IFT-7002 Fondements de l'apprentissage machine

Modèle d'apprentissage général et le compromis biais-complexité

Shai Shalev-Shwartz
The Hebrew University of Jerusalem

Traduit et adapté par Mario Marchand
Université Laval

Hiver 2024

Sommaire

Le modèle PAC agnostique

- On a supposé que les étiquettes étaient générées par un $f \in \mathcal{H}$
- Cette supposition peut s'avérer trop forte !
- Maintenant, soyons plus réaliste en considérant que les étiquettes sont générées par une distribution (que nous ne connaissons pas).

Le modèle PAC agnostique

- Pour le modèle PAC, $\mathcal{D}$ est une distribution sur le domaine $\mathcal{X}$
- Maintenant, considérons que $\mathcal{D}$ est une distribution sur $\mathcal{X} \times \mathcal{Y}$
- Le risque $L_{\mathcal{D}}(h)$ d'une hypothèse h est alors redéfini comme suit :

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x,y) : h(x) \neq y\})$$

- Le critère de "approximativement correct" est alors remplacé par

$$L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

PAC vs. PAC agnostique

	PAC	PAC agnostique
Distribution :	$\mathcal{D}$ sur $\mathcal{X}$	$\mathcal{D}$ sur $\mathcal{X} \times \mathcal{Y}$
Étiquetage :	$f \in \mathcal{H}$	vient de $\mathcal{D}$
Risque	$L_{\mathcal{D},f}(h) = \mathcal{D}(\{x : h(x) \neq f(x)\})$	$L_{\mathcal{D}}(h) = \mathcal{D}(\{(x,y) : h(x) \neq y\})$
Échantillon S :	$(x_1, \ldots, x_m) \sim \mathcal{D}^m$ $\forall i,\ y_i = f(x_i)$	$((x_1,y_1), \ldots, (x_m,y_m)) \sim \mathcal{D}^m$
Objectif :	$L_{\mathcal{D},f}(A(S)) \leq \epsilon$	$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

Au-delà de la classification binaire

Autres problèmes d'apprentissage :

- Catégorisation multi-classe : $\mathcal{Y}$ est un ensemble fini représentant $|\mathcal{Y}|$ differentes classes.
  - e.g., $\mathcal{X}$ est l'espace des documents et $\mathcal{Y} = \{\text{Actualité}, \text{Sports}, \text{Biologie}, \text{Médecine}\}$
- Régression : $\mathcal{Y} = \mathbb{R}$.
  - e.g., on désire prédire le poids d'un bébé à sa naissance en fonction de la mesure (par ultrasons) de la circonférence du crâne, de la circonférence de l'abdomen et de la longueur du fémur.

Fonctions de perte

- Soit $Z = \mathcal{X} \times \mathcal{Y}$
- Ayant une hypothèse $h \in \mathcal{H}$, et un exemple $(\mathbf{x}, y) \in Z$, quelle est la qualité de la prédiction de h sur $(\mathbf{x}, y)$?
- Fonction de perte : $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$
- Exemples :
  - Perte 0-1 : $\ell(h, (x, y)) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{if } h(x) = y \end{cases}$
  - Perte quadratique : $\ell(h, (x, y)) = (h(x) - y)^2$
  - Valeur absolue de la différence : $\ell(h, (x, y)) = |h(x) - y|$
  - Matrice de coûts : $\ell(h, (x, y)) = C_{h(x),y}$ où C est une matrice $|\mathcal{Y}| \times |\mathcal{Y}|$

Le modèle général PAC

Nous désirons "probablement et approximativement" résoudre :

$$\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \quad \text{tel que} \quad L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell(h, z)] \ .$$

- L'apprenant connaît $\mathcal{H}$, Z et ℓ
- L'apprenant ne connaît pas $\mathcal{D}$, mais a accès à un échantillon $S \sim \mathcal{D}^m$
- En utilisant S, l'apprenant A produit une hypothèse $A(S)$
- Lorsque $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, nous désirons, avec probabilité au moins $1 - \delta$ sur les tirages de S selon $\mathcal{D}^m$, que l'on ait

$$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$$

PAC agnostique : définition formelle

Définition (Critère PAC agnostique)

Un classe d'hypothèses $\mathcal{H}$ est apprenable au sens PAC agnostique, relativement à un ensemble $Z = \mathcal{X} \times \mathcal{Y}$ et une une fonction de perte $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, s'il existe une fonction $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ et un algorithme d'apprentissage A satisfaisant la propriété suivante : pour tout $\epsilon, \delta \in (0,1)$, $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ et distribution $\mathcal{D}$ sur Z, nous avons

$$\mathcal{D}^m \left(\left\{ S \in Z^m \ : \ L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \right) \ \geq \ 1 - \delta$$

Outline

La propriété de convergence uniforme

Définition (convergence uniforme)

$\mathcal{H}$ possède la *propriété de convergence uniforme* (relativement à ℓ) s'il existe une fonction $m^{\mathsf{UC}}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ telle que pour tout $\epsilon, \delta \in (0,1)$, pour toute distribution $\mathcal{D}$, et pour tout $m \geq m^{\mathsf{UC}}_{\mathcal{H}}(\epsilon, \delta)$, nous avons

$$\mathcal{D}^m \left(\{ S \in Z^m : \ \forall h \in \mathcal{H}, \ \ |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \} \right) \geq 1 - \delta$$

Donc, avec probabilité $\geq 1 - \delta$, $L_S(h)$ est une bonne estimation de $L_{\mathcal{D}}(h)$ pour tout $h \in \mathcal{H}$ lorsque $\mathcal{H}$ satisfait la propriété de convergence uniforme.

