Leçon n°3 : ERM I, Risques convexifiés, Décomposition du risque, Inégalités de concentration, Applications pour ERM

Contexte : $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d $\in \mathcal{X} \times \mathcal{Y}$

$$R(f) = \mathbb{E}\left(\ell(Y, f(X))\right) \qquad (X, Y) \perp \sigma\left((X_1, Y_1), \dots, (X_n, Y_n)\right)$$

de même distribution que (X_1, Y_1).

ERM : $\hat{f} \in \underset{f}{\operatorname{argmin}} \left(\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) \right)$

$\hat{R}(f) \approx R(f)$

Question : Peut-on contrôler $R(\hat{f}) - R^*$?

I. Convexification du risque

Pour des problèmes structurés (par exemple la classification binaire $\mathcal{Y} = \{-1, 1\}$), le problème de minimisation du risque empirique associé à la perte du 0-1 (i.e $\ell(y, f) = \mathbb{1}_{y \neq f}$) est souvent NP difficile.

Pour cette raison (et potentiellement des considérations d'optimisation et de statistique), il peut être judicieux de changer la fonction de perte ℓ de sorte à satisfaire certaines propriétés.

② Propriétés désirables :

- Différentiab. p.p. (SGD, ...)
- Smoothness
- Convexité
- ...

Étape 1 : On reparamétrise $f : \mathcal{X} \to \{-1, 1\}$ par $g : \mathcal{X} \to \mathbb{R}$ de la manière suivante

$$\forall x \in \mathcal{X},\ f(x) = \operatorname{sgn}(g(x))$$

et $$\operatorname{sgn}(y) = \begin{cases} 1 \text{ si } y > 0 \\ -1 \text{ si } y < 0 \\ \underbrace{\mathrm{Unif}(\{-1, 1\})}_{\perp \text{ des autres quantités}} \text{ si } y = 0 \end{cases}$$

remarque : $\operatorname{sgn}(\cdot)$ est en fait un noyau conditionnel de probas.

On voit alors que $\underbrace{R(g) = R(f)}_{\text{abus de notation}}$

$$= \mathbb{E}\left(\mathbb{1}_{g(x) \neq 0} \mathbb{1}_{f(x) \neq y}\right) + \mathbb{E}\left(\mathbb{1}_{g(x) = 0} \mathbb{1}_{f(x) \neq y}\right)$$

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

$$= \mathbb{E}\left[\Phi_{0-1}(y g(x))\right]$$

(3)

où $\Phi_{0-1}(v) = \begin{cases} 1 & \text{si } v < 0 \\ \frac{1}{2} & \text{si } v = 0 \\ 0 & \text{si } v > 0 \end{cases}$

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

$$R_\Phi(g) = \mathbb{E}\left(\Phi(y g(x))\right)$$

exemples:

- Perte quadratique $\Phi(v) = (v-1)^2$

  alors $\Phi(y g(x)) = (y g(x) - 1)^2$

  $= (y g(x) - y^2)^2$

  $= y^2 (g(x) - y)^2$

  $= (g(x) - y)^2$

- Hinge $\Phi(v) = \max(1-v, 0)$ (c.f. SVM)

- Exponential loss: $\Phi(v) = \exp(-v)$ (c.f. Adaboost)

- Logistic loss: $\Phi(v) = \log(1 + e^{-v})$

  ↳ $\Phi(y g(x)) = \log(1 + e^{-y g(x)})$

  $= -\log\left(\frac{1}{1 + e^{-y g(x)}}\right)$

  $= -\log\left(\sigma(y g(x))\right)$

avec $\sigma(v) = \frac{1}{1+e^{-v}}$: fonction sigmoïde.

<u>remarque</u> : $\Phi(y g(x)) = -\log\left(\underbrace{\sigma(y g(x))}_{\text{vraisemblance conditionnelle}}\right)$ peut

être vu comme l'opposé de la vraisemblance conditionnelle dans le modèle probabiliste

$$\mathbb{P}(y=1 \mid x) = \sigma(g(x))$$

$$\mathbb{P}(y=0 \mid x) = 1 - \sigma(g(x))$$

Pourquoi $R_{\mathcal{I}}(g) \neq R_{\mathcal{I}_{0-1}}(g)$?

↳ voir Bach 2024 (section 4.1).

II - Décomposition du risque

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left(\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \right)$$

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

- L'erreur d'approximation :
  - Est déterministe
  - Dépend de la régularité du pb et de la classe F

- L'erreur d'estimation :
  - Est de nature stochastique

Décomposition de l'erreur d'estimation :

Notons $f_{opt} \in \underset{f \in F}{\operatorname{argmin}}\, R(f)$.

