ENSAE
3ème année

Apprentissage en ligne et agrégation
Pierre Alquier

TP n° 1

Exercice 1. Jeux de données *wine*.

1) Ouvrir (avec R ou Python) les deux jeux de données proposés (*white* et *red*).
2) Implémenter l'algorithme de gradient en ligne avec des prédicteurs linéaires $f_\theta(x) = \langle \theta, x \rangle$ et la fonction de perte absolue $\ell(y', y) = |y - y'|$. Tester cet algorithme sur le jeu de données *white* :
   — comparer plusieurs versions, avec un pas η_t fixe ou dépendant du temps ;
   — comparer cette méthode de prédiction à la méthode élémentaire qui prédit $\hat{y}_t = 5$ pour tout t ;
   — proposer des représentations graphiques.
3) Tester maintenant cet algorithme sur le jeu de données *red* :
   — comparer à la méthode élémentaire qui prédit $\hat{y}_t = 5$ pour tout t ;
   — comparer à un prédicteur statique fabriqué uniquement sur le jeu de données *white* ;
   — comparer l'effet de considérer *white* et *red* comme deux problèmes différents, ou comme un seul problème (sans ré-initiliser le paramètre au début du jeu de données *red*).

Exercice 2. Frank Rosenblatt a proposé en 1956 un algorithme appelé **Perceptron** pour la classification linéaire dans le cas réalisable. Cet algorithme est considéré comme le premier réseau de neurones (à une seule couche).

On suppose que l'on a une suite $(x_t, y_t)_{t \geq 1}$ avec les proprietés suivantes. Tout d'abord, $x_t \in \mathcal{B}_d(0, R) = \{x \in \mathcal{R}^d : \|x\| \leq R\}$. D'autre part, on pose pour $w \in \mathbb{R}^d$ et $b \in \mathbb{R}$, pour tout x, $f_{w,b}(x) = w^T x + b$. Enfin, on suppose qu'il existe $(\tilde{w}, \tilde{b})$ tels que pour tout t, $y_t = \text{sgn}[f_{\tilde{w},\tilde{b}}(x_t)]$. Quitte à renormaliser, on suppose que $\|\tilde{w}\| = 1$. En fait, on suppose même qu'il existe $\gamma > 0$ tel que

$$y_t f_{\tilde{w},\tilde{b}}(x_t) \geq \gamma$$

pour tout t. On dit qu'on peut séparer les $+1$ et les -1 avec une marge γ. L'algorithme est alors défini par :

init : $w_0 = 0$, $b_0 = 0$.
prev. t : $\hat{y}_t = \mathrm{sgn}[f_{w_t,b_t}(x_t)]$.
update : $w_{t+1} = w_t + y_t x_t$ et $b_{t+1} = b_t + y_t R^2$ si $y_t \hat{y}_t \leq 0$; $w_{t+1} = w_t$ et $b_{t+1} = b_t$ sinon.

Pour alléger les notations, il est commode de poser $X_t = (x_t, R)$, $W_t = (w_t, b/R)$ et $\tilde{W} = (\tilde{w}, \tilde{b}/R)$.

1) Pourquoi est-il raisonnable de supposer $\tilde{b} \leq R$?
2) Par récurrence, démontrer que $\forall t \in \mathbb{N}$, on a

$$W_{t+1}^T \tilde{W} \geq \gamma \sum_{i=1}^{t} \mathbf{1}(\hat{y}_i \neq y_i).$$

3) Démontrer, également par récurrence, que pour tout t,

$$\|W_{t+1}\|^2 \leq 2R^2 \sum_{i=1}^{t} \mathbf{1}(\hat{y}_i \neq y_i).$$

4) En déduire le Théorème de Novikoff :

$$\forall t \in \mathbb{N}, \quad \sum_{i=1}^{t} \mathbf{1}(\hat{y}_i \neq y_i) \leq \left(\frac{2R}{\gamma}\right)^2.$$

Exercice 3. On suppose données une famille dénombrable de fonctions $(f_n, n \in \mathbb{N}^*)$ avec $f_i : \mathcal{X} \to \{0, +1\}$ et une suite d'observations $(x_t, y_t)_{t \geq 1}$ vérifiant $x_t \in \mathcal{X}$ et $\exists i_0, \forall t, y_t = f_{i_0}(x_t)$. Il s'agit en gros du modèle d'apprentissage de Gold (celui-ci se restreignant au cas de fonctions f_n récursives totales).

1) On définit une adaptation de l'algorithme **Consistent** à ce contexte : $V(1) = \mathbb{N}^*$, et à chaque étape t, $i(t) = \min V(t)$, $\hat{y}_t = f_{i(t)}(x_t)$ et $V(t+1) = \{i \in V(t) : f_i(x_t) = y_t\}$. Vérifier (c'est assez immédiat) que

$$\forall T \in \mathbb{N} : \quad \sum_{t=1}^{T} \mathbf{1}(y_t \neq \hat{y}_t) \leq i_0 - 1.$$

2) On définit maintenant une version bayésienne de l'algorithme **Halving**, que l'on notera **Bayesian Halving** : on fixe une loi de probabilité π sur $\mathbb{N}^*$ et on pose $V_1 = \mathbb{N}^*$. A chaque étape t, on pose

$$\hat{y}_t = \begin{cases} 1 \text{ si } \frac{\sum_{i \in V_t} \pi(i) f_i(x_t)}{\pi(V_t)} \geq \frac{1}{2} \\ \\ 0 \text{ sinon,} \end{cases}$$

puis $V_{t+1} = \{i \in V_t : f_i(x_t) = y_t\}$.

(a) Démontrer que

$$\forall T \in \mathbb{N} : \quad \sum_{t=1}^{T} \mathbf{1}(y_t \neq \hat{y}_t) \leq \log_2\left(\frac{1}{\pi(i_0)}\right).$$

(b) En choisissant π de façon adéquate, démontrer le Théorème de Barzdin et Freivalds qui dit qu'il existe deux constantes $a > 0$, $b > 0$ (à préciser) et une stratégie de prédictions vérifiant

$$\forall T \in \mathbb{N} : \quad \sum_{t=1}^{T} \mathbf{1}(y_t \neq \hat{y}_t) \leq a + b \log_2(i_0)$$

(la stratégie utilisée par ces deux auteurs était légèrement différente, mais reposait aussi sur l'idée de **Halving**).