Mathématiques du Machine Learning ERM1

Contexte : $(x, y), (x_1, y_1), \dots, (x_n, y_n)$ iid $\in \mathcal{X} \times \mathcal{Y}$.

$$\underbrace{R(f) = \mathbb{E}\left(\ell(y, f(x))\right)}_{\text{Risque théorique}} \text{ "}\approx\text{" } \underbrace{\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, f(x_i)\right)}_{\text{Risque empirique}}$$

Minimisation du risque empirique : $\hat{f} \in \underset{f \in F}{\arg\min}\ \hat{R}(f)$

Question : Que peut-on dire sur $R(\hat{f}) - \left\{R^* = \inf_f R(f)\right\}$?

I. Convexification du risque

Pour un problème de classification binaire $(\mathcal{Y} = \{-1, 1\})$, la fonction de perte usuelle est celle du 0-1 :

$$\ell(y, z) = \mathbb{1}_{y \neq z}.$$

⚠ Le problème devient donc combinatoire.

Solution : Pour éviter ce problème, on commence par rajouter de la régularité en effectuant une convexification du risque

- Étape 1 : Reparamétrisation en un test de signe

On reparamétrise $f : \mathcal{X} \to \{-1, 1\}$ par $g : \mathcal{X} \to \mathbb{R}$:

$$\forall x \in \mathbb{R},\ f(x) = \mathrm{sgn}(g(x)) \text{ où } \mathrm{sgn}(y) = \begin{cases} 1 & \text{si } y > 0 \\ -1 & \text{si } y < 0 \\ \mathrm{Unif}(\{-1, 1\}) & \text{si } y = 0 \end{cases}$$

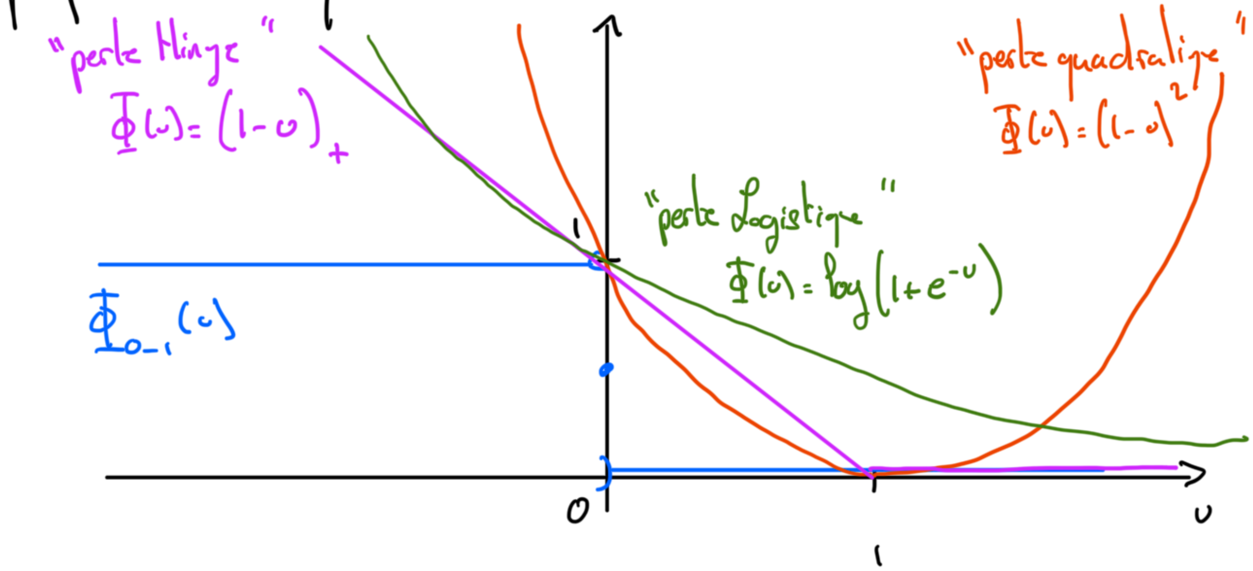
remarque : Pour gérer le cas en 0, on a remplacé une fonction par une fonction aléatoire. En pratique, ce ne sera pas un problème car on ne devrait jamais tomber sur 0.

On a alors $\mathcal{R}(f) = \mathbb{E}\left(\mathbb{1}_{g(x)\neq 0}\, \mathbb{1}_{f(x)\neq y}\right) + \mathbb{E}\left(\mathbb{1}_{g(x)=0}\, \mathbb{1}_{f(x)\neq y}\right)$

$$= \mathbb{E}\left(\mathbb{1}_{g(x)y<0}\right) + \frac{1}{2}\,\mathbb{E}\left(\mathbb{1}_{g(x)=0}\right)$$

$$= \mathbb{E}\left(\Phi_{0-1}\left(y g(x)\right)\right) \text{ où } \Phi_{0-1}(u) = \begin{cases} 1 \text{ si } u<0 \\ \frac{1}{2} \text{ si } u=0 \\ 0 \text{ si } u>0 \end{cases}$$

• Étape 2: On remplace Φ_{0-1} par une fonction avec de meilleures propriétés numériques.



