Apprentissage statistique

Chapitre 8 : Convexification du problème binaire

Lucie Le Briquer

13 mars 2018

ERM :

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(x_i) \neq y_i} \right] = \hat{R}_n(f)$$

- $\mathcal{F}$ peut être finie de très grand cardinalité
- n finie peut être grand

$\hat{\mathcal{R}}_n(f)$ est à valeurs discrètes.
Pour pouvoir résoudre un problème, on cherche à le convexifier. On peut convexifier la perte et/ou convexifier l'ensemble $\mathcal{F}$.

1ère étape (convexification de $\mathcal{F}$)
Si $f \in \mathcal{F} \colon \mathcal{X} \longrightarrow \{0,1\}$, on les symétrise en $\tilde{f} = 2f - 1$ à valeurs dans $\{-1,1\}$. Donc on peut supposer $f \in \mathcal{F} \colon \mathcal{X} \longrightarrow \{-1,1\}$. Si f est à valeurs dans $\{-1,1\}$,

$$\mathbb{1}_{f(X) \neq Y} \;\Leftrightarrow\; \mathbb{1}_{f(X)Y < 0}$$

On va considérer non plus des classifieurs $\mathcal{X} \longrightarrow \{-1,1\}$, mais des classifieurs dit *soft* à valeurs dans $\mathbb{R}$ (ou $[-1,1]$). À chaque classifieur soft $f \colon \mathcal{X} \longrightarrow \mathbb{R}$, on associe un classifeur *binaire* $\operatorname{sgn}(f) \in \{-1,1\}$.

Exemples.

- linéaire : $f_a(x) = a^T x$ pour $a \in \mathcal{A} \subset \mathbb{R}^d$
- majoritaire :

$$f_\lambda(x) = \sum_{i=1}^{M} \lambda_i f_i(x) \qquad \text{pour } (f_1, \ldots, f_M) \text{ fixé et } \lambda \in \mathcal{S}^{n-1}$$

- base :

$$f_\theta(x) = \left\{ \sum_{i \in \mathbb{N}} \theta_i f_i(x),\ \theta_i \in \mathbb{R}^{\mathbb{N}} \text{ et } (f_1, \ldots, f_i, \ldots) \text{ fixés} \right\}$$

Dorénavant on suppose que l'on a une famille convexe de classifieurs soft.

2ème étape (convexifier la perte)
Le problème :

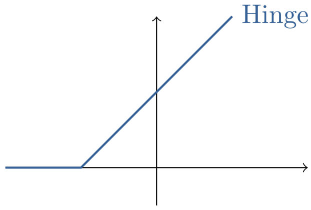
$$\min_{f \in \mathcal{F}_{\text{soft}}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i)Y_i<0}$$

n'est pas convexe. On va donc rempacler $\mathbb{1}_{-Z \geqslant 0}$ par une fonction convexe.

Définition 1 (fonction *surrogate*)
φ est une fonction *surrogate* si :
- φ est convexe
- φ est croissante
- $\varphi(0) = 1$, $\varphi \geqslant 0$

Exemples.

- Hinge : $\max(0, x+1)$



- exp : $\exp(\eta x)$
- log-it : $\log_2(1+e^{\eta x})$

Au lieu de minimiser le risque empirique, on considère le $\varphi-$risque empirique :

$$\hat{\mathcal{R}}_{\varphi}(f) = \frac{1}{n} \sum_{i=1}^{n} \varphi(-f(X_i)Y_i)$$

et le $\varphi-$risque $\mathcal{R}_{\varphi}(f) = \mathbb{E}[\varphi(-f(X)Y)]$. On note :

$$f_{\varphi}^{*} = \underset{f\colon \mathcal{X} \longrightarrow \mathbb{R}}{\operatorname{argmin}} \mathcal{R}_{\varphi}(f) \quad \text{et} \quad \hat{f}_{\varphi} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{\mathcal{R}}_{\varphi}(f)$$

Théorème 1
Si φ est différentiable, alors $\operatorname{sgn}(f_{\varphi}^{*}) = f^{*}$.

Preuve.

$$f^{*}(x) = 1 \;\Leftrightarrow\; \eta(x) \geqslant \frac{1}{2}$$

f_{φ}^{*} est la fonction qui minimise $\mathbb{E}[\varphi(-f(X)Y)]$. On raisonne poit par point et :

$$f_{\varphi}^{*}(x) = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}}\ \eta(x)\varphi(-\alpha) + (1-\eta(x))\varphi(\alpha)$$

Notons $H_\eta(\alpha) = \eta\varphi(-\alpha) + (1-\eta)\varphi(\alpha)$ alors argmin $H_\eta(\alpha)$ est obtenu par :

$$-\eta\varphi'(-\alpha) + (1-\eta)\varphi'(\alpha) = 0 \quad \Leftrightarrow \quad \frac{\eta}{1-\eta} = \frac{\varphi'(\alpha)}{\varphi'(-\alpha)}$$

