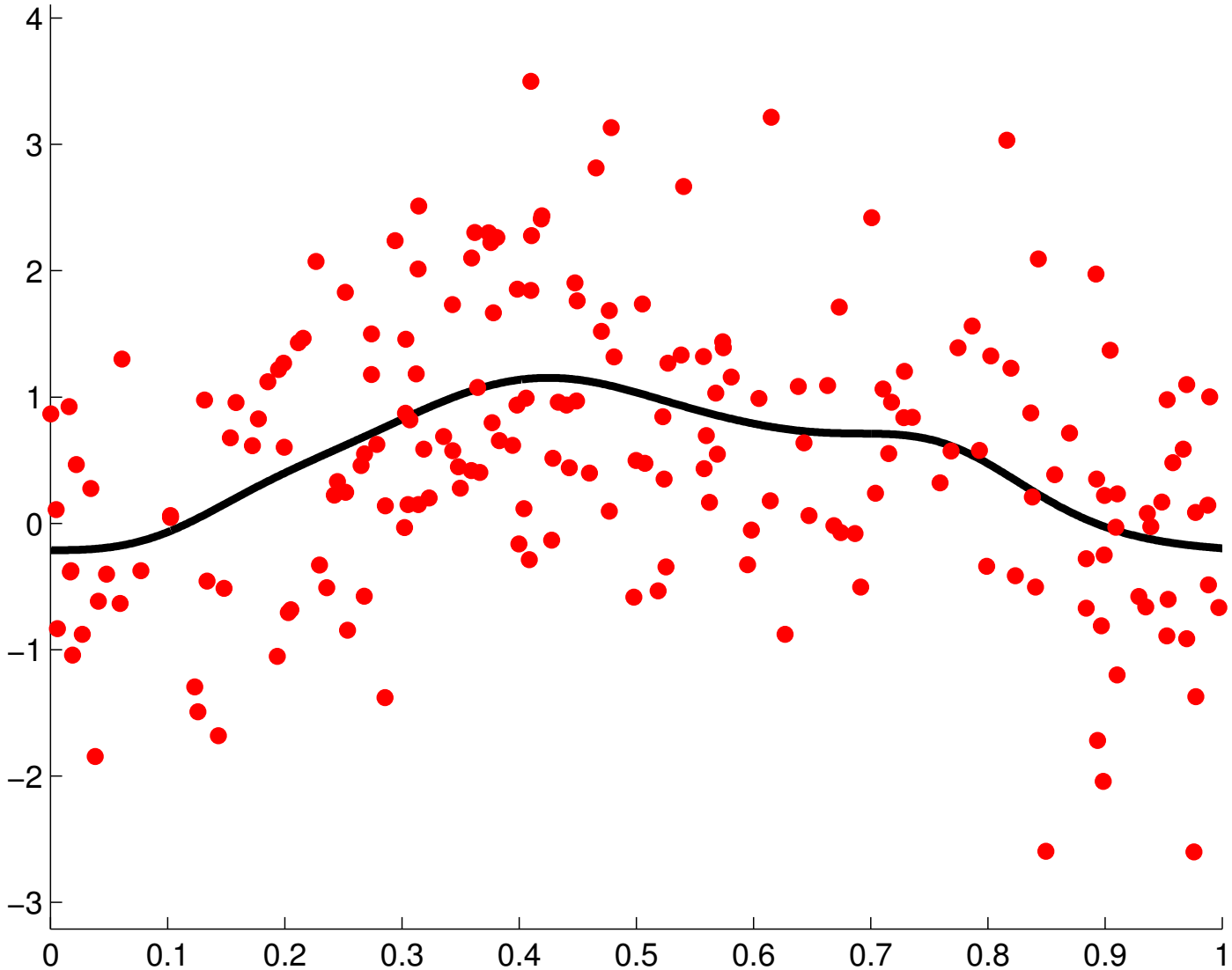
$$W_i^{k-\mathrm{ppv}}(x_{1\dots n}; x) := \frac{1}{k} \times \mathbb{1}_{\left\{x_i \in \{k\text{-ppv de } x \text{ parmi } x_{1\dots n}\}\right\}}$$

- Noyaux ($\mathcal{X} = \mathbb{R}^p$) :

$$W_i^{K,h}(x_{1\dots n}; x) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)}$$

Convention : $\frac{0}{0} = 0$.