II. Décomposition du risque

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* = \underbrace{\left\{ \mathcal{R}(\hat{f}) - \inf_{f' \in F} \mathcal{R}(f') \right\}}_{\text{Erreur d'estimation}} + \underbrace{\left\{ \inf_{f' \in F} \mathcal{R}(f') - \mathcal{R}^* \right\}}_{\text{Erreur d'approximation}}$$

<u>Question :</u> Comment se comporte l'erreur d'estimation ?

Notons $f_{opt} \in \arg\min_{f' \in F} \mathcal{R}(f)$: meilleur estimateur dans la classe F.

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f_{opt}) = \left\{ \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) \right\} + \left\{ \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{opt}) \right\} + \left\{ \hat{\mathcal{R}}(f_{opt}) - \mathcal{R}(f_{opt}) \right\}$$

$$\leq \boxed{2 \sup_{f \in F} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right|} \ (*) + \text{Erreur d'optimisation}$$

L'erreur d'optimisation dépend de l'algorithme d'entraînement utilisé.
Il peut souvent être aussi petit que l'on veut. Le reste de cette leçon étudie le premier terme.

III. Concentration : Méthode des différences bornées

Une manière de contrôler (*) est d'utiliser la méthode générique des différences bornées :

Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes sur $\mathcal{X}$ et $f : \mathcal{X}^n \to \mathbb{R}$ tq

$\forall i, \forall \underbrace{(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)}_{\mathbb{X}_i}, \underbrace{(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)}_{\mathbb{X}_i'}$ dans le support de la distribution de $(X_1, \dots, X_n)$,

$$\left| f(\mathbb{X}_i) - f(\mathbb{X}_i') \right| \leq c_i .$$

<u>Question :</u> Que peut-on dire de $f(X_1, \dots, X_n)$ vs $\mathbb{E}\left(f(X_1, \dots, X_n)\right)$?

1) Inégalité de Markov.

Soit A une variable <u>positive</u> dans L^1 et $a > 0$.

$$a \mathbb{1}_{A \geqslant a} \leq A \text{ (p.s.)} \Rightarrow \mathbb{E}\left(a \mathbb{1}_{A \geqslant a}\right) \leq \mathbb{E}(A)$$
$$\Rightarrow \mathbb{P}(A \geqslant a) \leq \frac{\mathbb{E}(A)}{a}$$

2) Transformation en martingale

Notons $\forall i$, $V_i = \mathbb{E}\left[f(X_1, \dots, X_n) \mid X_1, \dots, X_i\right) - \underbrace{\mathbb{E}\left[f(X_1, \dots, X_n) \mid X_1, \dots, X_{i-1}\right)}_{= \mathbb{E}\left(f(X_1, \dots, X_n)\right) \text{ si } i=1}$

Notons $M_i = \sum_{j=1}^{i} V_i$.

- M_i est une martingale pour la filtration $\left(\sigma(X_1, \dots, X_{i-1})\right)_i$.

En effet, $\mathbb{E}(V_i | X_1, \dots, X_{i-1}) = \mathbb{E}(\mathbb{E}(f(X_{1:n}) | X_{1:i}) | X_{1:i-1})$

$$- \mathbb{E}(\mathbb{E}(f(X_{1:n}) | X_{1:i-1}) | X_{1:i-1})$$

$$= \mathbb{E}(f(X_{1:n}) | X_{1:i-1}) - \mathbb{E}(f(X_{1:n}) | X_{1:i-1})$$

$$= 0 .$$

- $\underline{f(X_1, \dots, X_n) - \mathbb{E}(X_1, \dots, X_n) = \sum_i V_i}$

3) Fonction génératrice des moments et règle de récurrence

$$\mathbb{P}(f - \mathbb{E}(f) \geq t) \leq \mathbb{P}\left(\exp(s(f - \mathbb{E}(f))) \leq \exp(st)\right)$$

$$\underset{\text{Markov}}{\leq} e^{-st}\, \mathbb{E}\left(\exp\left(s \sum_i V_i\right)\right)$$

$$= e^{-st}\, \mathbb{E}\left(\prod_i \exp(s V_i)\right)$$

$$= e^{-st}\, \mathbb{E}\left(\left(\prod_{i=1}^{n-1} \exp(s V_i)\right) \exp(s V_n)\right)$$

$$= e^{-st}\, \mathbb{E}\left(\left(\prod_{i=1}^{n-1} \exp(s V_i)\right) \underbrace{\mathbb{E}\left(\exp(s V_n) | X_{1:n-1}\right)}_{(*)}\right)$$

On va travailler là dessus $(*)$
et généraliser par récurrence.

4) Borner les incréments

<u>Lemme</u> : $\forall i, \ \exists \alpha_i \leq \beta_i$ tq $V_i \in [\alpha_i, \beta_i]$ p.s. et $|\beta_i - \alpha_i| \leq c_i$.

