Apprentissage statistique: TD2
Online Learning

Emile Contal
http://econtal.perso.math.cnrs.fr/teaching

22 janvier 2015

Exercice 1.

1. **(Perceptron)** On rappelle le cadre et les notations du problème de classification supervisée séquentielle :

   - $(x_t, y_t)_{t\geq 1}$ une suite d'observations dans $\mathbb{R}^d \times \{-1, +1\}$,
   - $\forall t \geq 1,\ \ \|x_t\| = 1$,
   - $\exists w^* \in \mathbb{R}^d$ t.q $\|w^*\| = 1$ et $\forall t \geq 1,\ y_t \cdot \langle w^*, x_t\rangle > 0$.

   On rappelle également l'algorithme du perceptron :

   Algorithm 1: Perceptron

   $w_1 \leftarrow \mathbf{0}$
   for $t = 1, \ldots, T$ **do**
   $\quad \widehat{y}_t \leftarrow \mathrm{sng}(\langle w_t, x_t\rangle)$
   $\quad$ **if** $\widehat{y}_t = y_t$ **then**
   $\quad\quad w_{t+1} \leftarrow w_t$
   $\quad$ **else**
   $\quad\quad w_{t+1} \leftarrow w_t + y_t x_t$

   (a) Proposez une généralisation de l'Algorithme 1 dans le cas où l'hyperplan séparateur ne passe pas nécessairement par l'origine. De même dans le cas où les x_t ne sont pas restreints à la sphère unité.

   (b) On définit pour tout $T \geq 1$, $\gamma_T = \min_{t\leq T} |\langle w^*, x_t\rangle|$. Montrer que $\langle w^*, w_{t+1}\rangle \geq \langle w^*, w_t\rangle + \gamma_T$ lorsque $\widehat{y}_t \neq y_t$.

   (c) Montrer que $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$.

   (d) Donner une borne inférieure pour $\langle w^*, w_{t+1}\rangle$ en fonction de γ_T et M_T, le nombre de points mal classés au bout de T itérations. Donner une borne supérieure pour $\|w_{t+1}\|$ en fonction de M_T. En déduire une majoration de M_T en fonction de γ_T.

2. **(Perceptron à vastes marges)** On définit l'algorithme du perceptron à vaste marge de la même façon que l'Algorithme 1, où la prédiction est remplacée par :

$$\widetilde{y}_t = \begin{cases} +1 & \text{si } \frac{\langle w_t, x_t \rangle}{\|w_t\|} > \frac{\gamma_T}{2} \\ -1 & \text{si } \frac{\langle w_t, x_t \rangle}{\|w_t\|} < -\frac{\gamma_T}{2} \\ 0 & \text{sinon,} \end{cases}$$

et l'on considère qu'un point est mal classé lorsque $\widetilde{y}_t = 0$.

   (a) Montrer que $\|w_{t+1}\| \leq \|w_t\| + \frac{1}{2\|w_t\|} + \frac{\gamma_T}{2}$.

   (b) En déduire une majoration $M_T \leq \frac{8}{\gamma_T^2}$ dans ce cas.

   (c) Commenter les avantages de l'algorithme à vaste marge.

3. **(Perceptron et noyau)** On ne suppose plus qu'il existe un hyperplan séparateur dans $\mathbb{R}^d$, mais qu'il existe un espace vectoriel $\mathcal{H}$ de dimension quelconque (potentiellement infinie), une fonction $\phi : \mathbb{R}^d \to \mathcal{H}$ et un vecteur $h^* \in \mathcal{H}$ tels que les points $\phi(x_t)$ soient séparés par h^*.

   (a) Soit $k(a, b) = \langle \phi(a), \phi(b) \rangle_{\mathcal{H}}$. Proposez une généralisation de l'Algorithme 1 utilisant k.

   (b) Commenter les avantages de cet algorithme.

Exercice 2.

1. **(Exponential Weighted Average)** On se donne $\mathcal{C}$ une famille de N experts proposant à chaque temps t des prédictions $\{\widehat{y}_{t,i}, 1 \leq i \leq N\}$. Soit $\ell : \mathbb{R}^2 \to [0, 1]$ une fonction de perte, convexe en son premier argument.

   Fixons l'horizon T. Nous cherchons à minimiser R_T, notre regret par rapport au meilleur expert :

$$R_T = \sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \min_i \sum_{t=1}^{T} \ell(\widehat{y}_{t,i}, y_t)$$

   où $(\widehat{y}_t)_{t \leq T}$ sont les prédictions de notre stratégie.

   On rappelle l'algorithme EWA de paramètre $\eta > 0$:

Algorithm 2: Exponential Weighted Average

$\forall i \leq N$, $w_{1,i} \leftarrow 1$
for $t = 1, \ldots, T$ **do**
  $\widehat{y}_t \leftarrow \dfrac{\sum_{i=1}^{N} w_{t,i} \widehat{y}_{t,i}}{\sum_{i=1}^{N} w_{t,i}}$
  $w_{t+1,i} \leftarrow w_{t,i} \exp\Big(- \eta \ell(\widehat{y}_{t,i}, y_t) \Big)$

(a) Soit $\Phi_t = \log \sum_{i=1}^{N} w_{t,i}$. Montrer que pour tout $1 < t \leq T$, on a $\Phi_t - \Phi_{t-1} \leq -\eta \ell(\widehat{y}_t, y_t) + \frac{\eta^2}{8}$.
Indice : voir les moyennes pondérées comme des espérances, et utiliser l'inégalité de Hoeffding prouvée au TD précédent.

(b) Montrer que $\Phi_T - \Phi_0 \geq -\eta \min_i \sum_{t=1}^{T} \ell(\widehat{y}_{t,i}, y_t) - \log N$.

(c) En déduire une borne sur R_T.

(d) Quel choix de η préconisez vous ?

2. **(Doubling Trick)** L'horizon T est maintenant supposé fini mais inconnu. Soient $(I_k)_{k\in\mathbb{N}}$ la partition de $\mathbb{R}^+$ définie par les intervalles $I_k = [2^k, 2^{k+1} - 1]$. On considère l'Algorithme 2 de paramètre η_k qui dépend de l'intervalle de temps I_k. On note L_{I_k} la perte subie sur l'intervalle I_k.

(a) Donner une majoration du regret sur I_k.

(b) On pose $n = \lceil \log_2(T+1) \rceil$. Déduire une majoration de $L_T = \sum_{k=0}^{n} L_{I_k}$.

(c) Choisissez η_k de sorte à optimiser la majoration précédente, puis montrer pour tout T :

$$R_T \leq \frac{1}{\sqrt{2}-1}\sqrt{T \log N} + \frac{1}{\sqrt{2}}\sqrt{\log N}\,.$$