Optimisation stochastique

Contents

$$E(h) = \int l(h(x), y) dP(x, y)$$

$$h = \underset{h:\mathcal{X}\to\mathcal{Y}}{\text{argmin}}\ E(h)$$

$$E_N(h) = \frac{1}{N}\sum_{n=1}^{N} l(h(x_n), y_n)$$

$$\mathcal{H} : \text{famille de prédicteurs (Dimension D)}$$

$$h^*_{\mathcal{H}} = \underset{h\in\mathcal{H}}{\text{argmin}}\ E(h)$$

$$h^*_N = \underset{h\in\mathcal{H}}{\text{argmin}}\ E_N(h)$$

1 Problème 1

On veut résoudre :

$$\underset{A\in\mathbb{R}^{MxN}}{inf} F(A) \tag{1}$$

avec :

$$F(A) = \frac{1}{2}\sum_{n=1}^{N} ||Ax_n - y_n||^2 = \sum_{n=1}^{N} F_n(A) \qquad F_n(A) = \frac{1}{2}||Ax_n - y_n||^2$$

$$\begin{aligned}
F_n(A+H) &= \frac{1}{2}||(A+H)x_n - y_n||^2 \\
&= \frac{1}{2}\langle (A+H)x_n - y_n, (A+H)x_n - y_n\rangle \\
&= \frac{1}{2}\langle Ax_n - y_n, Ax_n - y_n\rangle + \langle Hx_n, Ax_n - y_n\rangle + \frac{1}{2}\langle Hx_n, Hx_n\rangle \\
&= F_n(A) + \langle Hx_n, Ax_n - y_n\rangle + \frac{1}{2}\langle Hx_n, Hx_n\rangle \\
&= F_n(A) + \langle Hx_n, Ax_n - y_n\rangle + \frac{1}{2}||Hx_n||^2
\end{aligned}$$

Donc :

$$\begin{aligned}
D_{F_n}(A)(H) &= \langle Hx_n, Ax_n - y_n\rangle_{\mathbb{R}^{MxN}} \\
&= \langle (Hx_n)^T, (Ax_n - y_n)^T\rangle_{\mathbb{R}^{NxM}} \\
&= \langle x_n^T H^T, (Ax_n - y_n)^T\rangle_{\mathbb{R}^{NxM}} \\
&= \langle H^T, x_n(Ax_n - y_n)^T\rangle_{\mathbb{R}^{NxM}} \\
&= \langle H, (Ax_n - y_n)x_n^T\rangle_{\mathbb{R}^{MxN}}
\end{aligned}$$

Donc :

$$\nabla F_n(A) = (Ax_n - y_n)x_n^T$$

2 Chapitre 2

2.1 ..

2.2 ..

2.2.1 ..

2.2.2 ..

$$h^* = \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ E(h) \qquad \text{idéal} \tag{2}$$

$$h^*_{\mathcal{H}_D} = \underset{h\in\mathcal{H}_D}{\operatorname{argmin}}\ E(h) \qquad \text{famille}$$

$$h^*_n = \underset{h\in\mathcal{H}_D}{\operatorname{argmin}}\ E_n(h) \qquad \text{Risque empirique}$$

<u>Optimisation :</u> $\tilde{h_n} \approx h^*_n$
On note ε les erreurs d'apprentissage :

$$\begin{aligned} \varepsilon &= E(\tilde{h_n}) - E(h^*_n) \geq 0 \\ &= E(\tilde{h_n}) - E(h^*_n) + E(h^*_n) - E(h^*_{\mathcal{H}_D}) + E(h^*_{\mathcal{H}_D}) - E(h^*) \end{aligned}$$

On pose :

$$\begin{aligned} \varepsilon opt &= E(\tilde{h_n}) - E(h^*_n) \\ \varepsilon_{est} &= E(h^*_n) - E(h^*_{\mathcal{H}_D}) \\ \varepsilon_{app} &= E(h^*_{\mathcal{H}_D}) - E(h^*) \end{aligned}$$

On a : $\varepsilon_{opt} \searrow$ avec N et $\varepsilon_{app} \searrow$ avec D

Erreurs d'approximation

Les erreurs d'approximation caractérisent la capacitéde la famille $\mathcal{H}_D$ à approcher le prédicteur idéal h^*.

Le choix de la famille $\mathcal{H}_D$ demande souvent un énorme travail d'analyse et de modélisation !