Exemple : noyaux (Nadaraya-Watson)



Consistance : Théorème de Stone

Théorème (2.2)

Hypothèses :

- $\mathcal{X} = \mathbb{R}^p$, *classification, coût 0–1*

(a) $\sum_{i=1}^n W_{i,n}(X_{1\ldots n}; X) \xrightarrow[n\to\infty]{(L^1)} 1 \quad et \quad \sum_{i=1}^n W_{i,n}(X_{1\ldots n}; X) \leqslant c_a \ p.s.$

(b) *« pas de sous-apprentissage »* : $\forall a > 0$,
$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^n W_{i,n}(X_{1\ldots n}; X) \mathbb{1}_{\|X_i - X\| \geqslant a}\right] = 0$

(c) *« pas de sur-apprentissage »* :
$\lim_{n\to\infty} \mathbb{E}\left[\max_{1\leqslant i\leqslant n} W_{i,n}(X_{1\ldots n}; X)\right] = 0$

(d) $\exists c_d > 0, \forall f \in L^1(P_X),\ \mathbb{E}\left[\sum_{i=1}^n W_{i,n}(X_{1\ldots n}; X) f(X_i)\right] \leqslant c_d \mathbb{E}\left[f(X)\right]$

Alors, la règle par moyennes locales $\widehat{f}$ *associée aux poids* $(W_{i,n})_{1\leqslant i\leqslant n}$ *est faiblement consistante pour* P.

Partitions : consistance universelle

Corollaire (2.1)

Hypothèses :

- $\mathcal{X} = \mathbb{R}^p$, *classification, coût 0–1*

(b') *« pas de sous-apprentissage »* :
$\lim_{n\to\infty} \max_{A\in\mathcal{A}_n} \{\mathrm{diam}(A)\} = 0$

(c') *« pas de sur-apprentissage »* : $\forall r > 0$,
$\lim_{n\to\infty} \frac{\mathrm{Card}\{A\in\mathcal{A}_n \,/\, A\cap\mathcal{B}(0,r)\neq\emptyset\}}{n} = 0$

Alors, $\widehat{f}^{\,\mathrm{part-class}}_{(\mathcal{A}_n)}$ *est faiblement universellement consistante.*

Partitions : consistance universelle

Corollaire (2.1)

Hypothèses :

- $\mathcal{X} = \mathbb{R}^p$, *classification, coût 0–1*

(b') *« pas de sous-apprentissage »* :
$\lim_{n\to\infty} \max_{A\in\mathcal{A}_n} \{\mathrm{diam}(A)\} = 0$

(c') *« pas de sur-apprentissage »* : $\forall r > 0$,
$\lim_{n\to\infty} \frac{\mathrm{Card}\{A\in\mathcal{A}_n \,/\, A\cap\mathcal{B}(0,r)\neq\emptyset\}}{n} = 0$

Alors, $\widehat{f}^{\,\mathrm{part-class}}_{(\mathcal{A}_n)}$ *est faiblement universellement consistante.*

Exemple : partitions cubiques avec

(b'') $h_n \to 0$

(c'') $nh_n^p \to 0$

Plan

1. Prévision
2. Régression et classification
3. Minimisation du risque empirique
4. Moyennes locales
5. On n'a rien sans rien
6. Conclusion : enjeux de l'apprentissage

Consistance universelle uniforme ?

- consistance faible universelle :

$$\sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}} \lim_{n\to+\infty} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell_P\left(f_P^\star, \widehat{f}(D_n)\right)\right] = 0$$

- consistance faible universelle uniforme :

$$\lim_{n\to+\infty} \sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell_P\left(f_P^\star, \widehat{f}(D_n)\right)\right] = 0$$

c'est-à-dire, vitesse d'apprentissage valable quelle que soit P ?

Consistance universelle uniforme ?

- consistance faible universelle :

$$\sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}} \lim_{n\to+\infty} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell_P\left(f_P^\star, \widehat{f}(D_n)\right)\right] = 0$$

- consistance faible universelle uniforme :

$$\lim_{n\to+\infty} \sup_{P \text{ loi sur } \mathcal{X}\times\mathcal{Y}} \mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell_P\left(f_P^\star, \widehat{f}(D_n)\right)\right] = 0$$

  c'est-à-dire, vitesse d'apprentissage valable quelle que soit P ?

