Éléments d'optimisation.

Machine Learning:

Trouver des paramètres $\hat{\theta}$ qui minimisent un critère empirique $\hat{\theta} \in \underset{\theta}{\arg\min}\, \hat{R}(\theta)$.

⚠ Il est parfois possible de résoudre ce problème avec $\nabla_\theta \hat{R}(\theta) = 0$, mais que se passe-t-il lorsque :

(i) On rajoute des contraintes,

(ii) L'équation a des solutions complexes ?

I. Dualité de Lagrange

Considérons le problème (appelé problème primal).

$$\begin{aligned} \min\ & f(\theta) \qquad \theta \in \Theta \\ \text{tq}\ & g_i(\theta) \leq 0 \qquad i = 1, \dots, m \\ & h_j(\theta) = 0 \qquad j = 1, \dots, p \end{aligned} \qquad (P)$$

Hyp : $(P) < +\infty$

1) Fonction Lagrangienne, Problème Dual, Dualité faible

$\forall \theta \in \Theta, (\mu_1, \dots, \mu_m) \in \mathbb{R}^m, (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$,

on définit

$$\mathcal{L}(\theta, \mu, \lambda) = f(\theta) + \sum_{i=1}^{m} \mu_i g_i(\theta) + \sum_{j=1}^{p} \lambda_j h_j(\theta) \quad \text{(Lagrangien)}.$$

Remarque : Si θ satisfait les contraintes de (P) et $\mu \geq 0$,

$$\mathcal{L}(\theta, \mu, \lambda) \leq f(\theta).$$

Définition :
$$p(\theta) \equiv \sup_{\mu \geq 0, \lambda} \mathcal{L}(\theta, \mu, \lambda) = \begin{cases} f(\theta) & \text{si } \theta \text{ satisfait } (P) \\ +\infty & \text{sinon.} \end{cases}$$

Proposition :
$$(P) \equiv \inf_{\theta \in \Theta} p(\theta) = \inf_{\theta \in \Theta} \sup_{\mu \geq 0, \lambda} \mathcal{L}(\theta, \mu, \lambda)$$

Le problème dual est obtenu par inversion du inf et du sup.

Définition :
$$d(\mu, \lambda) \equiv \inf_{\theta \in \Theta} \mathcal{L}(\theta, \mu, \lambda).$$

$$(D) \equiv \sup_{\mu \geq 0, \lambda} d(\mu, \lambda)$$

Théorème (Dualité faible) : $(D) \leq (P)$.

démonstration :

Si θ satisfait les contraintes de (P), et $\mu \geq 0$,

$$\mathcal{L}(\theta, \mu, \lambda) \leq f(\theta)$$

donc $d(\mu, \lambda) \leq (P)$

donc $\sup\limits_{\mu \geq 0, \lambda} d(\mu, \lambda) \leq (P)$ $\square$

Exercice : Trouver le dual de $\inf c^T x$

tq $Ax \leq b$

Solution : $L(x, \mu) = c^T x + \mu^T (Ax - b) = -b^T \mu + (A^T \mu + c)^T x$

donc $d(\mu) = \inf\limits_x L(x, \mu) = -b^T \mu + \inf\limits_x (A^T \mu + c)^T x$

$$= \begin{cases} -b^T \mu & \text{si } A^T \mu + c = 0 \\ -\infty & \text{sinon.} \end{cases}$$

donc $(D) = \sup -b^T \mu$

tq $\mu \geq 0$

$A^T \mu + c = 0$

2) Lagrangien et points de selle

<u>Définition</u> : On dit que $(\theta^*, \mu^*, \lambda^*)$ est un point de selle du Lagrangien si
~~satisfaisant les contraintes~~ ≥ 0
de (P) $\in \Theta$

$$L(\theta^*, \mu, \lambda) \leq L(\theta^*, \mu^*, \lambda^*) \leq L(\theta, \mu^*, \lambda^*) \quad \forall \lambda, \forall \mu \geq 0$$

et $\forall \theta$ satisfaisant les contraintes de (P).

<u>Théorème</u> : $(\theta^*, \mu^*, \lambda^*)$ est un point de selle ssi

$$\begin{cases} \bullet\ L(\theta^*, \mu^*, \lambda^*) = \inf_{\theta \in \Theta} L(\theta, \mu^*, \lambda^*) \\ \bullet\ \forall i, g_i(\theta^*) \leq 0,\ \forall j, h_j(\theta^*) = 0 \\ \bullet\ \forall i, \mu_i^* g_i(\theta^*) = 0 \end{cases}$$

<u>démonstration</u> :

$(\Rightarrow)$ Si $g_i(\theta^*) > 0$, $\mu_i^* \to +\infty \Rightarrow \infty \leq L(\theta^*, \mu^*, \lambda^*)$.

$\Rightarrow \forall i, g_i(\theta^*) \leq 0$

Si $h_j(\theta^*) \neq 0$, on obtient une absurdité similaire

$\Rightarrow \forall j, h_j(\theta^*) = 0$

La troisième contrainte est immédiate.