La convergence uniforme suffit pour apprendre

Théorème (La convergence uniforme suffit pour apprendre)

- *Si $\mathcal{H}$ possède la propriété de convergence uniforme (relativement à ℓ) avec la fonction $m_{\mathcal{H}}^{UC}$, alors $\mathcal{H}$ est apprenable au sens PAC agnostique avec une complexité d'échantillon $m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.*
- *Dans ce cas, $\mathrm{ERM}_{\mathcal{H}}$ est un algorithme d'apprentissage pour $\mathcal{H}$ au sens PAC agnostique.*

Preuve: Soit $h^* \in \mathrm{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ et $h_S \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)$. Si $\mathcal{H}$ satisfait la propriété de convergence uniforme, alors lorsque $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta)$, nous avons avec probabilité $\geq 1 - \delta$

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \tfrac{\epsilon}{2} \leq L_S(h^*) + \tfrac{\epsilon}{2} \leq L_{\mathcal{D}}(h^*) + \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = L_{\mathcal{D}}(h^*) + \epsilon\,.$$

Puisque $h_S = \mathrm{ERM}_{\mathcal{H}}(S)$ et $L_{\mathcal{D}}(h^*) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, on a

$$\mathcal{D}^m \left(\left\{ S \in Z^m : \; L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) \; \leq \; \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \right) \geq 1 - \delta$$

Les classes finies sont apprenables au sens PAC agnostique

Nous allons démontrer le théorème suivant :

Théorème

Soit $\mathcal{H}$ une classe finie et soit une fonction de perte à valeur dans $[0, 1]$. Alors, $\mathcal{H}$ est apprenable au sens PAC agnostique en utilisant $\mathrm{ERM}_{\mathcal{H}}$ avec la complexité d'échantillon satisfaisant

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

Preuve: En raison du dernier théorème, il suffit de démontrer que $\mathcal{H}$ possède la propriété de convergence uniforme avec

$$m_{\mathcal{H}}^{\mathsf{UC}}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

Preuve (suite)

- Pour démontrer que $\mathcal{H}$ possède la propriété de convergence uniforme, il suffit de démontrer que :

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta \ ,$$

- ou, de manière équivalente, il suffit de démontrer que :

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta \ .$$

- Par la borne de l'union, nous avons :

$$\begin{aligned}
&\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&\quad = \ \mathcal{D}^m(\cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&\quad \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \\
&\quad \leq \ |\mathcal{H}| \max_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \ .
\end{aligned}$$

Preuve (suite)

- Rappel : $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$ et $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$.
- Soit $\theta_i = \ell(h,z_i)$.
- Alors, pour tout i, $\mathbb{E}[\theta_i] = L_{\mathcal{D}}(h)$

Lemme (Inégalité de Hoeffding)

Soit $\theta_1,\ldots,\theta_m$ une séquence de variables aléatoires i.i.d. telle que pour tout i, $\mathbb{E}[\theta_i] = \mu$ et $\mathbb{P}[a \le \theta_i \le b] = 1$. Alors, pour tout $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \le 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right) .$$

Cela implique que pour h fixe, on a

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \le 2\exp\left(-2\,m\,\epsilon^2\right) .$$

Preuve (suite)

Nous avons démontré que :

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq 2\,|\mathcal{H}|\,\exp\left(-2\,m\,\epsilon^2\right)$$

Alors, si $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$, le terme à droite est $\leq \delta$ tel que désiré. ■

Le truc de discrétisation

- Supposons que $\mathcal{H}$ est paramétrisé par d nombres (e.g., les demi-espaces à $d-1$ dimensions avec seuil).
- Supposons que nous utilisions b bits pour coder chaque nombre (e.g., $b = 32$)
- Alors $|\mathcal{H}| \leq 2^{db}$, et donc

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2db \log 2 + 2 \log(2/\delta)}{\epsilon^2} \right\rceil .$$

- Pas très élégant, mais utile pour borner supérieurement la complexité d'échantillon.

Convergence uniforme des classes de VCdim finies

Les classes $\mathcal{H}$ dont $\mathrm{VCdim}(\mathcal{H}) = d < \infty$ possèdent la propriété de convergence uniforme.

Théorème (Convergence uniforme et VCdim)

Considérez la fonction de perte zéro-un. Soit $\mathcal{H}$ une classe de classificateurs binaires de $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Il existe alors une fonction $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ et des constantes C_1, C_2 satisfaisant

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2},$$

telles que pour tout $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$, on a

$$\mathcal{D}^m\{S : |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon, \forall h \in \mathcal{H}\} \geq 1 - \delta$$

.

Les classes de VCdim finies sont apprenables

En conséquence, le dernier théorème admet le corollaire suivant.

Corollaire (généralisation du théorème fondamental de l'apprentissage)

Considérez la fonction de perte zéro-un. Soit $\mathcal{H}$ une classe de classificateurs binaires de $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. $\mathcal{H}$ est apprenable au sens PAC agnostique. Plus spécifiquement, pour tout $m \geq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$ et pour tout $h_S \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)$, on a

$$\mathcal{D}^m\{S : L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon, \} \geq 1 - \delta \,.$$

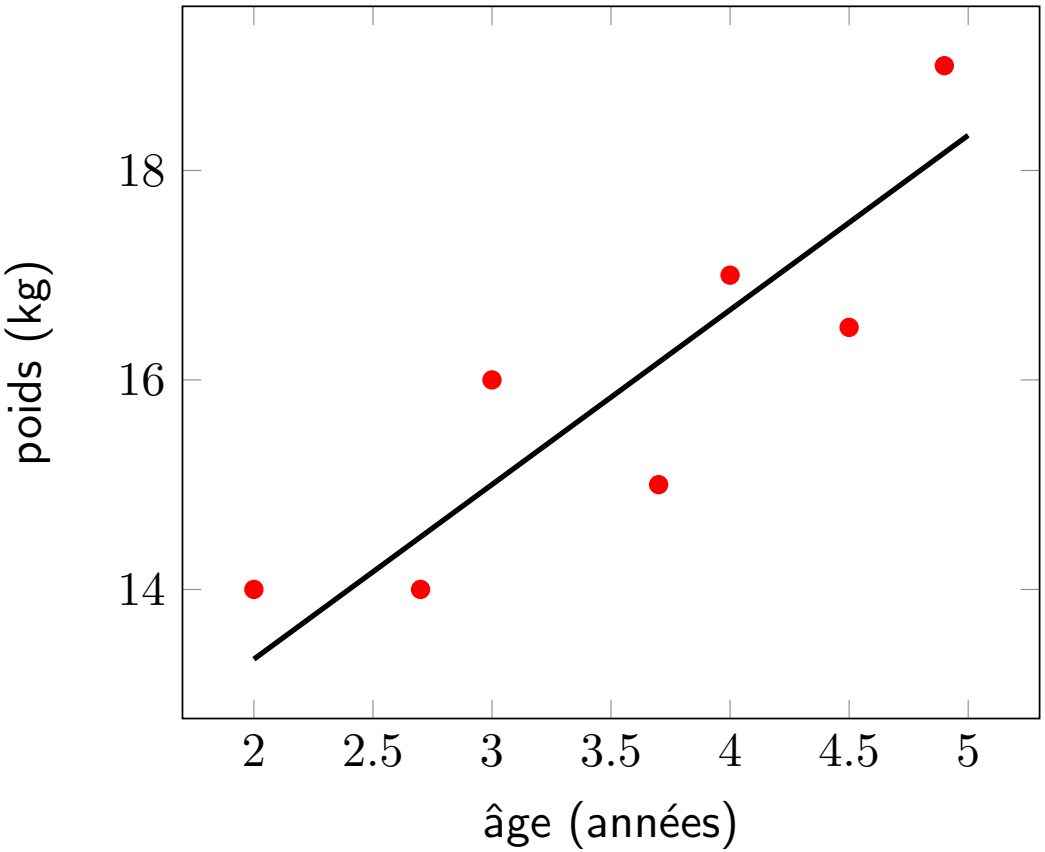
Remarques :