- Oui si $\mathcal{X}$ est fini.
- Non sinon.

On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$

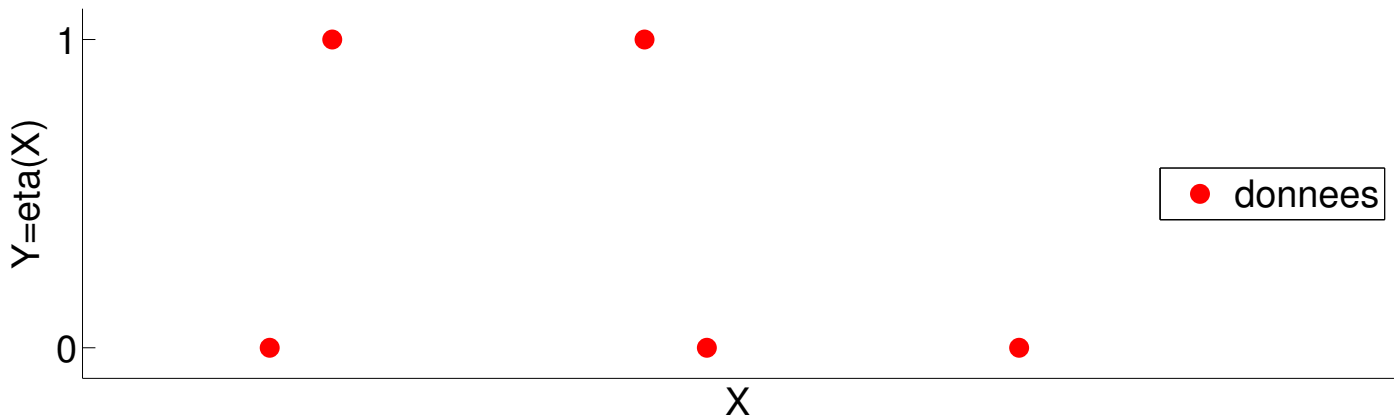
On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$



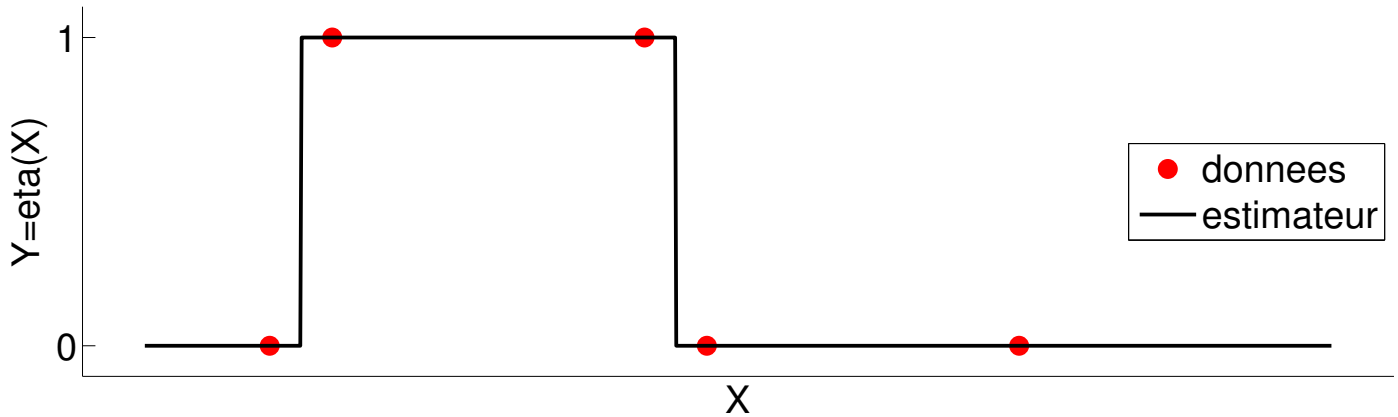
On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$



