TD 7 - Kernel et Boosting

Préambule : quelques éléments de topologie et d'analyse

- un espace vectoriel E sur $\mathbb{R}$ est un ensemble d'éléments tel qu'il est possible de faire des combinaisons linéaires de ses éléments (E est muni d'une opération d'addition et d'une opération de multiplication par un scalaire) ;
- une fonction $Q : E \times E \to \mathbb{R}$ est un produit scalaire ssi :
  1. elle est symétrique : $Q(x, y) = Q(y, x)$;
  2. elle est bilinéaire : $Q(\lambda_1 x_1 + \lambda_2 x_2, y) = \lambda_1 Q(x_1, y) + \lambda_2 Q(x_2, y)$;
  3. elle est positive $Q(x, x) \geq 0$ et $Q(x, x) = 0 \iff x = 0$

  On notera souvent $Q(x, y) =< x, y >_E$ et la norme d'un produit scalaire $\|x\|_Q = \sqrt{Q(x, x)}$;
- un espace de Hilbert est est un espace vectoriel complet muni d'un produit scalaire ;
- un noyau est une fonction $k : X \times X \to \mathbb{R}$ tel qu'il existe un espace de Hilbert $\mathcal{H}$ et une fonction (de projection ou *feature map*) $\phi : X \to \mathcal{H}$ telle que $\forall x, x' \in X, k(x, x') =< \phi(x), \phi(x') >_{\mathcal{H}}$

Exercice 1 – Noyaux

Q 1.1 Montrez que si K et K' sont deux noyaux (i.e. il existe ϕ et ϕ' telles que $K(x, y) =< \phi(x), \phi(y) >$, $K'(x, y) =< \phi'(x), \phi'(y) >$) :

Q 1.1.1 cK est un noyau pour $c \in \mathbb{R}^+$

Q 1.1.2 $K + K'$ est un noyau ;

Q 1.1.3 KK' est un noyau ;

Q 1.1.4 $(1+ < x, x' >)^d$ est un noyau.

Exercice 2 – RKHS

Soit $x_1, \cdots, x_n \in X$, une fonction $k : X \times X \to \mathbb{R}$, la matrice de Gram de K est la matrice $\mathcal{K} := k_{i,j} = k(x_i, x_j)$. Une matrice est dite définie semi-positive si $\forall c_i \in \mathbb{R}, \sum_{i,j} c_i c_j k_{i,j} \geq 0$. Dans ce cas, la fonction est dite également définie positive.

Q 2.1 Exprimez $\sum_{i,j} c_i c_j k_{i,j}$ par un produit scalaire. Montrez qu'un noyau est défini positif.

Q 2.2 Le but de cette question est de montrer la contraposée, qu'une fonction symétrique semi définie positive $k : X \times X \to \mathbb{R}$ est un noyau. Pour cela, il nous faut trouver un espace hilbertien $\mathcal{H}$, un produit scalaire $Q : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ et une projection $\phi : X \to \mathcal{H}$ telle que $k(x, y) = Q(\phi(x), \phi(y)) \ \ \forall x, y \in X$. On va considérer $\mathcal{H}$ l'espace vectoriel engendré par les fonctions de la forme $y \to k(y, x)$ pour tout $x \in X$. Un élément de $\mathcal{H}$ est donc une fonction de $X \to \mathbb{R}$.
Soit $\Phi : X \to \mathcal{H} \ := k(., x)$ un mapping de X aux fonctions de $\mathcal{H}$, $\Phi(x)(x') = k(x', x)$. Soient $\alpha_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}$, $x_i \in X$, $x'_i \in X$ pour $i \in \{1..n\}$. On définit :

$$f(.) = \sum_{i=1}^{n} \alpha_i \Phi(x_i)(.), \quad g(.) = \sum_{i=1}^{n} \beta_i \Phi(x'_i)(.), \quad Q(f, g) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k(x_i, x'_j)$$

f et g sont bien dans $\mathcal{H}$, vu que ce sont des combinaisons linéaires d'éléments de $\mathcal{H}$.

Q 2.2.1 Montrez que Q peut s'exprimer uniquement à l'aide des β_j et $f(x'_j)$, ou des α_i et $g(x_i)$

Q 2.2.2 Montrez que $Q(f,g)$ est un produit scalaire de f et g (on pourra alors remplacer $Q(f,g)$ par $\langle f,g\rangle$). Pour cela, il s'agit de démontrer que :

- Q est symétrique
- Q est bilinéaire
- $Q(f,f) \geq 0$ (on montrera dans la dernière question que $Q(f,f) = 0 \iff f = 0$).