- La VCdim peut être généralisée aux classes $\mathcal{H}$ de fonctions avec fonctions de perte à valeurs réelles (Voir chapitre 5 du livre de Vapnik, 1998) et un théorème similaire s'applique alors à ces cas.
- La convergence uniforme est suffisante mais non nécessaire. Nous verrons plus loin qu'il est possible d'apprendre avec une classe de fonctions ne possédant pas la propriété de convergence uniforme.

Outline

Régression linéaire

- $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}$, $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \mathbf{w} \in \mathbb{R}^d\}$
- Exemple pour $d = 1$: prédire le poids d'un enfant à partir de son âge.



La perte quadratique

- La perte zéro-un n'est pas appropriée pour la régression.
- Utilisons la perte quadratique : $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$.
- Minimisation du risque empirique s'écrit alors :

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

- Soit X la matrice $d \times m$ telle que sa $i$ème colonne est $\mathbf{x}_i$.
- Soit $\mathbf{y}$ le vecteur tel que sa $i$ème composante est y_i.
- Alors la minimisation du risque empirique s'écrit :

$$\min_{\mathbf{w} \in \mathbb{R}^d} \|X^\top \mathbf{w} - \mathbf{y}\|^2$$

Gradient et Optimisation

- La dérivée f' d'une fonction $f : \mathbb{R} \to \mathbb{R}$ évaluée à x est définie par

$$f'(x) \stackrel{\text{def}}{=} \lim_{\Delta \to 0} \frac{f(x+\Delta) - f(x)}{\Delta} .$$

- Si x minimise $f(x)$ alors $f'(x) = 0$.
- Soit $f : \mathbb{R}^d \to \mathbb{R}$.
- Son gradient, $\nabla f(\mathbf{x})$, est un vecteur de dimension d tel que sa $i$ème composante est la dérivée (évaluée en $a = 0$) de la fonction scalaire $g(a) \stackrel{\text{def}}{=} f((x_1, \ldots, x_{i-1}, x_i + a, x_{i+1}, \ldots, x_d))$.
- La dérivée de g s'appelle la dérivée partielle de f, dénotée par $\partial f / \partial x_i$.
- Si $\mathbf{x}$ minimise $f(\mathbf{x})$ alors $\nabla f(\mathbf{x}) = (0, \ldots, 0) \stackrel{\text{def}}{=} \mathbf{0}$.

Le Jacobien et la règle d'enchainement

- Le Jacobien d'une fonction $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ évalué à $\mathbf{x} \in \mathbb{R}^d$, dénoté $J_{\mathbf{x}}(\mathbf{f})$, est une matrice $m \times d$ telle que sa $i$ème ligne est $\nabla f_i(\mathbf{x})$.
- Si $m = 1$ alors $J_{\mathbf{x}}(f) = [\nabla f(\mathbf{x})]^\top$ (un vecteur ligne).
- Si $\mathbf{f}(\mathbf{w}) = A\mathbf{w}$ pour $A \in \mathbb{R}^{m,d}$ alors $J_{\mathbf{w}}(\mathbf{f}) = A$.
- Règle d'enchainement : Soit $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ et $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^d$, le Jacobien de la composition $(\mathbf{f} \circ \mathbf{g}) : \mathbb{R}^k \to \mathbb{R}^m$, évalué à $\mathbf{x}$, est donné par

$$J_{\mathbf{x}}(\mathbf{f} \circ \mathbf{g}) = J_{\mathbf{g}(\mathbf{x})}(\mathbf{f}) J_{\mathbf{x}}(\mathbf{g}) \ .$$

Les moindres carrés

- De retour à notre problème de minimisation du risque empirique :

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|X^\top \mathbf{w} - \mathbf{y}\|^2$$

- Soit $\mathbf{g}(\mathbf{w}) = X^\top \mathbf{w} - \mathbf{y}$ et $\mathbf{f}(\mathbf{v}) = \frac{1}{2}\|\mathbf{v}\|^2 = \frac{1}{2}\sum_{i=1}^m v_i^2$.
- $\operatorname{argmin}_{\mathbf{w}} \mathbf{f}(\mathbf{g}(\mathbf{w}))$ est donné par $\mathbf{w}$ satisfaisant $J_{\mathbf{w}}(\mathbf{f}\circ\mathbf{g}) = \mathbf{0}^\top$.
- Or, $J_{\mathbf{w}}(\mathbf{g}) = X^\top$ et $J_{\mathbf{v}}(\mathbf{f}) = (v_1, \ldots, v_m)$.
- La règle d'enchainement nous dit que

$$J_{\mathbf{w}}(\mathbf{f}\circ\mathbf{g}) = J_{\mathbf{g}(\mathbf{w})}(\mathbf{f})J_{\mathbf{w}}(\mathbf{g}) = \mathbf{g}(\mathbf{w})^\top X^\top = (X^\top \mathbf{w} - \mathbf{y})^\top X^\top .$$

- En imposant que $J_{\mathbf{w}}(\mathbf{f}\circ\mathbf{g}) = (0, \ldots, 0)$, nous obtenons

$$(X^\top \mathbf{w} - \mathbf{y})^\top X^\top = \mathbf{0}^\top \quad \Rightarrow \quad XX^\top \mathbf{w} = X\mathbf{y} .$$

