TP °2

Le but de ce TP est de revisiter certaines notions vues dans le chapitre 4 qui introduit l'apprentissage automatique et l'optimisation du point de vue pratique.

Rappel
Notions : Différentiabilité, différentes variantes d'algorithmes de descente de gradient, Matlab/Scilab ou n'importe quel outil de programmation et de visualisation.

1 Préliminaire sur les espaces de Hilbert à noyau reproduisant

Nous commençons par présenter quelques rappels sur les espaces de Hilbert, les noyaux et les noyaux reproduisants. Ensuite, les caractéristiques de tels noyaux sont données. Puis nous définissons la notion des espaces de Hilbert à noyau reproduisant. Ensuite, nous introduisons un des éléments fondamentaux des méthodes à noyaux qui est le théorème de Représentation.

Définition 1 *(Espace Hilbert) L'espace de Hilbert est un espace complet muni d'un produit scalaire, c'est-à-dire qu'il s'agit d'un espace de Banach muni d'un produit scalaire.*

Définition 2 *(noyau défini positif) Un noyau défini positif (n.d.p.) sur l'ensemble Ω est une fonction $K : \Omega \times \Omega \to \mathbb{R}$ symétrique:*

$$\forall (x, y) \in \Omega^2, \quad K(x, y) = K(y, x),$$

et qui satisfait, pour tout $N \in \mathbb{N}, (x_1, ..., x_N) \in \Omega^N$ et $(a_1, ..., a_N) \in \mathbb{R}^{\mathbb{N}}$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(x_i, x_j) \geq 0$$

Théorème 1 *Si K est un n.d.p. sur un espace Ω quelconque, alors il existe un espace de Hilbert H muni du produit scalaire $\langle ., . \rangle_H$ et une application $\phi : \Omega \mapsto H$ tels que :*

$$\forall (x, y) \in \Omega^2, \quad K(x, y) = \langle \phi(x), \phi(y) \rangle_H.$$

1.0.1 Espace de Hilbert à noyau reproduisant (RKHS)

Soit Ω un espace quelconque, et $(H, \langle ., . \rangle_H)$ un espace de Hilbert de fonction.

Définition 3 *Une fonction $K : \Omega \mapsto \mathbb{R}$ est appelée un noyau reproduisant (noté n.r.) si et seulement si :*

- *H contient toutes les fonctions de la forme*
$$\forall x \in \Omega, \quad K_x : y \mapsto K(x, y)$$
- *Pour tout $x \in \Omega$ et $f \in H$, on a:*
$$f(x) = \langle f, K_x \rangle_H$$

Si un n.r. existe, H est appelé un espace de Hilbert à noyau reproduisant (RKHS).

1.0.2 Propriétés des n.r. et RKHS

Théorème 2 *(Aronszajn, 1950)*

- *Si un n.r. existe, il est unique.*
- *Un n.r. existe si et seulement si $\forall x \in \Omega$, la fonctionnelle $f \mapsto f(x)$ (de H dans $\mathbb{R}$) est continue.*
- *Un n.r. est un noyau d.p.*
- *Réciproquement, si K est d.p., alors il existe un RKHS ayant K pour n.r.*
- *Si K est un n.r., il vérifie la propriété reproduisante:*
$$\forall (x, y) \in \Omega^2, \quad \langle K_x, K_y \rangle_H = K(x, y)$$

.

1.0.3 Théorème de Représentation

Nous pouvons constater que sous certaines conditions, la solution optimale d'un problème d'optimisation dans H peut s'écrire sous la forme d'une combinaison de noyaux, indépendamment de la dimension de H. Cette constatation est formulée par le théorème suivant :

Théorème 3 *(Théorème de Représentation) Soient un espace non vide Ω, un noyau défini positif K sur $\Omega\times\Omega$, soient $(x_1, y_1), ..., (x_d, y_d) \in \Omega \times \mathbb{R}$, une fonction réelle g strictement monotone croissante sur $[0, \infty[$, et une fonction coût arbitraire c. Soit H l'espace de Hilbert associé au noyau K, avec $K_{x_i} = K(x_i, .)$. Toute fonction $f \in H$ minimisant la fonctionnelle régularisée*

$$c\left((x_1, y_1, f(x_1)), ..., (x_d, y_d, f(x_d))\right) + g\left(\|f\|\right),$$

