TD 6

Exercice 1 – 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 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 1.2 Exprimer l'erreur ϵ_t en fonction d'un coût donné $l(x, y)$ et des $w_t(i)$.

Q 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 1.4 Nous allons montrer d'abord que le choix de α_t conduit à minimiser Z_t.

Q 1.4.1 Exprimer Z_t et ϵ_t en fonction de $w_t(i)$, α_t et $y_i h_t(x_i)$.

Q 1.4.2 Exprimer $\frac{\partial Z_t}{\partial a_t}$. En déduire la valeur de α_t qui minimise Z_t.

Q 1.4.3 Donner l'expression de Z_t en fonction de ϵ_t pour α_t optimal.

Q 1.4.4 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 1.5 Nous allons montrer maintenant que Z est une borne supérieure de l'erreur 0-1.

Q 1.5.1 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 1.6 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 1.6.1 Conclure sur la décroissance exponentielle de l'erreur.

Exercice 2 – Entropie

Q 2.1 On lance un dé truqué n fois de manière indépendante, la probabilité de chaque chiffre étant $p_k, k = 1..6$.

Q 2.1.1 Exprimer la probabilité d'obtenir la suite $(x_1, \ldots, x_n)$ en fonction des p_k et des n_k - le nombre de fois où k apparaît dans le tirage.

Q 2.1.2 Vers quelle expression tend n_k lorsque n tend vers l'infini ? En déduire une expression de la probabilité d'une suite "typique" en fonction de l'entropie $H(p) = -\sum_k p_k log_2(p_k)$. Que remarquez vous pour des valeurs fortes/faibles de H ?

Q 2.2 Quelques propriétés de la fonction entropie. On appelle vecteur de probabilité un vecteur qui représente une distribution de probabilité discrète à n modalités : $\mathbf{p} = (p_1, \ldots, p_n)$ tel que $\sum_{i=1}^n p_i = 1$ et $p_i \geq 0$.

Q 2.2.1 Montrer que pour $H(\mathbf{p}) \geq 0$

Q 2.2.2 Soit $\mathbf{p}$ et $\mathbf{q}$ deux vecteurs de probabilité de même dimension.
Montrer d'abord que $log(x) \leq x - 1$ en utilisant le fait que la fonction log est concave. En déduire que $-\sum_{i=1}^n p_i log(p_i) \leq -\sum_{i=1}^n p_i log(q_i)$.

Q 2.2.3 En déduire que $H(\mathbf{p}) \leq \log(n)$ pour $\mathbf{p}$ de dimension n.

Q 2.3 On appelle distance de Kullback-Leibler la fonction $D(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^n p_i log(\frac{p_i}{q_i})$

Q 2.3.1 Montrer l'inégalité de Jensen : si f est une fonction convexe dérivable, $\mathbf{p}$ un vecteur de probabilité et t_i des réels quelconques, alors $\sum_{i=1}^n p_i f(t_i) \geq f(\sum_{i=1}^n p_i t_i)$ (indication : procéder par récurrence et penser à poser $p'_i = \frac{p_i}{1-p_n}$).

Q 2.3.2 En déduire que pour une fonction convexe, $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. Trouver une autre démonstration pour la question 2.2.2

Q 2.3.3 Montrer que $D(\mathbf{p}||\mathbf{q}) \geq 0$ avec égalité ssi $\mathbf{p} = \mathbf{q}$. Est-ce que D est une distance ?

Q 2.3.4 On considère $x_1, \ldots x_N$ des observations i.i.d. dans $\mathcal{X}$ un ensemble discret fini. La distribution empirique observée est définie par $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^N \delta(x - x_i)$ avec δ la fonction de dirac (0 partout sauf en 0 où elle vaut 1). Soit p_θ une distribution paramétrisée par θ. Montrer que maximiser la vraisemblance revient à minimiser $D(\hat{p}||p_\theta)$.