- Si $XX^\top$ est inversible, la solution de ce système d'équations linéaires est donnée par

$$\mathbf{w} = (XX^\top)^{-1} X\mathbf{y} .$$

Les moindres carrés

Petit truc non rigoureux pour se rappeler de cette solution :

- Nous désirons obtenir $X^\top \mathbf{w} \approx \mathbf{y}$
- Multipliez les deux côtés par X pour obtenir $XX^\top \mathbf{w} \approx X\mathbf{y}$
- Multipliez les deux côtés par $(XX^\top)^{-1}$ pour obtenir :

$$\mathbf{w} = (XX^\top)^{-1} X\mathbf{y}$$

- Mais qu'arrive-t-il si $(XX^\top)^{-1}$ n'existe pas ?
- Nous allons voir que, dans ce cas, il existe une infinité de solutions, dont celle obtenue par la **pseudo-inverse** de $XX^\top$.

On a toujours au moins une solution !

- Notez que

$$XX^\top \mathbf{w} = \sum_{k=1}^{m} \mathbf{x}_k \mathbf{x}_k^\top \mathbf{w} = \sum_{k=1}^{m} \mathbf{x}_k \langle \mathbf{x}_k, \mathbf{w} \rangle = \sum_{k=1}^{m} c_k \mathbf{x}_k \ .$$

- Donc $XX^\top$ projette tout $\mathbf{w} \in \mathbb{R}^d$ dans l'espace généré par les vecteurs colonnes de X.
- Si cet espace est de dimension $< d$, alors il existe plusieurs $\mathbf{w} \in \mathbb{R}^d$ projetés sur le même point $XX^\top \mathbf{w}$.
- Dans ce cas $(XX^\top)^{-1}$ n'existe pas.
- Par contre,

$$X\mathbf{y} = \sum_{k=1}^{m} y_k \mathbf{x}_k$$

  est dans l'espace généré par les vecteurs colonnes de X.
- Il existe alors une infinité de solutions de $XX^\top \mathbf{w} = X\mathbf{y}$ obtenues par les $\mathbf{w}$ projetés sur le même point $X\mathbf{y}$. Trouvons l'une de ces solutions !

Décomposition SVD de $XX^\top$

- Puisque $XX^\top$ est une matrice $d \times d$ symétrique semi-définie positive, ses valeurs propres $\lambda_1, \ldots, \lambda_d$ sont toutes non négatives et l'on peut écrire

$$XX^\top = \sum_{k=1}^{d} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top = V D V^\top .$$

- Ceci constitue la décomposition SVD de $XX^\top$. Chaque colonne de V est un vecteur propre $\mathbf{v}_k$ de $XX^\top$, et D est une matrice diagonale constituée des valeurs propres $\lambda_1, \ldots, \lambda_d$.
- Les vecteurs propres ici sont orthonormés, *i.e.*, $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{i,j}$.
- Notez que $XX^\top$ projette tout $\mathbf{w} \in \mathbb{R}^d$ dans un espace de dimension $< d$ (et n'est donc pas inversible) ssi il existe k tel que $\lambda_k = 0$.
- Trouvons alors un $\mathbf{w}$ solutionnant $XX^\top \mathbf{w} = X\mathbf{y}$ qui se trouve dans l'espace générés par les $\mathbf{v}_k$ tel que $\lambda_k > 0$, *i.e.*, écrivons

$$\mathbf{w} = \sum_{k:\lambda_k > 0} \beta_k \mathbf{v}_k .$$

La pseudo-inverse

- La pseudo-inverse de $XX^\top$, noté $\left(XX^\top\right)^+$, est définie par

$$\left(XX^\top\right)^+ \stackrel{\text{def}}{=} \sum_{k:\lambda_k>0} \frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^\top .$$

- On a alors

$$\begin{aligned}\left(XX^\top\right)^+ XX^\top &= \sum_{i:\lambda_i>0}\sum_{j:\lambda_j>0} \frac{\lambda_j}{\lambda_i}\mathbf{v}_i\mathbf{v}_i^\top\mathbf{v}_j\mathbf{v}_j^\top \\ &= \sum_{i:\lambda_i>0}\sum_{j:\lambda_j>0} \frac{\lambda_j}{\lambda_i}\delta_{i,j}\mathbf{v}_i\mathbf{v}_j^\top = \sum_{i:\lambda_i>0}\mathbf{v}_i\mathbf{v}_i^\top .\end{aligned}$$

- Puisqu'il s'agit de l'opérateur identité lorsque toutes les valeurs propres sont non nulles, nous avons $\left(XX^\top\right)^+ = \left(XX^\top\right)^{-1}$ lorsqu'il n'existe pas k t.q. $\lambda_k = 0$, *i.e.*, lorsque $\left(XX^\top\right)^{-1}$ existe.

La solution générale

- Donc, pour tout $\mathbf{w} = \sum_{k:\lambda_k>0} \beta_k \mathbf{v}_k$, on a

$$\left(XX^\top\right)^+ XX^\top \mathbf{w} \;=\; \sum_{i:\lambda_i>0} \mathbf{v}_i \mathbf{v}_i^\top \sum_{k:\lambda_k>0} \beta_k \mathbf{v}_k \;=\; \mathbf{w}\,.$$

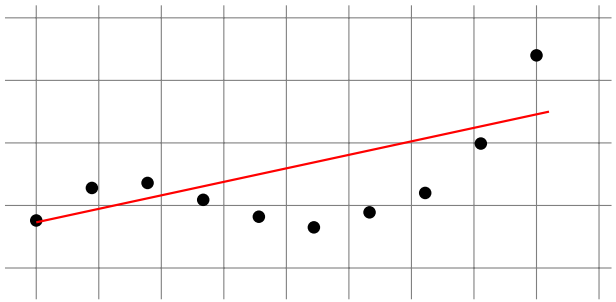
- Lorsque $\mathbf{w}$ est solution de $XX^\top \mathbf{w} = X\mathbf{y}$, on a que $\mathbf{w}$ est donné par

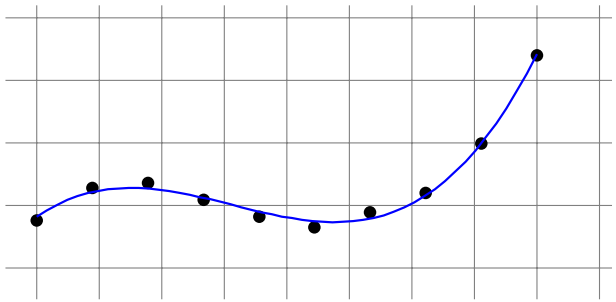
$$\mathbf{w} \;=\; \left(XX^\top\right)^+ X\mathbf{y}\,.$$

- Il s'agit donc de la solution de norme Euclidienne minimale et elle se trouve dans l'espace généré par les vecteurs colonnes de X.