Alors $R(\hat{f}) - R(f_{opt})$

$$\leq \underbrace{\left(R(\hat{f}) - \hat{R}(\hat{f})\right)}_{\leq \sup_{f \in F} \left|\hat{R}(f) - R(f)\right|} + \left(\hat{R}(\hat{f}) - \hat{R}(f_{opt})\right) + \underbrace{\left(\hat{R}(f_{opt}) - R(f_{opt})\right)}_{\leq \sup_{f \in F} \left|\hat{R}(f) - R(f)\right|}$$

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

(erreur d'optim ↓)

Remarque : $\hat{R}(f) - R(f) = \underbrace{\hat{R}(f)}_{= \frac{1}{n} \sum \ldots} - \mathbb{E}\left(R(f)\right)$

III. Concentration : Méthode des différences bornées.

- Soient $X_1, \dots, X_n$ des V.A.s indépendantes (dans $\mathcal{X}$)
- Soit f une fonction de $\mathcal{X}^n$ dans $\mathbb{R}$ tq,

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

$$|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i.$$

Que peut-on dire de $f(X_1, \dots, X_n)$ vs $\mathbb{E}(f(X_1, \dots, X_n))$.

1) Inégalité de Markov

Soit A une VA positive dans L^1. Alors $\forall a > 0$,

~~$\mathbb{E}(A) = \mathbb{E}(A \mathbb{1}_{A \geq a}) + \mathbb{E}(A \mathbb{1}_{A < a})$~~

$$\mathbb{E}(a \mathbb{1}_{A \geq a}) \leq \mathbb{E}(A) \quad (\text{car p.s. } a \mathbb{1}_{A \geq a} \leq A).$$

i.e $\mathbb{P}(A \geq a) \leq \dfrac{\mathbb{E}(A)}{a}$.

2) Transformation en somme de martingale

Notons $V_i = \mathbb{E}(f(X_1, \dots, X_n) \mid X_1, \dots, X_i)$

$$- \mathbb{E}(f(X_1, \dots, X_n) \mid X_1, \dots, X_{i-1}).$$

④ Alors :

- Lemme : $\left(\sum_{k=1}^{i} V_k\right)$ est une martingale pour la filtration $\left(\sigma(X_1, \dots, X_{i-1})\right)$.

preuve : $\mathbb{E}(V_{i+1} | X_1, \dots, X_i)$

$$= \mathbb{E}\left(\mathbb{E}(f | X_1, \dots, X_{i+1}) - \mathbb{E}(f | X_1, \dots, X_i) \,\middle|\, X_1, \dots, X_i \right)$$

$$= \mathbb{E}\left(\mathbb{E}(f | X_1, \dots, X_{i+1}) \,\middle|\, X_1, \dots, X_i \right)$$

$$- \mathbb{E}\left(\mathbb{E}(f | X_1, \dots, X_i) \,\middle|\, X_1, \dots, X_i \right)$$

$$= \mathbb{E}(f | X_1, \dots, X_i) - \mathbb{E}(f | X_1, \dots, X_i)$$

$$= 0$$

□

- De plus, $\boxed{f - \mathbb{E}(f) = \sum_i V_i}$

3) Fonction génératrice des moments et règle de récurrence

Soit $t > 0$, Soit $s \geq 0$, $\exp(\cdot)$ croissante

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

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

Markov

(9)

De plus,

$$\mathbb{E}\left(e^{s\sum_i V_i}\right) = \mathbb{E}\left(\prod_i e^{sV_i}\right)$$

$$= \mathbb{E}\left(\mathbb{E}\left(\left(\prod_{i=1}^{n-1} e^{sV_i}\right) e^{sV_n} \,\middle|\, X_1, \dots, X_{n-1}\right)\right)$$

$$= \mathbb{E}\left(\left(\prod_{i=1}^{n-1} e^{sV_i}\right) \mathbb{E}\left(e^{sV_n} \,\middle|\, X_1, \dots, X_n\right)\right)$$

1) Borner les incréments

Lemme : $\forall i, \exists \alpha_i \leq \beta_i$ t.q. $V_i \in [\alpha_i, \beta_i]$ p.s. et

$$\beta_i - \alpha_i \leq c_i .$$

Preuve : $V_i = \mathbb{E}(f \mid X_1, \dots, X_i) - \mathbb{E}(f \mid X_1, \dots, X_{i-1})$

$$= \mathbb{E}\left(\int f(X_1, \dots, X_i, x_{i+1}, \dots, x_n)\, \mathbb{P}^{(X_1,\dots,X_i)}(dx_{i+1}, \dots, dx_n)\right.$$

$$\left. - \int f(X_1, \dots, X_{i-1}, x_i, \dots, x_n)\, \mathbb{P}^{(X_1,\dots,X_{i-1})}(dx_i, \dots, dx_n) \,\middle|\, X_1, \dots, X_{i-1}\right)$$

Indep ↓

$$= \mathbb{E}\left(\int f(X_1, \dots, X_i, x_{i+1}, \dots, x_n)\, \mathbb{P}(dx_{i+1}) \dots \mathbb{P}(dx_n)\right.$$

$$\left. - \int f(X_1, \dots, X_{i-1}, x_i, \dots, x_n)\, \mathbb{P}(dx_i) \dots \mathbb{P}(dx_n) \,\middle|\, X_1, \dots, X_{i-1}\right)$$

(9)

$$= \mathbb{E}\Big(\int \Big(f(X_1, \dots, X_{i-1}, X_i, x_{i+1}, \dots, x_n) - f(X_1, \dots, X_{i-1}, x_i, x_{i+1}, \dots, x_n) \Big) \mathbb{P}(dx_i) \mathbb{P}(dx_{i+1}) \dots \mathbb{P}(dx_n) \;\Big|\; X_1, \dots, X_{i-1} \Big)$$

De plus, f est bornée sur $Supp(X_1) \times \dots \times Supp(X_n)$ (conséquence directe de l'hypothèse sur f.

