

①

Lesson n°3: EMN I, Risques convexifiés, Décomposition du risque;
Inégalités de concentration; Applications pour EMN

Contexte: $(x_1, y_1), \dots, (x_n, y_n)$ i.i.d. $\in X \times Y$

$$m(f) = \mathbb{E}(P(y, f(x))) \quad (x, y) \perp \sigma((x_1, y_1), \dots, (x_n, y_n))$$

de même distribution que (x_i, y_i) .

$$\text{EMN: } \hat{f} \in \arg\min \left(\hat{m}(\hat{f}) = \frac{1}{n} \sum_{i=1}^n P(x_i, y_i) \right)$$

$\approx m(f)$

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

I. Convexification du risque

Pour des problèmes structurés (par exemple la classification binaire $y = \{-1, 1\}$), le problème de minimisation du risque empirique admet la perte du 0-1 (i.e. $P(y, f) = \prod_{y \neq f} 1$)

~~sous-jacente~~ 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 P de sorte à satisfaire certaines propriétés.

②

Propriétés désirables:

- Dérivabilité (SGD, ...)

- Smoothness

- Convexité

- ...

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

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

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

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

On voit alors que $R(g) = R(f)$

abus de notation

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

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

$$= E\left[\sum_{i=1}^n (gg(x_i))\right]$$

③ où $\hat{\Phi}_{\alpha-1}(u) = \begin{cases} 1 & \text{si } u < 0 \\ \frac{1}{2} & \text{si } u = 0 \\ 0 & \text{si } u > 0 \end{cases}$

Étape 2: On remplace $\hat{\Phi}_{\alpha-1}$ par une fonction avec de meilleures propriétés numériques

$$M_{\bar{g}}(y) = E(\bar{\Phi}(yg(z)))$$

exemples:

- Partie quadratique $\bar{\Phi}(u) = (u-1)^2$

- abs. $\bar{\Phi}(yg(z)) = (yg(z)-1)^2$
 $= (yg(z)-y)^2$
 $= y^2(yg(z)-y)$
 $= (y(z)-y)^2$

- Hinge $\bar{\Phi}(u) = \max(1-y, 0)$ (c.f. SVM)

- Exponential Loss: $\bar{\Phi}(u) = \exp(-u)$ (c.f. AdaBoost)

- Logistic Loss: $\bar{\Phi}(u) = \log(1 + e^{-u})$

- $\bar{\Phi}(yg(z)) = \log(1 + e^{-yg(z)})$
 $= -\log\left(\frac{1}{1 + e^{-yg(z)}}\right)$
 $= -\log(\sigma(yg(z)))$

④

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

remarque: $\mathbb{E}(yg(u)) = -\underbrace{\text{Poy}(\sigma(yg(u)))}_{\text{Prise en compte conditionnelle}}$ peut

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

$$\text{P}(y=1|u) = \sigma(g(u))$$

$$\text{P}(y=0|u) = 1 - \sigma(g(u))$$

Pourquoi $M_{\hat{g}}(g) \approx M_{\hat{g}_{0,1}}(g)$?

↳ voir Bach 2024 (section 4.1).

II - Décomposition du risque

$$\hat{f} \in \operatorname{argmin}_{f \in F} \left(\hat{M}(f) = \frac{1}{n} \sum_{i=1}^n R(y_i, f(u_i)) \right)$$

$$\text{Alors: } M(\hat{f}) - M^* = \left\{ M(\hat{f}) - \inf_{f \in F} M(f) \right\} + \left\{ \inf_{f \in F} M(f) - M^* \right\}$$

erreurs d'estimation

erreurs d'approximation

- (5)
 - 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 $\hat{f}_{opt} \in \arg\min_{f \in F} \mathcal{R}(f)$.