Ajustement polynômial

- Parfois, les prédicteurs linéaires ne sont pas suffisamment expressifs.
- Montrons qu'il est possible d'ajuster un polynôme en utilisant la régression linéaire.





Ajustement polynomial

- Considérons d'abord $\mathcal{X} = \mathbb{R}$ et les fonctions polynomiales de degré n :

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

- Objectif : ayant les données $S = ((x_1, y_1), \ldots, (x_m, y_m))$, trouvez un polynôme de degré n minimisant le risque empirique
- Réduction à la régression linéaire :
- Soit $\boldsymbol{\psi} : \mathbb{R} \to \mathbb{R}^{n+1}$ tel que $\boldsymbol{\psi}(x) \stackrel{\text{def}}{=} (1, x, x^2, \ldots, x^n)$
- Soit $\mathbf{a} = (a_0, a_1, \ldots, a_n)$. Observez que :

$$p(x) = \sum_{i=0}^{n} a_i x^i = \langle \mathbf{a}, \boldsymbol{\psi}(x) \rangle$$

- Pour trouver $\mathbf{a}$, il suffit de résoudre les moindres carrés par rapport à $S = ((\boldsymbol{\psi}(x_1), y_1), \ldots, (\boldsymbol{\psi}(x_m), y_m))$

Ajustement polynomial

- Le même truc s'applique si $\mathcal{X} = \mathbb{R}^d$ au lieu de $\mathbb{R}$. Dans ce cas, considérons
- $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{N(d,n)}$ tel que :

$$\boldsymbol{\psi}(\mathbf{x}) = (1, \{x_i\}_{i=1}^d, \{x_i x_j\}, \{x_i x_j x_k\}, \ldots)$$

  contenant tous les produits d'au plus n composantes de $\mathbf{x} \in \mathbb{R}^d$
- Donc $N(d,n) = 1 + d + d^2 + \ldots + d^n = (d^{n+1} - 1)/(d-1) \in O(d^n)$ pour tout $d > 1$.
- Notez que dans ce cas, X est un matrice $N(d,n) \times m$.
- Donc $XX^\top$ est $N(d,n) \times N(d,n)$ et nécessite alors un temps en $O(d^{3n})$ pour son inversion.
- Nous verrons plus loin qu'il est possible d'utiliser un *noyau* à la place de $\boldsymbol{\psi}$; ce qui nécessitera (uniquement) l'inversion d'une matrice $m \times m$, peu importe la valeur de n.

Outline

Prédire la probabilité d'appartenir à une classe

- On a K classes : $\mathcal{Y} = \{1, \ldots, K\}$; alors $|\mathcal{Y}| = K$.
- Pour tout $x \in \mathcal{X}$, on désire prédire la probabilité que x appartienne à la classe $i \in \mathcal{Y}$.
- L'approche de la régression logistique consiste à construire $\mathbf{h} = (h_1, \ldots, h_K) : \mathcal{X} \to [0, 1]^K$ tel que $h_i(x)$ représente la probabilité que x appartienne à la classe i.
- C'est donc une approche de régression car il faut construire K fonctions à valeur dans $[0, 1]$. Cependant, il faut aussi satisfaire

$$\sum_{i=1}^{K} h_i(x) = 1, \ \forall x \in \mathcal{X} \, .$$

Fonction de perte logarithmique

- Lorsque $\mathcal{X} = \mathbb{R}^d$, l'approche la plus courante consiste à utiliser un vecteur $\mathbf{w}_i \in \mathbb{R}^d$ par fonction h_i et de choisir

$$h_i(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} e^{\langle \mathbf{w}_i, \mathbf{x} \rangle}, \quad Z(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{K} e^{\langle \mathbf{w}_i, \mathbf{x} \rangle},$$

  ce qui nous assure d'avoir $\sum_{i=1}^{K} h_i(\mathbf{x}) = 1, \ \forall \mathbf{x} \in \mathbb{R}^d$.
- La perte $\ell(\mathbf{h}, (\mathbf{x}, y))$ subit par le prédicteur $\mathbf{h}$ sur l'exemple $(\mathbf{x}, y)$ est donnée par la fonction logarithmique

$$\ell(\mathbf{h}, (\mathbf{x}, y)) \stackrel{\text{def}}{=} \log \left(\frac{1}{h_y(\mathbf{x})} \right) = \log Z(\mathbf{x}) - \langle \mathbf{w}_y, \mathbf{x} \rangle .$$

- Ainsi la perte sur $(\mathbf{x}, y)$ de $\mathbf{h}$ sera élevée lorsque $h_y(\mathbf{x}) \ll 1$.

Minimiser le risque empirique

- On peut démontrer que cette fonction de perte est convexe en $(\mathbf{w}_1, \ldots, \mathbf{w}_K)$ et, conséquemment, minimiser le risque empirique

$$\frac{1}{m} \sum_{i=1}^{m} \log \left(\frac{1}{h_{y_i}(\mathbf{x}_i)} \right) ,$$

  s'effectue efficacement à l'aide de la descente de gradient que l'on verra plus loin.