admet une représentation de la forme

$$f = \sum_{i=1}^{d} \alpha_i K_{x_i}.$$

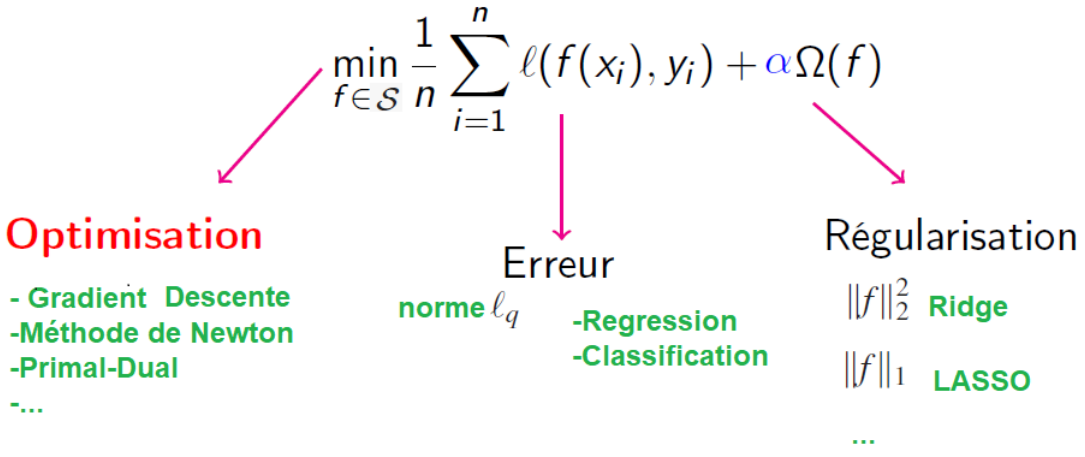
1.0.4 Exemples de RKHS

Des exemples très courants de noyaux RKHS symétriques sont

- linéaire :
$$K(x, x') = x^T x' + c, \quad x, x' \in \mathbb{R}^d, \quad c \in \mathbb{R}$$
- polynômial :
$$K(x, x') = (\alpha x^T x' + c)^d, \quad x, x' \in \mathbb{R}^d, \quad c \in \mathbb{R}$$
- Gaussien :
$$K(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2}), \quad x, x' \in \mathbb{R}^d$$
- exponentiel :
$$K(x, x') = \exp(-\frac{\|x - x'\|}{2\sigma^2}), \quad x, x' \in \mathbb{R}^d$$
- Laplacien :
$$K(x, x') = \exp(-\frac{\|x - x'\|}{\sigma}), \quad x, x' \in \mathbb{R}^d$$
- *Sinc* noyau :
$$K(x, x') = \prod_{i=1}^{n} \frac{\sin(\sigma(x_i - x'_i)}{x_i - x'_i}), \quad x, x' \in \mathbb{R}^d$$

L'apprentissage supervisé : cas général

Données d'apprentissage : $(X_i, y_i)_{i=1,...,n} \in (\mathcal{X} \times \mathcal{Y})^n$.
Modèle : $y = f(x)$, pour un certain $f \in \mathcal{S}$.
Problème à résoudre pour l'apprentissage :



Trouver un modèle qui permet d'accomplir la généralisation, contrôler la complexité, réduire l'erreur, et avec un temps raisonnable.

Nous allons étudier du point de vue pratique l'impact des trois ingrédients cités dans le schéma. A savoir, l'optimisation, la fonction de perte (mesure d'erreur) et la régularisation.

On considère le couple de variables aléatoires (X, Y), tel que X suit une loi uniforme sur $[0, 1]$ et $Y = f^*(X) + \varepsilon$ où $f^*(x) = \cos(2x)$ et ε est un bruit Gaussien indépendant de X de variance unité et de moyenne nulle.

2 Impact du choix de l'espace d'hypothèse

Exemple 1

On se place dans le cadre où la fonction de perte est quadratique $\|\cdot\|^2$, et $\alpha = 0$. Nous choisirons comme outil d'optimisation le gradient descente. Notre objectif est d'étudier l'influence du choix d'un sous espace d'hypothèse $\mathcal{S}$. Nous considérons que

$$\mathcal{S} = \left\{ f : \mathbb{R}^d \mapsto \mathbb{R}, \mid f(x) = \sum_{i=1}^{n} \omega_i K(x, x_i), \quad x \in \mathbb{R} \text{ et } \omega \in \mathbb{R}^n \right\}.$$

Le problème d'apprentissage supervisé revient à chercher $\hat{f} \in \mathcal{S}$ qui réalise le minimum du risque empirique

$$\mathcal{J}(\omega_1, \ldots, \omega_n) := \frac{1}{n} \sum_{i=1}^{n} \|f(x_i) - y_i\|^2.$$

1. Calculer le gradient

$$\nabla \mathcal{J}_\omega(\omega_1, \ldots, \omega_n) = \left(\partial_{\omega_1} \mathcal{J}(\omega_1, \ldots, \omega_n), \ldots, \partial_{\omega_n} \mathcal{J}(\omega_1, \ldots, \omega_n)\right)^T.$$

2. On considère que $K(x, x') = x^T x' + c$ noyau linéaire. En utilisant l'algorithme 1 de gradient descente, estimer $\hat{f}$ en se basant sur les points des données d'entraînement. Afficher le résultat obtenu sur le même graphique que les données. Puis essayer l'estimateur $\hat{f}$ sur les données de test.