(⇐) Immédiat.

□

Théorème : Si $(\theta^*, \mu^*, \lambda^*)$ est un point de selle, alors

- θ^* est une solution de (P)
- (μ^*, λ^*) est une solution de (D)
- (D) = (P).

démonstration :

Si $(\theta^*, \mu^*, \lambda^*)$ est un point de selle, alors le théorème précédent assure que θ^* satisfait les contraintes de (P). De plus, on peut écrire

$$f(\theta^*) = f(\theta^*) + \sum_{i=1}^{n} \underbrace{\mu_i^* g_i(\theta^*)}_{=0} + \sum_{j=1}^{p} \lambda_j^* \underbrace{f_j(\theta^*)}_{=0}$$

$$= L(\theta^*, \mu^*, \lambda^*)$$

$$= \inf_{\theta \in \Theta} L(\theta, \mu^*, \lambda^*)$$

$$= d(\mu^*, \lambda^*)$$

d'où le résultat par dualité faible. □

3) Le cas convexe, Condition de Slater

Définition : Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est dite convexe sur $C \subset \mathbb{R}^d$ si

$\forall x_1, x_2 \in C$, $\forall \lambda \in [0,1]$

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

Remarque : Pour que cette définition fasse sens, il faut que

$\forall x_1, x_2 \in C$, $\forall \lambda \in [0,1]$, $\lambda x_1 + (1-\lambda) x_2 \in C$.

Un tel ensemble est dit convexe.

Définition : Le problème (P) est dit convexe si

- f est convexe
- $H = \mathbb{R}^d$
- $\forall i$, g_i est convexe
- $\forall j$, h_j est affine.

Théorème (Condition de Slater)

Si (P) est convexe et qu'il existe un point strictement faisable ($\exists x$ tq $\forall i$, $g_i(x) < 0$ et $\forall j$, $h_j(x) = 0$)

alors (D) = (P).

II. Conditions de KKT

💡 L'étude des points de selle du Lagrangien permet de dvlp les conditions de KKT

1) Qualification des contraintes.

- Contraintes linéaires (ou affines)
- (LICQ) [Linearly independent constraints qualifications]

Un point θ^* ~~[illegible]~~ satisfait les contraintes LICQ si

$$\{\nabla h_j, \forall j\} \cup \{\nabla g_i \mid g_i(\theta^*) = 0\}$$ est linéairement indépendante.

- (Slater) Un point θ^* satisfait les contraintes de Slater si les conditions du théorème du même nom sont vérifiées (Remarque que cette contrainte dépend uniquement de (P)).

2) Conditions de KKT.

Si θ^* est un optimum local de (P) et si θ^* satisfait une des conditions du paragraphe précédent, alors $\exists \mu, \lambda$ tq

(stationnarité) • $\nabla f(\theta^*) + \sum_{i} \mu_i g_i(\theta_i^*) + \sum_j \lambda_j h_j(\theta^*) = 0$

(F.(P)) • $\forall i,\ g_i(\theta^*) \leq 0,\ \forall j,\ h_j(\theta^*) = 0$

(F.(D)) • $\forall i,\ \mu_i \geq 0$

(complémentarité) • $\forall i,\ \mu_i g_i(\theta^*) = 0$

Exercice Étudier $\min x_2$

$$\text{tq} \quad \begin{cases} (x_1-1)^2 + x_2^2 \leq 1 \\ (x_1+1)^2 + x_2^2 \leq 1 \end{cases}$$

3) Conditions Suffisantes

⚠ Souvent, les conditions de KKT ne sont que des conditions ~~n~~ nécessaires

Cependant, elles sont suffisantes si (P) est convexe et satisfait les ~~conditions~~ hypothèses de la condition de Slater.

III Algorithmes d'optimisation

Parfois, il n'est pas possible d'obtenir des solutions en forme close, on utilise alors des algorithmes d'approximation.

Descente de gradient :

Pour résoudre $\min_{\theta} f(\theta)$, on utilise l'algorithme

$$\theta_{t+1} \leftarrow \theta_t - \alpha_{t+1} \nabla f(\theta_t) \quad \text{(Descente de gradient)}$$

En effet, comme (sous des hypothèses faibles)

$$f(\theta+\Delta\theta)=f(\theta)+\langle\nabla f(\theta),\Delta\theta\rangle+o(\|\Delta\theta\|),$$

l'algorithme suit localement la direction de décroissance maximale.

Exercice :

Décrire l'algorithme de descente de gradient pour le problème de SVM

$$\hat{\theta}\in\arg\min\ \frac{1}{n}\sum_{i=1}^{n}(1-y_i\theta^\top x_i)_+ +\lambda\|\theta\|_2^2$$

Algorithme du gradient stochastique

Pour accélérer l'algorithme, il possible de remplacer le gradient par un estimateur, plus facile à calculer.

Exercice :

Que devient l'algorithme précédent si à chaque étape, le gradient est calculé sur un sous-ensemble du jeu de données uniquement ?