- Notez que pour la classification binaire ($K = 2$), on a

$$\begin{aligned} h_1(\mathbf{x}) &= \frac{e^{\langle \mathbf{w}_1, \mathbf{x} \rangle}}{e^{\langle \mathbf{w}_1, \mathbf{x} \rangle} + e^{\langle \mathbf{w}_2, \mathbf{x} \rangle}} = \frac{1}{1 + e^{-\langle \mathbf{w}_1 - \mathbf{w}_2, \mathbf{x} \rangle}} \\ h_2(\mathbf{x}) &= \frac{e^{\langle \mathbf{w}_2, \mathbf{x} \rangle}}{e^{\langle \mathbf{w}_2, \mathbf{x} \rangle} + e^{\langle \mathbf{w}_1, \mathbf{x} \rangle}} = \frac{1}{1 + e^{-\langle \mathbf{w}_2 - \mathbf{w}_1, \mathbf{x} \rangle}} \end{aligned}$$

Classification binaire et perte logistique

- En utilisant $\mathbf{w} \stackrel{\text{def}}{=} \mathbf{w}_1 - \mathbf{w}_2$, on obtient

$$\begin{aligned} h_1(\mathbf{x}) &= \frac{1}{1+e^{-\langle \mathbf{w}, \mathbf{x} \rangle}} \stackrel{\text{def}}{=} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle) \\ h_2(\mathbf{x}) &= \frac{1}{1+e^{\langle \mathbf{w}, \mathbf{x} \rangle}} = 1 - h_1(\mathbf{x}) \end{aligned}$$

- La fonction $\sigma(a) = 1/(1+\exp(-a))$ est appelée une sigmoïde (en forme de "s").
- On a : $\sigma(-\infty) = 0$, $\sigma(+\infty) = 1$, et $\sigma(0) = 1/2$.
- Maintenant, utilisons $\mathcal{Y} = \{+1, -1\}$ à la place de $\mathcal{Y} = \{1, 2\}$ et $\mathbf{w}$ à la place de $\mathbf{h}$. On obtient alors la perte logistique :

$$\begin{aligned} \ell(\mathbf{w}, (\mathbf{x}, y)) &= \mathbb{1}_{[y=+1]} \log\left(\frac{1}{h_+(\mathbf{x})}\right) + \mathbb{1}_{[y=-1]} \log\left(\frac{1}{h_-(\mathbf{x})}\right) \\ &= \mathbb{1}_{[y=+1]} \log\left(1+e^{-\langle \mathbf{w}, \mathbf{x} \rangle}\right) + \mathbb{1}_{[y=-1]} \log\left(1+e^{\langle \mathbf{w}, \mathbf{x} \rangle}\right) \\ &= \log\left(1+e^{-y\langle \mathbf{w}, \mathbf{x} \rangle}\right) . \end{aligned}$$

Régression logistique

- La régression logistique se résume alors à trouver $\mathbf{w}$ minimisant le risque empirique

$$\frac{1}{m}\sum_{i=1}^{m}\log\left(1+e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right).$$

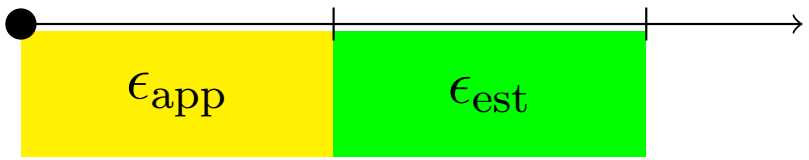
- Puisque la fonction de perte logistique est convexe, trouver $\mathbf{w}$ minimisant le risque empirique se fait efficacement par la descente de gradient (voir plus tard)
- La fonction $\sigma(\langle\mathbf{w},\mathbf{x}\rangle)$ ainsi obtenue s'interprète comme la probabilité, selon $\mathbf{w}$, que l'étiquette de $\mathbf{x}$ soit $+1$.
- La régression logistique et la régression linéaire s'utilisent le plus souvent en ajoutant $\lambda\|\mathbf{w}\|^2$ au risque empirique (avec $\lambda \approx 1/\sqrt{m}$, voir plus loin) et deviennent des algorithmes d'apprentissage très performants lorsqu'utilisés avec une représentation appropriée pour $\mathcal{X}$ où un noyau approprié (voir plus loin).

Outline

Décomposition de l'erreur

- Considérons un algorithme d'apprentissage A pour un $\mathcal{H}$. Soit $h_S \stackrel{\text{def}}{=} A(S)$. Nous pouvons décomposer le risque de h_S comme suit :

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}}$$

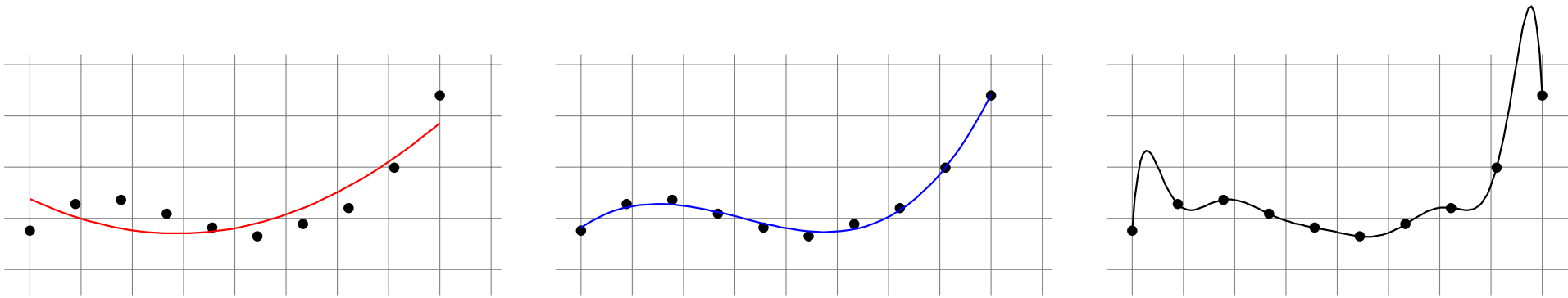


- **L'erreur d'approximation :** $\epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$:
  - La portion du risque due au fait que l'apprenant se restreint à $\mathcal{H}$.
  - Mesure le degré de biais inductif que l'on a en se restreignant à $\mathcal{H}$.
  - Ne dépend pas de S.
  - Diminue lorsque l'on augmente la complexité (la taille, ou le VC) de $\mathcal{H}$.
- **L'erreur d'estimation :** $\epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}$:
  - La portion du risque qui dépends de S.
  - Diminue avec $|S|$ et devrait tendre vers 0 lorsque $|S| \to \infty$.
  - Augmente avec la complexité de $\mathcal{H}$.