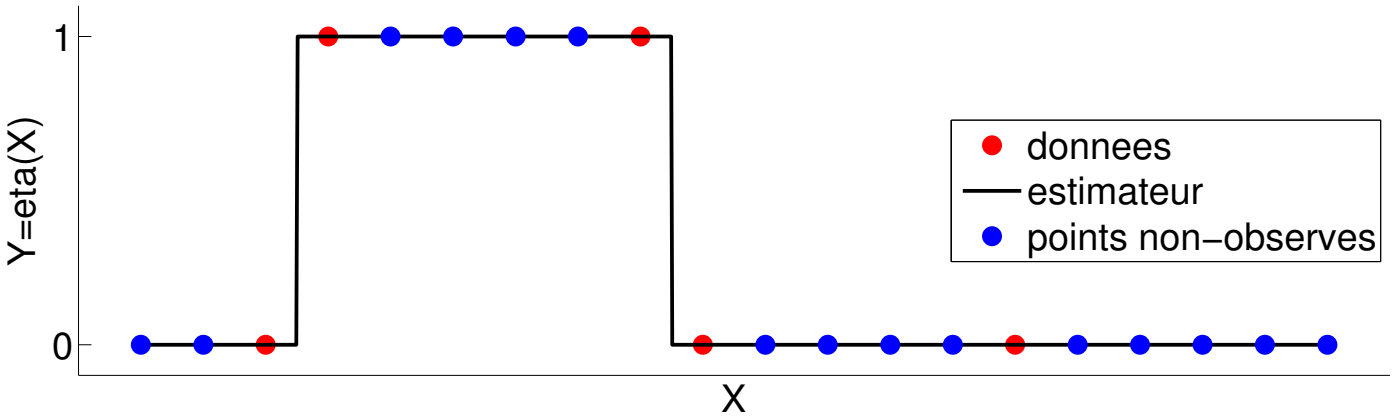
On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$



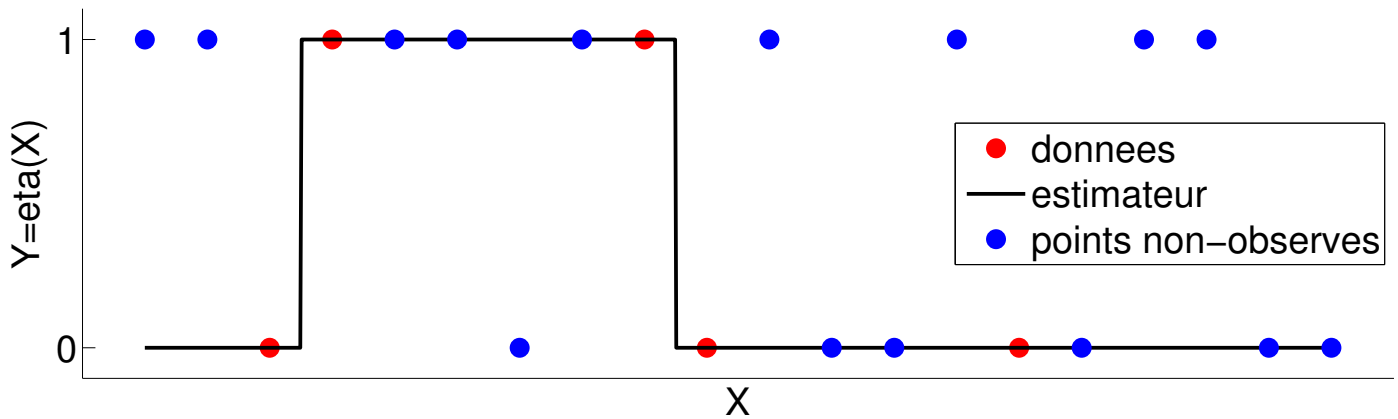
On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$