3. Afficher les résultats si on cherche à estimer les modèles suivants :

   (a) $f^*(x) = |x|, \quad x \in [-1, 1]$
   (b) $f^*(x) = 3\|x\|_2^3 - 2\|x\|_2^2 + 3\|x\|_2 + 3, \quad x \in [-1, 1]^3$
   (c) $f^*(x) = \sin(x_1 + x_2) \quad x \in [-2, 2]^2$
   (d) $f^*(x) = 3\|x\|_2^3 - 2\|x\|_2^2 + 3\|x\|_2 + 3, \quad x \in [-1, 1]^3$
   (e) $f^*(x) = \left| \frac{1}{3}(x_1 + x_2)^3 - \frac{1}{4}(x_1 + x_2) \right|, \quad x \in [-1, 1]^2$

4. Afficher le risque empirique en fonction des itérations.

5. Calculer l'erreur relative $\frac{\|f^* - y\|_2}{\|y\|_2}$ pour chaque case, puis regrouper les résultats dans un tableau.

6. On répète le même travail en considérant le noyau polynômial.

7. On répète le même travail en considérant le noyau Gaussien. Vérifier l'influence du paramètre σ sur les résultats de prédiction.

8. Comparer les résultats de prédiction obtenus en utilisant les différents noyaux (Gaussien, exponentiel, Laplacien, *Sinc*).

9. En utilisant le noyau qui donne le meilleur résultat d'après le tableau de la question précédente, appliquer l'approche pour la prédiction d'un modèle issu de données réelles.

3 Impact du choix de la régularisation

Notre objectif ici est d'étudier l'influence du terme de régularisation $\Omega(f)$, ainsi que le coefficient α.

Exercice 2:

1. Cas de régularisation de Tikhonov (Regression ridge). Reprendre l'approche étudiée dans l'exercice 1 en considérant cette fois-ci que $\alpha \neq 0$. Calculer le gradient de $\mathcal{J}$ dans ce cas. Étudier l'influence du choix de α sur la convergence de l'algorithme et sur la qualité de prédiction.

2. Cas de régularisation L^1 (Regression LASSO $\Omega(f) = \|f\|_1$). Dans ce cadre on ne peut pas utiliser l'algorithme de gradient car $\Omega(f)$ n'est pas différentiel. Nous avons donc deux façons pour résoudre ce problème d'apprentissage. On utilise soit un algorithme qui n'a pas besoin de calcul du gradient, soit un algorithme basé sur le calcul du proximal. Puisque nous nous intéressons à l'influence de la régularisation, nous allons opter pour une méthode de lissage. Cela consiste à pénaliser le module de la façon suivante :

$$\text{for } \gamma > 0 \quad \|f\|_1^\gamma = \sqrt{(f+\gamma)^2}.$$

Comme $\|f\|_1^\gamma$ est différentiable, on peut appliquer l'algorithme de gradient descente. Reprendre l'approche étudiée dans la question précédente. Calculer le gradient de $\mathcal{J}^\gamma$ dans ce cas. Étudier l'influence du choix de α sur la convergence de l'algorithme et sur la qualité de prédiction. Enfin, tester le rôle de γ.

3. Elastic net regularization : il consiste à utiliser une combinaison de régularisation Ridge et LASSO de la façon suivante

$$\Omega(f) = \lambda_1 \|f\|_2^2 + \lambda_2 \|f\|_1.$$

En utilisant la même pénalisation, reprendre l'approche étudiée dans la question précédente. Calculer le gradient de $\mathcal{J}^\gamma$ dans ce cas. Étudier l'influence du choix de λ_1 et λ_2 sur la convergence de l'algorithme et sur la qualité de prédiction.

4 Impact du choix de l'algorithme d'optimisation

Notre objectif ici est d'étudier l'intérêt de l'algorithme d'optimisation sur la qualité de prédiction des modèles. Nous allons tester 5 algorithmes d'optimisation. A savoir, l'algorithme de gradient descente classique, sa variante stochastique, son accélération Nestrov et l'algorithme de Newton, puis la méthode proximale pour les fonctions coût non-différentielles.

5 Impact du choix de la fonction de perte

Notre objectif ici est d'étudier l'influence du choix de la fonction de perte ℓ. Nous allons nous restreindre à quatre fonctions de perte. A savoir, quadratique $\ell(f) = \|f\|^2$, perte absolue ($\ell(f) = |f|$), ϵ-insensitive loss $\ell(f) = \max(0, |f| - \epsilon)$, Huber loss

$$\ell(f) = \begin{cases} \frac{1}{2} f^2 & \text{si } |f| \leq \delta \\ \delta(|f| - \frac{1}{2}\delta) & \text{sinon} \end{cases}$$