Alors en notant

$$\beta_i = \mathbb{E}\Big(\sup_{\zeta \in Supp(X_i)} \int f(X_1, \dots, X_{i-1}, \zeta, x_{i+1} \dots x_n) \dots \Big)$$

et $\alpha_i = \text{« } \inf_{\zeta \in Supp(X_i)} \text{ »}$

nous avons $-\infty < \alpha_i \leq \beta_i < +\infty$ et $\leq C_i$.

$$(\beta_i - \alpha_i) = \mathbb{E}\Big(\sup_{\zeta, \zeta' \in Supp(X_i)} \int \Big(f(X_1, \dots, X_{i-1}, \zeta, x_{i+1}, \dots, x_n) - f(X_1, \dots, X_{i-1}, \zeta', x_{i+1}, \dots, x_n) \Big) \mathbb{P}(dx_i) \dots \mathbb{P}(dx_n) \;\Big|\; X_1, \dots, X_{i-1} \Big)$$

$\leq C_i$.

□

(10) Corollaire : $\mathbb{E}\left(\prod_i e^{sV_i}\right) \leq \exp\left(s^2 \sum_{i=1}^{n} c_i^2 / 8\right)$

preuve : Soit i. Notons $\varphi(s) = \ln\left(\mathbb{E}\left(e^{sV_i} \mid V_1, \dots, V_{i-1}\right)\right)$.

Alors $\varphi'(s) = \dfrac{\mathbb{E}\left(V_i e^{sV_i} \mid V_1, \dots, V_{i-1}\right)}{\mathbb{E}\left(e^{sV_i} \mid V_1, \dots, V_{i-1}\right)}$

et $\varphi''(s) = \dfrac{\mathbb{E}\left(V_i^2 e^{sV_i} \mid V_1, \dots, V_{i-1}\right)}{\mathbb{E}\left(e^{sV_i} \mid \dots\right)} - \left(\dfrac{\mathbb{E}\left(V_i e^{sV_i} \mid \dots\right)}{\mathbb{E}\left(e^{sV_i} \mid \dots\right)}\right)^2$

et on reconnaît l'expression de la variance de V_i conditionnelle pour densité conditionnelle à $X_1, \dots, X_{i-1}$: $x \mapsto e^{sx} / \mathbb{E}\left(e^{sV_i} \mid X_1, \dots, X_{i-1}\right)$

d'où : $\varphi''(s) = \operatorname{var}_s \mathbb{E}\left((V_i - \xi)^2 \times \text{densité modifiée} \mid V_1, \dots, V_{i-1}\right)$

$\leq \dfrac{(\beta_i - \alpha_i)^2}{4}$ (en prenant $\xi = \text{milieu du segment} = \dfrac{\beta_i + \alpha_i}{2}$).

*

Le résultat est alors immédiat par la formule de récurrence

* d'où, d'après Taylor Lagrange :

$$\varphi(s) \leq \frac{(\beta_i - \alpha_i)^2 s^2}{8}.$$

4) Conclusion

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

et minimiser en s $\left(s^2 = \dfrac{4t}{\sum_i (\beta_i - \alpha_i)^2}\right)$ donne

$$\boxed{\mathbb{P}\left(\sum_i V_i \geq t\right) = \mathbb{P}\left(\hat{P}(f) - \mathbb{E}(f) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_i (\beta_i - \alpha_i)^2}\right)}$$

IV. Application I : EMM avec $\mathcal{F}$ finie.

Hypothèse 1 : $P(y, f(x)) \in [0, P_\infty] \quad \forall f \in \mathcal{F}, \forall (x,y) \in \mathrm{Supp}$

Hypothèse 2 : $|\mathcal{F}| \leq +\infty$

Alors, par borne d'union,

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

$$\leq \sum_{f \in \mathcal{F}} 2 \exp\left(-2nt^2 / P_\infty^2\right)$$

McDiarmid

$$\leq 2|\mathcal{F}| \exp\left(-2nt^2 / P_\infty^2\right)$$

(12) Donc, en notant $\delta = 2|\mathcal{F}| \exp\left(-2n t^2 / P_\infty^2\right)$, avec prob $1-\delta$,

$$\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \leq t = \frac{P_\infty}{\sqrt{2n}} \sqrt{\log \frac{2|\mathcal{F}|}{\delta}}$$

$$= \frac{P_\infty}{\sqrt{2n}} \sqrt{\log(2|\mathcal{F}|) + \log \frac{1}{\delta}}$$

$$\leq P_\infty \sqrt{\frac{\log(2|\mathcal{F}|)}{2n}} + \frac{P_\infty}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}}$$