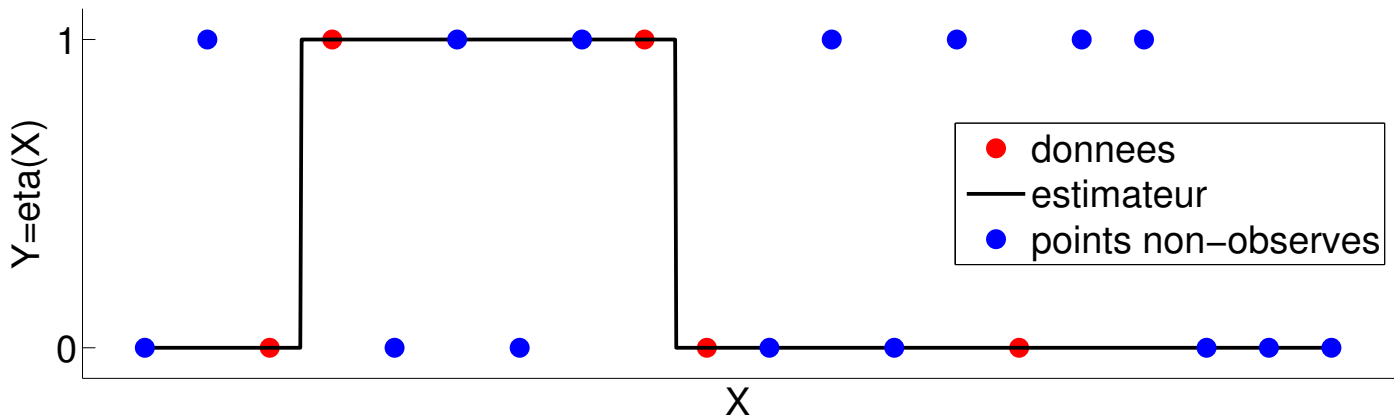
On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$



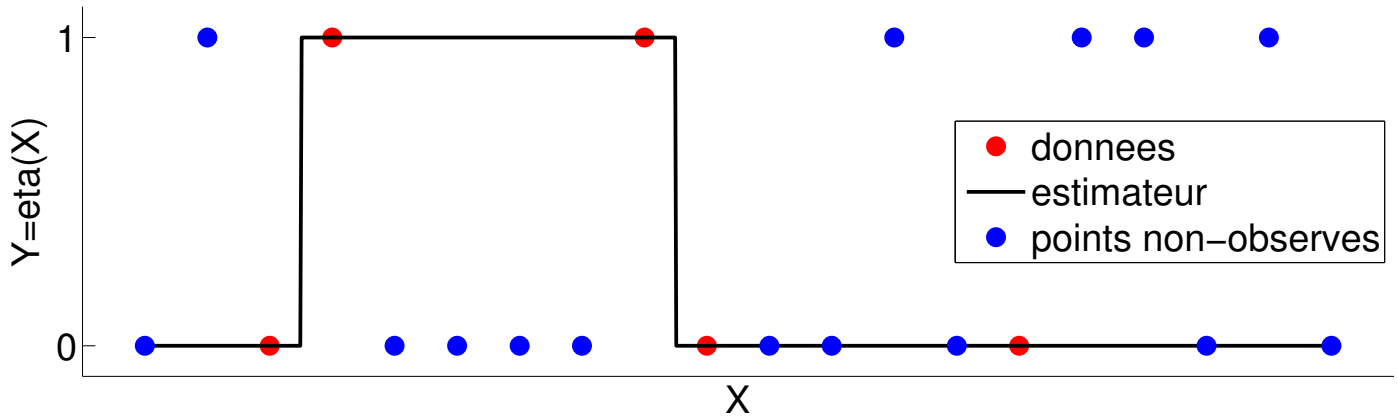
On n'a rien sans rien

Théorème (2.3)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \right\} \geqslant \frac{1}{2} \cdot$$

$\mathcal{M}_1(\mathcal{X} \times \{0,1\})$: ensemble des mesures de probabilité sur $\mathcal{X} \times \{0,1\}$



Classification sur $\mathcal{X}$ fini

Règle de majorité $\widehat{f}^{\,\mathrm{maj}}$:
$\widehat{f}^{\,\mathrm{maj}}\big((x_i, y_i)_{1 \leqslant i \leqslant n}; x\big)$ réalise un vote majoritaire parmi $\{y_i \,/\, x_i = x\}$

Proposition (2.14)

En classification, avec le coût 0–1, si $\mathcal{X}$ est fini, pour tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P\big(f_P^\star, \widehat{f}^{\,\mathrm{maj}}(D_n)\big) \right] \right\} \leqslant \sqrt{\frac{\ln(2)\,\mathrm{Card}(\mathcal{X})}{2n}} \,.$$

Classification sur $\mathcal{X}$ fini

Règle de majorité $\widehat{f}^{\,\mathrm{maj}}$:
$\widehat{f}^{\,\mathrm{maj}}\big((x_i, y_i)_{1 \leqslant i \leqslant n}; x\big)$ réalise un vote majoritaire parmi $\{y_i \,/\, x_i = x\}$

Proposition (2.14)

En classification, avec le coût 0–1, si $\mathcal{X}$ est fini, pour tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P\big(f_P^\star, \widehat{f}^{\,\mathrm{maj}}(D_n)\big) \right] \right\} \leqslant \sqrt{\frac{\ln(2)\,\mathrm{Card}(\mathcal{X})}{2n}} \,.$$

$\Rightarrow$ $\widehat{f}^{\,\mathrm{maj}}$ est uniformément universellement consistante.