$$\begin{aligned}
 & \text{Alors } \mathcal{R}(\hat{f}) - \mathcal{R}(f_{opt}) \\
 & \leq \underbrace{\left(\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) \right)}_{\downarrow} + \underbrace{\left(\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{opt}) \right)}_{\downarrow} + \underbrace{\left(\hat{\mathcal{R}}(f_{opt}) - \mathcal{R}(f_{opt}) \right)}_{\downarrow} \\
 & \leq 2 \sup_{f \in F} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| + \text{Erreur d'optimisation}.
 \end{aligned}$$

$$\begin{aligned}
 \text{Remarque: } \hat{\mathcal{R}}(f) - \mathcal{R}(f) &= \underbrace{\hat{\mathcal{R}}(f) - \mathbb{E}(\mathcal{R}(f))}_{= \frac{1}{n} \sum \dots}
 \end{aligned}$$

⑥

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,

$V(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n), (x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)$
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 $E(f(X_1, \dots, X_n))$.

1) Inégalité de Markov

Soit A une VA dans \mathbb{L}' . Alors $\forall a > 0$,

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

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

$$\text{i.e. } P(A \geq a) \leq \frac{E(A)}{a}.$$

2) Transformation en somme de martingale

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

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

② Plus:

○ Lemme: (V_i) est une martingale pour la filtration $(\sigma(x_1, \dots, x_i))$.

preuve: $\mathbb{E}(V_{i+1} | x_1, \dots, x_i)$

~~Def P(X, Y)~~

$$= \mathbb{E}\left(\mathbb{E}(f | x_1, \dots, x_{i+1}) - \mathbb{E}(f | x_1, \dots, x_i) | x_1, \dots, x_i\right)$$

$$= \mathbb{E}\left(\mathbb{E}(f | x_1, \dots, x_{i+1}) | x_1, \dots, x_i\right)$$

$$- \mathbb{E}\left(\mathbb{E}(f | x_1, \dots, x_i) | x_1, \dots, x_i\right)$$

$$= \mathbb{E}(f | x_1, \dots, x_i) - \mathbb{E}(f | x_1, \dots, x_i)$$

$$= 0$$

□

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

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

$$\leq e^{-st} \mathbb{E}[e^{s \sum_i V_i}]$$

Notation

③

De plus,

$$\begin{aligned}
 \mathbb{E}\left(e^{s\sum_i V_i}\right) &= \mathbb{E}\left(\prod_i e^{sV_i}\right) \\
 &= \mathbb{E}\left(\mathbb{E}\left(\prod_{i=1}^{n-1} e^{sV_i} \middle| X_1, \dots, X_{n-1}\right)\right) \\
 &= \mathbb{E}\left(\prod_{i=1}^{n-1} e^{sV_i}\right) \mathbb{E}\left(e^{sV_n} \middle| X_1, \dots, X_n\right)
 \end{aligned}$$

④ 1) Borne les incréments

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

Preuve: $V_i = \mathbb{E}(f|x_1, \dots, x_i) - \mathbb{E}(f|x_1, \dots, x_{i-1})$

$$\begin{aligned}
 &= \mathbb{E}\left(\int f(x_1, \dots, x_{i-1}, x_i, \dots, x_n) P^{(x_1, \dots, x_{i-1})}(dx_{i+1}, \dots, dx_n)\right. \\
 &\quad \left.- \int f(x_1, \dots, x_{i-1}, x_i, \dots, x_n) P^{(x_1, \dots, x_{i-1})}(dx_i, \dots, dx_n)\right)
 \end{aligned}$$

Indep

$$\begin{aligned}
 &\stackrel{*}{=} \mathbb{E}\left(\int f(x_1, \dots, x_{i-1}, x_i, \dots, x_n) P(dx_{i+1}) \dots P(dx_n)\right. \\
 &\quad \left.- \int f(x_1, \dots, x_{i-1}, x_i, \dots, x_n) P(dx_i) \dots P(dx_n)\right) \\
 &\quad (x_1, \dots, x_{i-1})
 \end{aligned}$$

(9)

$$= \mathbb{E} \left(S \left(P(x_1, \dots, x_{i-1}, x_i, \alpha_{i+1}, \dots, x_n) - P(x_1, \dots, x_{i-1}, x_i, \alpha_{i+1}, \dots, x_n) \right) P(dx_i) P(dx_{i+1}) \dots P(dx_n) \mid x_1, \dots, x_{i-1} \right)$$

De plus, P est bornée sur $\text{Supp}(x_1) \times \dots \times \text{Supp}(x_n)$
 (conséquence directe de l'hypothèse sur P).

