Mathématiques du Machine Learning ERM2

<u>Contexte</u> Contrôle de l'erreur d'estimation

$$\sup_{f\in F} |\hat{M}(f) - M(f)| \le \sup_{f\in F} \hat{M}(f) - M(f) + \sup_{f\in F} M(f) - \hat{M}(f).$$

I. Lien entre Forte Proba - Espérance.

<u>Hypothèse</u> : $\forall f \in F$, $\forall (x,y) \in$ Support, $0 \le \ell(y, f(x)) \le \ell_\infty$.

<u>Remarque</u> : $\forall f \in F$, $\hat{M}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$.

en changeant un couple (x_i, y_i) par n'importe quel autre couple dans le support, $\hat{M}(f)$ change au plus de ℓ_∞ / n.

Donc $\sup_{f\in F} \hat{M}(f) - M(f)$ change également au plus de ℓ_∞ / n

et $\sup_{f\in F} M(f) - \hat{M}(f)$ change également au plus de ℓ_∞ / n.

On peut donc appliquer la <u>méthode des différences bornées</u> :

$\forall \delta \in (0,1)$, chacune des inégalités suivantes est vérifiée avec probabilité au moins $1-\delta$:

$$\left| \sup_{f\in F} \left(M(f) - \hat{M}(f)\right) - \mathbb{E}\left(\sup_{f\in F} \left(M(f) - \hat{M}(f)\right)\right) \right| \le \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log\left(\frac{1}{\delta}\right)}$$

$$\left| \sup_{f\in F} \left(\hat{M}(f) - M(f)\right) - \mathbb{E}\left(\sup_{f\in F} \left(\hat{M}(f) - M(f)\right)\right) \right| \le \frac{\ell_\infty}{\sqrt{2n}} \sqrt{\log\left(\frac{1}{\delta}\right)}$$

Donc : pour obtenir des bornes avec forte proba, on peut se contenter de travailler sur borner les espérances

$$\mathbb{E}\left(\sup_{f\in F}\left(M(f)-\hat{M}(f)\right)\right) \text{ et } \underbrace{\mathbb{E}\left(\sup_{f\in F}\left(\hat{M}(f)-M(f)\right)\right)}_{*}.$$

II. Complexité de Rademacher

Notons $\forall i,\ z_i=(x_i,y_i)$ et $z=(x,y)$.

Symétrisation : Soient $z_1',\dots,z_n'$ iid, de même loi que z et indépendants de $z_1,\dots,z_n,z$ (des copies indépendantes des données).

Nous avons, $\forall f$,

$$\hat{M}(f)-M(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i,f(x_i)) - \mathbb{E}_z\left(\ell(y,f(x))\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\ell(y_i,f(x_i)) - \mathbb{E}_{z_1',\dots,z_n'}\left(\frac{1}{n}\sum_{i=1}^{n}\ell(y_i',f(x_i'))\right)$$

$$= \underbrace{\mathbb{E}_{z_1',\dots,z_n'}\left(\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,f(x_i)) - \frac{1}{n}\sum_{i=1}^{n}\ell(y_i',f(x_i'))\right)}$$

donc, $\sup_{f\in F}\left(\hat{M}(f)-M(f)\right) = \sup_{f\in F}$ //

$$\leq \mathbb{E}_{z_1',\dots,z_n'}\left(\sup_{f\in F}\left(\frac{1}{n}\sum_{i=1}^{n}\ell(y_i,f(x_i)) - \frac{1}{n}\sum_{i=1}^{n}\ell(y_i',f(x_i'))\right)\right)$$

et enfin, $(*) \leq \mathbb{E}_{\substack{z_1,\dots,z_n\\ z_1',\dots,z_n'}}\left(\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\ell(y_i,f(x_i)) - \sum_{i=1}^{n}\ell(y_i',f(x_i'))\right)}_{\text{Loi symétrique}}\right)$

Nous avons quasiment terminé ! Soient $\varepsilon_1, \ldots, \varepsilon_n$ des variables aléatoires iid de loi $\mathrm{Rad}(1/2)$ $\left(\begin{array}{l} = 1 \text{ avec proba } \frac{1}{2} \\ = -1 \text{ avec proba } \frac{1}{2} \end{array}\right)$ qui sont indépendantes des autres quantités du cours.

Remarquons que $\forall i$, $\underbrace{\ell(y_i, f(x_i)) - \ell(y_i', f(x_i'))}_{\alpha_i}$ a une loi symétrique.

Alors, $\forall i$, α_i a la même loi que $\varepsilon_i \alpha_i$.