Classification sur $\mathcal{X}$ fini

Règle de majorité $\widehat{f}^{\,\mathrm{maj}}$:
$\widehat{f}^{\,\mathrm{maj}}\big((x_i, y_i)_{1 \leqslant i \leqslant n}; x\big)$ réalise un vote majoritaire parmi $\{y_i \,/\, x_i = x\}$

Proposition (2.14)

En classification, avec le coût 0–1, si $\mathcal{X}$ est fini, pour tout $n \geqslant 1$:

$$\sup_{P \in \mathcal{M}_1(\mathcal{X} \times \{0,1\})} \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P\big(f_P^\star, \widehat{f}^{\,\mathrm{maj}}(D_n)\big) \right] \right\} \leqslant \sqrt{\frac{\ln(2)\,\mathrm{Card}(\mathcal{X})}{2n}} \,.$$

Attention aux constantes : on peut avoir $\mathrm{Card}(\mathcal{X}) \geqslant n$
$\Rightarrow$ attention aux résultats asymptotiques et aux $\mathcal{O}(\cdot)$ qui peuvent cacher de telles constantes dans des termes de 1er ou 2e ordre.

On n'a rien sans rien ?

Vitesse universelle à constante près :

$$\exists c(P) \in \mathbb{R}, \quad \forall n \geqslant 1, \quad \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \leqslant c(P) u_n \xrightarrow[n \to +\infty]{} 0 \, ?$$

On n'a vraiment rien sans rien

Vitesse universelle à constante près : Impossible !

$$\exists c(P) \in \mathbb{R}, \quad \forall n \geqslant 1, \quad \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \leqslant c(P) u_n \xrightarrow[n \to +\infty]{} 0 \, ?$$

Théorème (2.4)

En classification, avec le coût 0–1, si $\mathcal{X}$ est infini, pour toute règle de classification $\widehat{f}$ et toute suite $(a_n)_{n \geqslant 1}$ décroissant vers zéro, avec $a_1 \leqslant 1/16$:

$$\exists P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}), \forall n \geqslant 1, \qquad \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell_P \left(f_P^\star, \widehat{f}(D_n) \right) \right] \geqslant a_n \, .$$

$\Rightarrow$ impossible d'avoir $\dfrac{c(P)}{\log \log n}$ comme borne de risque universelle !

Plan

1. Prévision
2. Régression et classification
3. Minimisation du risque empirique
4. Moyennes locales
5. On n'a rien sans rien
6. Conclusion : enjeux de l'apprentissage

Risque minimax

Définition (2.8)

$\mathcal{P}$: ensemble de lois de probabilité sur $\mathcal{X} \times \mathcal{Y}$.

Risque minimax sur $\mathcal{P}$ avec n observations :

$$\mathcal{R}_{\text{minimax}}(\mathcal{P}, n) := \inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\mathcal{R}_P\big(\widehat{f}(D_n)\big) - \mathcal{R}_P^\star \right]$$

Risque minimax

Définition (2.8)

$\mathcal{P}$: ensemble de lois de probabilité sur $\mathcal{X} \times \mathcal{Y}$.

Risque minimax sur $\mathcal{P}$ avec n observations :

$$\mathcal{R}_{\text{minimax}}(\mathcal{P}, n) := \inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\mathcal{R}_P\big(\widehat{f}(D_n)\big) - \mathcal{R}_P^\star \right]$$

On n'a rien sans rien (classification, coût 0–1) :

$$\forall n \geqslant 1, \qquad \mathcal{R}_{\text{minimax}}\Big(\mathcal{M}_1\big(\mathcal{X} \times \{0,1\}\big)\Big) = \frac{1}{2}$$

$\Rightarrow$ « pile ou face » optimal (en pire cas) !

Risque minimax et adaptation

S classe de Vapnik-Chervonenkis, $h \in [0,1]$.