Compromis biais-complexité

- Comment choisir $\mathcal{H}$?

degré 2 degré 3 degré 10



- Un $\mathcal{H}$ peu complexe donne une grande erreur d'approximation (grand biais inductif) et une faible erreur d'estimation.
- Un $\mathcal{H}$ complexe donne une faible erreur d'approximation (petit biais inductif) et une grande erreur d'estimation.
- Pour un S donné, il faut donc choisir un $\mathcal{H}$ de complexité intermédiaire nous donnant le meilleur compromis biais-complexité.

Outline

Validation

- Nous avons appris (sur S) une hypothèse h.
- Nous désirons estimer le vrai risque de h.
- Une solution simple : tirer un nouvel échantillon $V = ((x_1, y_1), \ldots, (x_{|V|}, y_{|V|}))$ i.i.d. selon $\mathcal{D}$.
- $L_V(h)$ est alors un estimateur non biaisé de $L_{\mathcal{D}}(h)$ car

$$\mathop{\mathbb{E}}_{V \sim \mathcal{D}^{|V|}} L_V(h) = L_{\mathcal{D}}(h)\,.$$

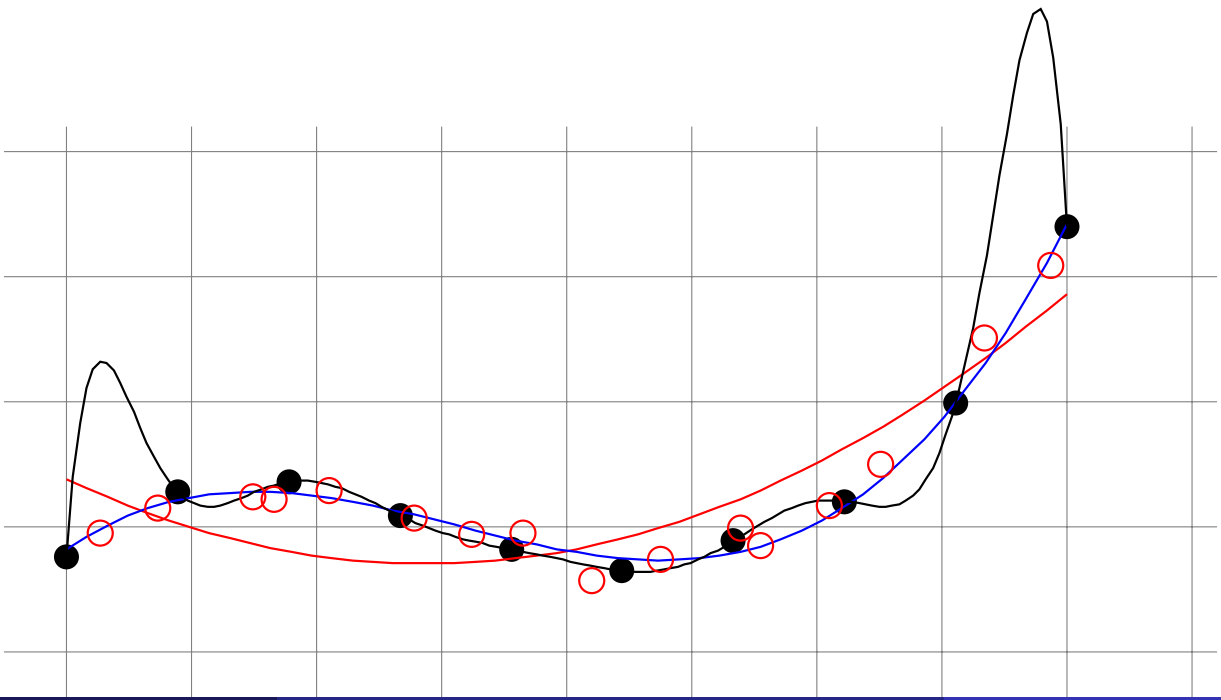
- Pour une fonction de perte ℓ à valeur dans $[0, 1]$, l'inégalité de Hoeffding nous donne

$$|L_V(h) - L_{\mathcal{D}}(h)| \;\leq\; \sqrt{\frac{\log(2/\delta)}{2\,|V|}}$$

avec probabilité $\geq 1 - \delta$ sur les tirages de V.

Validation pour la sélection du modèle

- Ajustement de polynômes de degrés 2,3, et 10 à partir de points.
- Les points noirs constituent l'ensemble d'entrainement S et les points rouges constituent l'ensemble de validation V.
- Le polynôme de degré 10 possède la plus faible erreur S.
- Le polynôme de degré 3 possède la plus faible erreur sur V.
- Possède-t-il le plus faible risque selon $\mathcal{D}$ parmi ces 3 polynômes ?



Validation pour la sélection du modèle — Analyse

- Soit $\mathcal{H} = \{h_1, \ldots, h_r\}$ l'ensemble des prédicteurs obtenus par minimisation du risque empirique sur S en utilisant différentes classes de prédicteurs.
- Soit V un ensemble de validation, différent de S, mais obtenu i.i.d. selon la même distribution $\mathcal{D}$.
- Choisir $h^* \in \mathrm{ERM}_{\mathcal{H}}(V)$.
- Selon l'inégalité de Hoeffding et la borne de l'union sur $\mathcal{H}$ (voir la preuve de l'apprenabilité de $\mathcal{H}$ fini au sens PAC-agnostique), on a

$$\mathcal{D}^{|V|}\left(\left\{V : |L_{\mathcal{D}}(h) - L_V(h)| \leq \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2|V|}}, \forall h \in \mathcal{H}\right\}\right) \geq 1-\delta$$

Validation pour la sélection du modèle — Analyse

Alors, avec probabilité $\geq 1-\delta$ sur les tirages de V, on a, pour tout $h \in \mathcal{H}$,