Largeur de Kolmogorov

Une méthode populaire pour construire des prédicteurs consiste à considérer un sous-espace vectoriel $\mathcal{H}_D$.

$$h \in \mathcal{H}_D \Leftrightarrow h = \sum_{d=1}^{D} \omega_d \psi_d$$

avec :

$$\psi_d : \mathcal{X} \to \mathcal{Y} \qquad \text{applications pré-définies}$$

Exemples :

- polynomes
- polynomes trigonométriques
- ondelettes
- splines

Supposons qu'on sache par avance que $h^* \in \mathcal{F}$ où $\mathcal{F}$ est une famille de fonctions connues.

La largeur de Kolmogorov de $\mathcal{F}$ permet de quantifier à quel point la famille $\mathcal{F}$ peut être approchée par un sous-espace vectoriel de dimension D.

Définition : La largeur de Kolmogorov de $\mathcal{F}$ est définie par :

$$\delta_d(\mathcal{F}, ||.||) = \inf_{\dim(\mathcal{H}_D)=d} \sup_{f \in \mathcal{F}} \inf_{h \in \mathcal{H}_D} ||h - f||$$

Proposition :

- $\forall \mathcal{F}, \delta_d(\mathcal{F}, ||.||) \searrow$ avec d
- $\forall \alpha, \delta_d(\alpha\mathcal{F}, ||.||) = |\alpha|\delta_d(\mathcal{F}, ||.||)$
- $\delta_d(\mathcal{F}, ||.||) = \delta_d(\text{conv}(F), ||.||)$ où $\text{conv}(F)$ est l'enveloppe convexe de $\mathcal{F}$
- $\mathcal{F}$ compact $\Rightarrow \delta_d(\mathcal{F}, ||.||) \underset{d\to+\infty}{\rightarrow} 0$ et $\mathcal{F}$ bornée.

Définition : (Espace de Soboléo)
On note :

$$\mathcal{W}_p^r([0,1]) = \{h \in \mathcal{C}^{r-1}([0,1]) \text{ tel que } h^{(r-1)}\text{absolument continue et } h^{(r)} \in L^p([0,1])\}$$

- $\mathcal{W}_2^0 = L^2$
- $\mathcal{W}_p^0 = L^p$
- $\mathcal{W}_\infty^1 =$ fonctions lipschitziennes

Théorème :

$$\delta_d(\mathcal{B}_p^r) \propto d^{-r} \text{ pour } 1 \leq p \leq +\infty$$

avec :

$$\mathcal{B}_p^r = \{h \in \mathcal{W}_p^r([0,1]) \text{ tel que } ||h^{(r)}||_{L^p} \leq 1\}$$

Les sous-espaces optimaux $\mathcal{H}_D$ sont :
blabla
Théorème :

$$\delta_d(\mathcal{B}_p^r, ||.||_2) = \mathcal{O}(d^{-r/p})$$

Fléau de la dimension !

$p = 10000$, $r = 100$ (très régulier). Erreur d'approximation $\mathcal{O}(10^{-1/100})$
Pour faire une erreur d'approximation de $\frac{1}{100}$, il faut : $D = 10^{200}$

Il faut donc un S.E.V de dimension 10^{200} pourobtenir une erreur d'approximation de $\frac{1}{100}$!

2.2.3 Expressivité des réseaux de neurones

Théorème d'approximation universelle :
Soit $f : \mathbb{R}^p \to \mathbb{R}^q$ une fonction continue.
Soit $K \subset \mathbb{R}^p$ un compact et $\varepsilon > 0$ une précision arbitraire.

Alors il existe f_ε de la forme :
$f_\varepsilon = W_2 \circ \sigma \circ W_1$, avec :

- W_1, W_2 des fonctions affines
- σ une fonction arbitraire qui n'est pas un polynome

telle que :

$$\sup_{x \in K} |f_\varepsilon(x) - f(x)| = ||f_\varepsilon - f||_{L^\infty(K)} \leq \varepsilon$$

Exemple : $\sigma = \text{ReLU} \implies \sigma(x) = \max(0, x)$

$p = q = 1$, $W_1(x) = \langle c, x\rangle + b$, $W_2(x) = \langle c', x\rangle + b'$

$W_2 \circ \sigma \circ W_1 = \sum_{d=1}^{D} c'_d \sigma(\langle c_d, x\rangle + b_d) + b'$

Donc la fonction f_ε est une fonction linéaire par morceaux qui contient D morceaux.

2.3 Erreurs d'estimation

Théorème :
Supposons que $\mathcal{H} = \{h = \sum_{d=1}^{D} \omega_d \psi_d\}$ où (ψ_d) est une famille arbitraire.
Dans ce cas, on a :

$$\sup_{h \in \mathcal{H}} |E(h) - E_N(h)| \leq c\sqrt{\frac{D}{N}}$$

où c est une constante.

Il faut beaucoup d'échantillons N pour avoir une erreur d'estimation faible !