- Proposition 2.15 :
  $\mathcal{R}_{\text{minimax}}\left(\{P \,/\, f^\star \in S\}, n\right) \propto \sqrt{\frac{V(S)}{n}}$
  (atteint en minimisant le risque empirique sur S)

Risque minimax et adaptation

S classe de Vapnik-Chervonenkis, $h \in [0,1]$.

- Proposition 2.15 :
  $\mathcal{R}_{\text{minimax}}\left(\{P \,/\, f^\star \in S\}, n\right) \propto \sqrt{\frac{V(S)}{n}}$
  (atteint en minimisant le risque empirique sur S)

- Proposition 2.16 : à un facteur $\ln(n)$ près,

$$\mathcal{R}_{\text{minimax}}\left(\left\{P \,/\, f^\star \in S \text{ et } |2\eta_P(X) - 1| \geqslant h \text{ p.s.}\right\}, n\right) \propto \sqrt{\frac{V(S)}{n}} \wedge \frac{V(S)}{nh}$$

(atteint en minimisant le risque empirique sur S)

Risque minimax et adaptation

S classe de Vapnik-Chervonenkis, $h \in [0,1]$.

- Proposition 2.15 :
  $\mathcal{R}_{\text{minimax}}\left(\{P \,/\, f^\star \in S\}, n\right) \propto \sqrt{\frac{V(S)}{n}}$
  (atteint en minimisant le risque empirique sur S)

- Proposition 2.16 : à un facteur $\ln(n)$ près,

$$\mathcal{R}_{\text{minimax}}\left(\left\{P \,/\, f^\star \in S \text{ et } |2\eta_P(X) - 1| \geqslant h \text{ p.s.}\right\}, n\right) \propto \sqrt{\frac{V(S)}{n}} \wedge \frac{V(S)}{nh}$$

(atteint en minimisant le risque empirique sur S)

Adaptation : même règle (quasi) optimale $\forall S$? $\forall h$? $\forall (h, S)$?

Conclusion : enjeux de l'apprentissage

- On ne peut pas apprendre « n'importe quoi »

$\Rightarrow$ Quelles hypothèses « intéressantes » sur P ?

Conclusion : enjeux de l'apprentissage

- On ne peut pas apprendre « n'importe quoi »

$\Rightarrow$ Quelles hypothèses « intéressantes » sur P ?

- Hypothèse $\Rightarrow$ vitesse optimale ? pour quelle(s) règle(s) ?

Conclusion : enjeux de l'apprentissage

- On ne peut pas apprendre « n'importe quoi »

$\Rightarrow$ Quelles hypothèses « intéressantes » sur P ?

- Hypothèse $\Rightarrow$ vitesse optimale ? pour quelle(s) règle(s) ?

- Règle d'apprentissage $\Rightarrow$ hypothèses implicites ?
  ($\Leftrightarrow$ identifier les lois P où elle fonctionne / ne fonctionne pas)
  - minimisation du risque empirique : $f^\star$ « proche » de S
  - moyennes locales (k-ppv) : $f^\star$ « régulière »
  - réseaux de neurones (profonds), forêts aléatoires, SVM : ???

Conclusion : enjeux de l'apprentissage

- On ne peut pas apprendre « n'importe quoi »

$\Rightarrow$ Quelles hypothèses « intéressantes » sur P ?

- Hypothèse $\Rightarrow$ vitesse optimale ? pour quelle(s) règle(s) ?

- Règle d'apprentissage $\Rightarrow$ hypothèses implicites ?
  ($\Leftrightarrow$ identifier les lois P où elle fonctionne / ne fonctionne pas)
  - minimisation du risque empirique : $f^\star$ « proche » de S
  - moyennes locales (k-ppv) : $f^\star$ « régulière »
  - réseaux de neurones (profonds), forêts aléatoires, SVM : ???

- Problèmes clés : choix des paramètres & coût de calcul
  Données massives $\Rightarrow$ nouveaux défis

Questions ?


Cornillon, P.-A. and Matzner-Løber, E. (2011).
Régression avec R.
Pratique R. Springer.


Devroye, L. P., Györfi, L., and Lugosi, G. (1996).
A Probabilistic Theory of Pattern Recognition, volume 31 of *Applications of Mathematics (New York)*.
Springer-Verlag, New New York.


Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The Elements of Statistical Learning.
Springer Series in Statistics. Springer, New York, second edition.
Data Mining, Inference, and Prediction.