Alors $(*) \leq \mathbb{E}_{\substack{z_1, \ldots, z_n \\ z_1', \ldots, z_n' \\ \varepsilon_1, \ldots, \varepsilon_n}} \left(\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(\ell(y_i, f(x_i)) - \ell(y_i', f(x_i')) \right) \right)$

$$\leq \mathbb{E}_{\substack{z_1, \ldots, z_n \\ \varepsilon_1, \ldots, \varepsilon_n}} \left(\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \ell(y_i, f(x_i)) \right) + \mathbb{E}_{\substack{z_1', \ldots, z_n' \\ \varepsilon_1, \ldots, \varepsilon_n}} \left(\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \ell(y_i', f(x_i')) \right)$$

$$= 2\, \mathbb{E}_{\substack{z_1, \ldots, z_n \\ \varepsilon_1, \ldots, \varepsilon_n}} \left(\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \ell(y_i, f(x_i)) \right).$$

De manière générale, si H est une classe de fonctions h et étant donné une observation $S = (z_1, \ldots, z_n)$, on définit

$$R_S(H) = \mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n} \left(\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i) \right).$$

On vient de prouver le résultat suivant :

Théorème (Symétrisation) :

Si H est une famille de fonctions, $z_1, \dots, z_n$ iid,

$$\underset{(z_1,\dots,z_n)}{\mathbb{E}} \sup_{h\in H} \left(\frac{1}{n}\sum_{i=1}^{n} h(z_i) - \underset{z}{\mathbb{E}}(h(z)) \right) \leq 2 \underbrace{\underset{(z_1,\dots,z_n)}{\mathbb{E}} \left(R_{(z_1,\dots,z_n)}(H) \right)}_{:= R_n(H)}$$

et

$$\underset{(z_1,\dots,z_n)}{\mathbb{E}} \sup_{h\in H} \left(\underset{z}{\mathbb{E}}(h(z)) - \frac{1}{n}\sum_{i=1}^{n} h(z_i) \right) \leq 2 R_n(H).$$

$R_S(H)$ = Complexité de Rademacher de H conditionnellement à S

$R_n(H)$ = Complexité de Rademacher de H sous la loi des données.

Proposition : Si g est G-Lipschitz, $R(g(H)) \leq G\, R(H)$

avec $g(H) = \{ g\circ h \,;\, h \in H \}$.

Preuve : Admis, voir Bach 2024.

III. Analyse des modèles linéaires

Modèles linéaires contraints : $F = \{ f_\theta : x \mapsto \theta^T \varphi(x) \,,\, \Omega(\theta) \leq D \}$

Ω : norme sur $\mathbb{R}^d$.

$$R_n(F) = \underset{\substack{x_1,\dots,x_n \\ \varepsilon_1,\dots,\varepsilon_n}}{\mathbb{E}} \left(\sup_{\substack{\theta \\ \Omega(\theta)\leq D}} \left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \theta^T \varphi(x_i) \right) \right)$$

$$= \mathbb{E}\left(\sup_{\Omega(\theta) \leq D} \frac{1}{n} \varepsilon^T \Phi \theta \right) \text{ avec } \varepsilon = \begin{pmatrix} \vdots \\ \varepsilon_n \end{pmatrix}, \quad \Phi = \begin{pmatrix} \varphi(x_1)^T \\ \vdots \\ \varphi(x_n)^T \end{pmatrix}$$

$$= \frac{D}{n} \mathbb{E}\left(\Omega^*\left(\Phi^T \varepsilon \right)\right)$$ où Ω^* est la norme duale de Ω.

<u>Cas $\Omega = \|\cdot\|_2$:</u>

$$M_n(f) = \frac{D}{n} \mathbb{E}\left(\|\Phi^T \varepsilon\|_2 \right)$$

$$\underset{\text{Jensen}}{\leq} \frac{D}{n} \sqrt{\mathbb{E}\left(\|\Phi^T \varepsilon\|_2^2 \right)}$$

$$= \frac{D}{n} \sqrt{\mathbb{E}\left(\varepsilon^T \Phi^T \Phi \varepsilon \right)}$$

$$= \frac{D}{n} \sqrt{\mathbb{E}\left(\mathrm{tr}\left(\varepsilon^T \Phi^T \Phi \varepsilon \right) \right)}$$

$$= \frac{D}{n} \sqrt{\mathbb{E}\left(\mathrm{tr}\left(\Phi^T \Phi \varepsilon \varepsilon^T \right) \right)}$$

$$= \frac{D}{n} \sqrt{\mathbb{E}\left(\mathrm{tr}\left(\Phi^T \Phi \underbrace{\mathbb{E}\left(\varepsilon \varepsilon^T \mid \Phi \right)}_{= I} \right) \right)}$$

$$= \frac{D}{n} \sqrt{\mathbb{E}\left(\mathrm{tr}\left(\Phi^T \Phi \right) \right)}$$

$$= \frac{D}{n} \sqrt{\mathbb{E}\left(\sum_i \|\varphi(x_i)\|^2 \right)}$$

$$= \frac{D}{\sqrt{n}} \sqrt{\mathbb{E}\|\varphi(x)\|^2}$$

Conclusion : Si $\forall y$ dans le support, $\ell(y, \cdot)$ est G Lipschitz,

$$\mathbb{E}\left(\mathcal{R}\left(f_{\hat{\theta}}\right)\right) \leq \inf_{\|\theta\|_2 \leq D} \mathcal{R}\left(f_{\theta}\right) + \frac{4GD}{\sqrt{n}} \sqrt{\mathbb{E}\|\varphi(x)\|^2} .$$