$X_{1:i-1}$ mesurables

<u>Preuve</u> : $V_i = \mathbb{E}(f | X_{1:i}) - \mathbb{E}(f | X_{1:i-1})$

$$= \int f(X_1,\dots,X_i,x_{i+1},\dots,x_n)\,\mathbb{P}_{|X_{1:i}}(dx_{i+1:n})$$

$$- \int f(X_1,\dots,X_{i-1},x_i,\dots,x_n)\,\mathbb{P}_{|X_{1:i-1}}(dx_{i:n})$$

$$\underset{(\text{Indépendance})}{=} \underbrace{\int \left(f(X_{1:i},x_{i+1:n}) - \int f(X_{1:i-1},x_{i:n})\mathbb{P}(dx_i)\right) \mathbb{P}(dx_{i+1:n})}_{:= \Delta_{|X_{1:i-1}}(X_i)}$$

et notons $\alpha_i := \sup_{x \in \operatorname{supp}(X_i)} \Delta_{|X_{1:i-1}}(x)$

et $\beta_i := \inf_{x \in \operatorname{supp}(X_i)} \Delta_{|X_{1:i-1}}(x)$.

Alors clairement $\alpha_i \leq V_i \leq \beta_i$

de plus,
$$|\beta_i - \alpha_i| = \sup_{x,x' \in \operatorname{supp}(X_i)} \int \overbrace{f(X_{1:i-1},x,x_{i+1:n}) - f(X_{1:i-1},x',x_{i+1:n})}^{\leq c_i} \mathbb{P}(dx_{i+1})\dots\mathbb{P}(dx_n)$$

$$\leq c_i$$

$\square$

Corollaire : $\mathbb{E}\left(\prod_i \exp(s V_i)\right) \leq \exp\left(s^2 \frac{\sum_i c_i^2}{8}\right)$

Preuve : Comme expliqué plus haut, on se focalise sur (*), le reste étant obtenu par récurrence

Montrons que $\mathbb{E}(\exp(sV_n)|X_{1:n-1}) \leq \exp(s^2 c_i^2/8)$.

Notons, $\varphi(s) = \ln\left(\mathbb{E}(\exp(sV_n)|X_{1:n-1})\right)$

Alors, $\varphi'(s) = \dfrac{\mathbb{E}(V_n \exp(s V_n) \mid X_{1:n-1})}{\mathbb{E}(\exp(s V_n) \mid X_{1:n-1})}$

et, $\varphi''(s) = \dfrac{\mathbb{E}(V_n^2 \exp(s V_n) \mid X_{1:n-1})}{\mathbb{E}(e^{s V_n} \mid X_{1:n-1})} - \left(\dfrac{\mathbb{E}(V_n \exp(s V_n) \mid X_{1:n-1})}{\mathbb{E}(\exp(s V_n) \mid X_{1:n-1})} \right)^2$

et on reconnaît l'expression de la variance de V_n sous la loi conditionnelle à $X_{1:n}$ de densité $v \mapsto \exp(sv) / \mathbb{E}(\exp(s V_n) \mid X_{1:n-1})$.

d'où $\varphi''(s) = \underset{\delta}{\arg\min} \int (V_n - \delta)^2 \times \text{densité modifiée}(V_n) \, dV_n$

$$\leq \frac{(\beta_n - \alpha_n)^2}{4} \text{ en prenant } \delta = \frac{\beta_n + \alpha_n}{2}$$

donc, d'après Taylor : $\varphi(s) \leq \dfrac{(\beta_i - \alpha_i)^2}{8} s^2$ $\square$

4) Conclusion

$$\mathbb{P}\left(\sum_i V_i \geq t\right) \leq \inf_{s>0} \exp(-st) \exp\left(s^2 \frac{\sum_i c_i^2}{8}\right)$$

et avec $s = \dfrac{4t}{\sum_i c_i^2}$, on obtient

$$\boxed{\mathbb{P}\big(f(X_{1:n}) - \mathbb{E}(f(X_{1:n}))\big) \leq \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right)}$$

IV. Application : ERM pour un ensemble d'hypothèses fini

Hypothèse 1 : $\ell(y, f(x)) \in [0, \ell_\infty]$ $\forall f \in F$, $\forall (x, y) \in$ Supp.

Hypothèse 2 : $|F| < +\infty$.

Par borne d'union,

$$\mathbb{P}\left(\sup_{f \in F} |\hat{R}(f) - R(f)| \geq t \right) \leq \sum_{f \in F} \mathbb{P}\left(|\hat{R}(f) - R(f)| \geq t \right)$$

$$\leq \sum_{f \in F} 2 \exp\left(- \frac{2nt^2}{\ell_\infty^2} \right)$$

$$\leq \underbrace{2 |F| \exp\left(-2nt^2 / \ell_\infty^2 \right)}_{= \delta}$$

et donc, avec probabilité au moins $1 - \delta$,

$$\sup_{f \in F} |\hat{R}(f) - R(f)| \leq \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log\left(\frac{2|F|}{\delta} \right)}$$