D'où,

$$\eta \geqslant \frac{1}{2} \;\Leftrightarrow\; \frac{\eta}{1-\eta} \geqslant 1 \;\Leftrightarrow\; \varphi'(\alpha) \geqslant \varphi'(-\alpha) \underset{\text{convexité}}{\Leftrightarrow} \alpha \geqslant -\alpha \;\Leftrightarrow\; \alpha \geqslant 0$$

Don' $\eta(x) \geqslant \frac{1}{2} \;\Leftrightarrow\; f^*_\varphi(x) \geqslant 0$ ou encore, $f^*(x) = 1 \;\Leftrightarrow\; \mathrm{sgn}(f^*_\varphi(x)) = 1$. □

Lemme 1 (de Zhang)

Pour tout $\eta > 0$, on note $\tau(\eta) = \min_{\alpha\in\mathbb{R}} H_\eta(\alpha)$, supposons qu'il existe $c > 0$ et $\gamma < 1$ tels que :

$$\left|\eta - \frac{1}{2}\right| \leqslant c(1-\tau(\eta))^\gamma$$

pour tout $\eta \in [0,1]$, alors,

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f^*) \leqslant 2c(\mathcal{R}_\varphi(f) - \mathcal{R}^*_\varphi)^\gamma$$

Preuve.

$$\begin{aligned}\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f^*) &= \mathbb{E}\left[2\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{\mathrm{sgn}(f)f^*<0}\right] \\ &\underset{\text{hyp}}{\leqslant} 2c\mathbb{E}\left[(1-\tau(\eta(x)))^\gamma \mathbb{1}_{\mathrm{sgn}(f)f^*<0}\right] \\ &\underset{\text{Jensen}}{\leqslant} 2c\mathbb{E}\left[1-\tau(\eta(x))\mathbb{1}_{\mathrm{sgn}(f)f^*<0}\right]^\gamma\end{aligned}$$

Le résultat vient du fait que :

$$\mathcal{R}_\varphi(f) - \mathcal{R}^*_\varphi \geqslant \mathbb{E}\left[(1-\tau(\eta(x)))\mathbb{1}_{\mathrm{sgn}(f)f^*<0}\right]$$

En effet f^* a le même signe que f^*_φ et $\mathrm{sgn}(f)$ a le même signe que f. Donc :

$$\mathbb{E}\left[(1-\tau(\eta(x)))\mathbb{1}_{\mathrm{sgn}(f)f^*<0}\right] = \mathbb{E}\left[(1-\tau(\eta(x)))\mathbb{1}_{f(X)f^*_\varphi(X)<0}\right]$$

$\mathcal{R}_\varphi(f) = \mathbb{E}[\varphi(-f(X)Y)] = \mathbb{E}[H_{\eta(X)}(f(X))]$ et $\mathcal{R}^*_\varphi = \mathbb{E}[\min_\alpha H_{\eta(X)}(\alpha)]$. Donc il suffit de montrer que $\forall x \in X$,

$$H_{\eta(x)}(f(x)) \geqslant (1-\tau(\eta(x)))\mathbb{1}_{f(x)f^*_\varphi(x)<0} + \min_\alpha H_{\eta(x)}(\alpha)$$

Si $\mathbb{1}_{f(x)f^*_\varphi(x)<0} = 0$ alors c'est bon. Sinon, supposons que $f(x)f^*_\varphi(x) < 0$. On veut donc montrer :

$$H_{\eta(x)}(f(x)) \geqslant (1-\tau(\eta(x))) + \tau(\eta(x)) = 1$$

Mais

$$\begin{aligned}H_{\eta(x)}(f(x)) &= \eta(x)\varphi(-f(x)) + (1-\eta(x))\varphi(f(x)) \\ &\geqslant \varphi\big(-\eta(x)f(x) + (1-\eta(x))f(x)\big) \\ &= \varphi\big((1-2\eta(x))f(x)\big)\end{aligned}$$

Comme f_φ^* a le même signe que f^* et f^* a le même signe que $2\eta(x) - 1$. Comme $f(x)f_\varphi^*(x) < 0$, $\text{sgn}(f_\varphi^*(x)) = -\text{sgn}(f^*) = -\text{sgn}(2\eta(x) - 1) = \text{sgn}(1 - 2\eta(x))$. Alors $(1 - 2\eta(x))f(x) \geqslant 0$ et donc comme φ est croissante et que $\varphi(0) = 1$, on a bien $H_{\eta(x)}(f(x)) \geqslant 1$. $\square$

Remarque.

- Si φ est Hinge, $c = \frac{1}{2}$ et $\gamma = 1$
- Si φ est exp, $c = \frac{1}{\sqrt{2}}$ et $\gamma = \frac{1}{2}$
- Si φ est log-it, $c = \frac{1}{\sqrt{2}}$ et $\gamma = \frac{1}{2}$