Alors on note

$$\beta_i = \mathbb{E} \left(\sup_{\gamma \in \text{Supp}(x_i)} \left\{ P(x_1, \dots, x_{i-1}, \gamma, x_{i+1}, \dots, x_n) \right\} \dots \right)$$

$$\text{et } d_i = \inf_{f \in \text{Supp}(x_i)} \dots$$

Nous avons $-\infty < \alpha_i \leq \beta_i < +\infty$ et c_i

$$(\beta_i - \alpha_i) = \mathbb{E} \left(\sup_{\gamma \in \text{Supp}(x_i)} \left\{ \left(P(x_1, \dots, x_{i-1}, \gamma, x_{i+1}, \dots, x_n) - P(x_1, \dots, x_{i-1}, \gamma, x_{i+1}, \dots, x_n) \right) P(dx_i) \dots P(dx_n) \right\} \mid x_1, \dots, x_{i-1} \right)$$

$$\leq c_i.$$

□

$$\textcircled{10} \quad \text{Condition : } \mathbb{E}\left(\prod_i e^{sV_i}\right) \leq \exp\left(s^2 \sum_{i=1}^n c_i^2 / g\right)$$

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

$$\text{Alors } \ell'(s) = \frac{\mathbb{E}(V_i e^{sV_i} | V_1, \dots, V_{i-1})}{\mathbb{E}(e^{sV_i} | V_1, \dots, V_{i-1})}$$

$$\text{et } \ell''(s) = \frac{\mathbb{E}(V_i^2 e^{sV_i} | V_1, \dots, V_{i-1}) - \left(\frac{\mathbb{E}(V_i e^{sV_i} | \dots)}{\mathbb{E}(e^{sV_i} | \dots)}\right)^2}{\mathbb{E}(e^{sV_i} | \dots)}$$

et on reconnaît l'expression de la variance de V_i conditionnelle pour densité conditionnelle à X_1, \dots, X_{i-1} : $\mathbb{P} \mapsto \frac{s^2}{\mathbb{E}(e^{sV_i} | X_1, \dots, X_{i-1})}$

$$\text{d'où : } \ell''(s) = \arg \min_s \mathbb{E}\left((V_i - s)^2 | \text{densité conditionnelle } (V_1, \dots, V_{i-1})\right)$$

$$\leq \frac{(\beta_i - \alpha_i)^2}{4} \quad (\text{en prenant } s = \text{moy. du segment } = \frac{\beta_i + \alpha_i}{2}).$$

* Le résultat est alors immédiat par la formule de récurrence

* d'ici, d'après Taylor Lagrange :

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

(11)

c) Conclusion

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

et minimiser est $s \left(s = \frac{4t}{\sum_i (\beta_i - \alpha_i)^2} \right)$ donc

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

IV. Application I: EM pour F fini.

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

Hypothèse 2: $|F| \leq +\infty$

Alors, par borne d'union,

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

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

McDiarmid

$$\leq 2|F| \exp\left(-2nt^2/P_\alpha^2\right)$$

(12)

Dans, on note $S = 2|f| \exp\left(-2\pi t^2/\rho_\infty^2\right)$, avec prob $1-S$,

$$\begin{aligned} \sup_{f \in F} |\hat{M}(f) - M(f)| &\leq t = \frac{\rho_\infty}{\sqrt{2\pi}} \sqrt{\log\left(\frac{2|f|}{S}\right)} \\ &= \frac{\rho_\infty}{\sqrt{2\pi}} \sqrt{\log(2|f|) + \log\frac{1}{S}} \\ &\leq \rho_\infty \sqrt{\frac{\log(2|f|)}{2\pi}} + \frac{\rho_\infty}{\sqrt{2\pi}} \sqrt{\log\frac{1}{S}} \end{aligned}$$