Q 2.2.3 Que vaut $Q(k(.,x),f)$? $Q(k(.,x),k(.,x'))$? Justifiez le nom de k : *reproducing kernel*.

Q 2.2.4 En admettant que $Q(f,g)^2 \leq Q(f,f)^2 Q(g,g)^2$, montrez que $|f(x)|^2 \leq k(x,x).Q(f,f)$. Concluez.

Exercice 3 – Noyaux sur les chaînes de caractères

Soit S une séquence de mots sur un alphabet $\mathcal{A}$ fini. Montrez que :

1. $K(x,x')$ = nombre de sous-chaînes de longueur 5 que x et x' ont en commun est un noyau ;
2. $K(x,x') = 1$ si x et x' ont au moins une sous-chaîne de longueur 5 en commun, 0 sinon, n'est pas un noyau (indice : considérez 3 chaînes x,x' et x'').

Exercice 4 – Boosting

Rappel de l'algorithme : on cherche à construire une combinaison de classifieurs faibles $f_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ de manière itérative, de manière à prendre mieux en compte à une itération donnée les erreurs des itérations précédentes. Pour cela, une distribution de poids sur les exemples est considérée et adaptée à chaque itération afin d'augmenter le poids des exemples mal classés, et de baisser le poids des exemples bien classés. Soit $D_t = (w_t(1), \ldots, w_t(n))$ la distribution des poids des exemples au pas t, D_1 correspondant à la distribution uniforme. L'algorithme consiste en l'itération de la procédure suivante :

1. Choisir h_t qui minimise l'erreur selon D_t
2. Calculer l'erreur ϵ_t associé au classifieur h_t selon D_t
3. Fixer $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$
4. Mettre à jour D_{t+1} : $w_{t+1}(i) = \frac{1}{Z_t} w_t(i) e^{(-\alpha_t y_i h_t(x_i))}$, avec Z_t facteur de normalisation

Q 4.1 Introduction

Q 4.1.1 Rappeler le principe et les différences entre le boosting et le bagging. Soit le jeu de données suivant : $Y^+ = \{(-3,-1),(-3,1),(3,-1),(3,1)\}, Y^- = \{(-1,-1),(-1,1),(1,-1),(1,1)\}$. En considérant comme classifieur faible des stumps (fonction de type $\mathbb{1}_{x_i<\theta_i}$, correspondant à un arbre de décision à 2 feuilles), quels sont les deux premiers classifieurs appris ? Sont-ils suffisant pour la classification parfaite ?

Q 4.1.2 Exprimer l'erreur ϵ_t en fonction d'un coût donné $l(x,y)$ et des $w_t(i)$.

Q 4.1.3 Comment varie α_t en fonction de ϵ_t ? Que se passe-t-il pour $w_{t+1}(i)$ si l'exemple i est bien classifié ? mal classifié ?

On va montrer dans la suite que l'algorithme optimise bien l'erreur d'apprentissage. Le principe de la démonstration consiste à montrer que à chaque pas t, l'erreur est borné par $Z = \prod_{j=1}^{t} Z_j$, et que ce produit converge vers 0.

Q 4.2 Nous allons montrer d'abord que le choix de α_t conduit à minimiser Z_t.

Q 4.2.1 Exprimer Z_t et ϵ_t en fonction de $w_t(i)$, α_t et $y_i h_t(x_i)$.

Q 4.2.2 Exprimer $\frac{\partial Z_t}{\partial a_t}$. En déduire la valeur de α_t qui minimise Z_t.

Q 4.2.3 Donner l'expression de Z_t en fonction de ϵ_t pour α_t optimal.

Q 4.2.4 Soit $\gamma_t = \frac{1}{2} - \epsilon_t$. Sachant que $1 - x \leq e^{-x}$, montrer que Z décroît exponentiellement en fonctio n de t.

Q 4.3 Nous allons montrer maintenant que Z est une borne supérieure de l'erreur 0-1.

Q 4.3.1 Exprimer $w_{t+1}(i)$ en fonction de $h_j(x), \alpha_j(i), Z_j,\ 1 \leq j \leq t$, puis en fonction de $f_t(x_i)$. En déduire une expression de $\sum_i w_t(i)$ en fonction des Z_j et $y_i f_t(x_i)$, puis une expression de $Z = \prod_j Z_j$ en fonction de $y_i f_t(x_i)$

Q 4.4 Montrez que l'erreur $0-1$ est bornée par le coût exponentiel $l(x, y) = e^{(-yf(x))}$. En déduire que Z est un majorant de l'erreur $0-1$.

Q 4.4.1 Conclure sur la décroissance exponentielle de l'erreur.