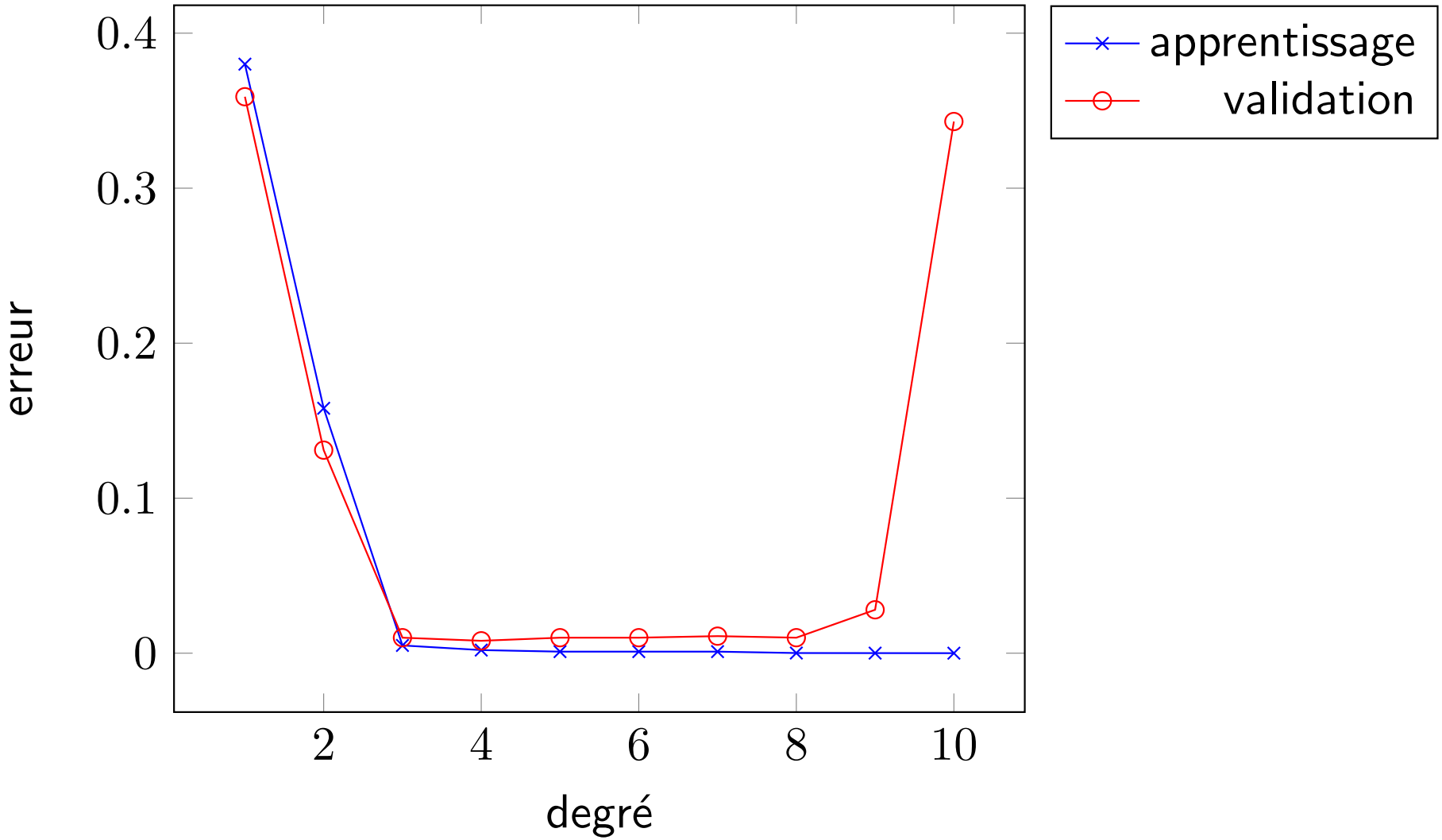
$$\begin{aligned} L_{\mathcal{D}}(h^*) &\leq L_V(h^*) + \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2|V|}} \leq L_V(h) + \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2|V|}} \\ &\leq L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2|V|}}\,. \end{aligned}$$

Donc, avec probabilité $\geq 1-\delta$ on a

$$L_{\mathcal{D}}(h^*) \leq \min_{h\in\mathcal{H}} L_D(h) + \sqrt{\frac{2\log(2|\mathcal{H}|/\delta)}{|V|}}\,.$$

Donc, $L_{\mathcal{D}}(h^*)$ est à ϵ près de $\min_{h\in\mathcal{H}} L_D(h)$ si $|V| \geq 2\log(2|\mathcal{H}|/\delta)/(\epsilon^2)$.

La courbe de sélection de modèle



Partition apprentissage/validation/test

- En pratique, il est usuel de partitionner notre ensemble de données en trois groupes distincts :
  - Ensemble d'apprentissage S : utilisé par l'algorithme d'apprentissage avec différents hyperparamètres (ou classes) afin de produire $\mathcal{H} = \{h_1, \ldots, h_r\}$.
  - Ensemble de validation V : Choisir h^* dans $\mathcal{H}$ en estimant l'erreur sur l'ensemble de validation V.
  - Ensemble test T : Estimer l'erreur de h^* sur l'ensemble test.

Validation croisée pour sélection de modèle

- Lorsqu'il y a peu de données, il est préférable de les partitionner en deux groupes (apprentissage S, test T) et faire de la validation croisée sur S pour choisir h^* (que l'on testera par la suite sur T).

`Validation croisée` k `fois`

entrée :

Échantillon d'apprentissage $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, algorithme d'apprentissage A, et un ensemble Θ de valeurs d'hyperparamètres

partitionner S en $S_1, S_2, \ldots, S_k$ avec $S_i \cap S_j = \emptyset$ et $|S_i| \approx |S_j| \, \forall i, j$

pour chaque $\theta \in \Theta$

pour $i = 1 \ldots k$

$h_{i,\theta} = A(S \setminus S_i; \theta)$

$\text{CV-error}(\theta) = \frac{1}{k} \sum_{i=1}^{k} L_{S_i}(h_{i,\theta})$

sortie :

$\theta^\star = \text{argmin}_\theta \, [\text{CV-error}(\theta)], \quad h_{\theta^\star} = A(S; \theta^\star)$

En résumé

- Le modèle PAC général
  - est agnostique (ne suppose pas de cible f étiquetant les exemples)
  - admet les fonctions de perte générales
- La convergence uniforme suffit pour apprendre une classe $\mathcal{H}$.
- Exemples de classes ayant la propriété de convergence uniforme :
  - Les classes finies avec l'utilisation d'une fonction de perte bornée
  - Les classes de classificateurs binaires avec une VCdim fini
- Utilisation des moindres carrés pour la régression linéaire et l'ajustement de fonctions polynomiales
- Utilisation de la régression logistique pour prédire la probabilité d'appartenance à une classe
- Décomposition de l'erreur et le compromis biais-complexité
- Validation de prédicteurs et la sélection de modèles