Apprentissage statistique: TD3
The Perceptron Algorithm and Bregman Divergence

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

18 janvier 2015

Exercice 1.
On se place dans le cadre de la classification linéaire 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 $\forall t \geq 1,\ y_t \cdot \langle w^*, x_t \rangle > 0$,
- $\gamma = \min_t |w^* \cdot x_t|$.

1. **(Rappels)** Rappelez les garanties théoriques sur le nombre d'erreurs M_T faites par l'algorithme du Perceptron et l'algorithme de Winnow après T itérations. Dans quels cas suggérez-vous d'utiliser l'algorithme de Winnow ?

2. **(Perceptron à vastes marges)** On définit l'algorithme du perceptron à marge γ de la même façon que l'Algorithme 1, où la prédiction est remplacée par :
$$\widehat{y}_t = \begin{cases} +1 & \text{si } \frac{w_t \cdot x_t}{\|w_t\|} > \frac{\gamma}{2} \\ -1 & \text{si } \frac{w_t \cdot x_t}{\|w_t\|} < -\frac{\gamma}{2} \\ 0 & \text{sinon,} \end{cases}$$
et l'on considère qu'un point est mal classé lorsque $\widehat{y}_t = 0$.

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

   (b) Rappeler la borne inférieure de $\|w_{t+1}\|$ en fonction de M_T et γ. En déduire une majoration $M_T \leq \frac{8}{\gamma^2}$.
   Indice : regarder le dernier $t \leq T$ où $\|w_t\| > \frac{2}{\gamma}$

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

Algorithm 1: Perceptron

$w_1 \leftarrow \mathbf{0}$
for $t = 1, \dots, T$ **do**
  Receive x_t
  $\widehat{y}_t \leftarrow \text{sgn}(w_t \cdot x_t)$
  Receive y_t
  if $\widehat{y}_t = y_t$ **then**
    $w_{t+1} \leftarrow w_t$
  else
    $w_{t+1} \leftarrow w_t + y_t x_t$

Algorithm 2: Winnow

$w_1 \leftarrow \mathbf{1}/N$
for $t = 1, \dots, T$ **do**
  Receive x_t
  $\widehat{y}_t \leftarrow \text{sgn}(w_t \cdot x_t)$
  Receive y_t
  if $\widehat{y}_t = y_t$ **then**
    $w_{t+1} \leftarrow w_t$
  else
    for $i = 1, \dots, N$ **do**
      $w_{t+1,i} \leftarrow \frac{w_{t,i} e^{\eta y_t x_{t,i}}}{\sum_{j=1}^{N} w_{t,j} e^{\eta y_t x_{t,j}}}$

3. **(Perceptron et noyau)**

   (a) Proposez une réécriture de l'Algorithme 1 telle que les poids sont attribués aux x_t, faisant apparaître des produits scalaires entre x_t et les $\{x_s\}_{s \leq T}$.

   (b) 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^*. Soit $k(a, b) = \langle \phi(a), \phi(b) \rangle_{\mathcal{H}}$. Adapter l'algorithme précédent à ce cas plus général.

   (c) Commenter les avantages de cet algorithme.

Exercice 2.

1. **Divergence de Bregman**
   Soit $F : \mathbb{R}^d \to \mathbb{R}$ une fonction strictement convexe dérivable au moins deux fois, f son gradient, et f^{-1} l'inverse de son gradient. La divergence de Bregman entre u et v dans $\mathbb{R}^d$ est l'erreur de l'approximation de Taylor de $F(u)$ en v :
   $$d_F(u,v) = F(u) - F(v) - f(v) \cdot (u - v)\,.$$
   (a) Montrer que l'Algorithme 3 est équivalent à l'Algorithme du Perceptron lorsque $F(u) = \frac{1}{2}\|u\|^2$ et $\eta = 1$. Calculer $d_F(u,v)$ dans ce cas.
   (b) Montrer que l'Algorithme 3 est équivalent à l'Algorithme de Winnow lorsque $F(u) = \sum_{i=1}^d \big(u_i \log u_i - u_i\big)$. Calculer $d_F(u,v)$ dans ce cas. Dans le cas où $\sum_{i=1}^d u_i = \sum_{i=1}^d v_i = 1$, comparer avec l'entropie relative $d_{KL}(u,v) = \sum_{i=1}^d u_i \log \frac{u_i}{v_i}$.
   (c) Montrer que $d_F(u,v) \geq 0$ avec égalité lorsque $u = v$, et que $d_F(\cdot, v)$ est convexe.
   (d) Montrer que $d_F(u,w) = d_F(u,v) + d_F(v,w) + (f(w) - f(v)) \cdot (v - u)$.

2. **p-Norm Perceptron**
   On considère maintenant l'Algorithme 3 avec $F(u) = \frac{1}{2}\|u\|_q^2$ avec $p \in [2,\infty]$ et $p^{-1} + q^{-1} = 1$. On suppose maintenant que $\forall t \geq 1$, $\|x_t\|_p^2 \leq 1$.
   (a) Calculer f et f^{-1}. En déduire $d_F(u,v) = F(u) + F(v) - f(v)v$.
   (b) Montrer que lorsque $f(v) = f(u) + x$, alors:
   $$d_F(u,v) = \frac{1}{2}\|f(v)\|_p^2 - \frac{1}{2}\|f(u)\|_p^2 - xu \leq \frac{p-1}{2}\|x\|_p^2\,.$$
   Indice pour l'inégalité : utiliser Taylor puis l'inégalité de Hölder avec la exposants conjugués $p/(p-2)$ et $p/2$.
   (c) Soit w^* un hyperplan séparateur de marge γ, $\min_t |w^* \cdot x_t| = \gamma$ et tel que $\|w^*\|_q^2 \leq 1$. Montrer que lorsque $\widehat{y}_t \neq y_t$,
   $$d_F(w^*, w_t) - d_F(w^*, w_{t+1}) \geq \eta\gamma - \eta^2 \frac{p-1}{2}\,.$$
   (d) Donnez une borne sur le nombre d'erreurs de l'Algorithme 3, puis optimiser η.

Algorithm 3: Online Learning with Bregman Divergence

$w_1 \leftarrow \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  Receive x_t
  $\widehat{y}_t \leftarrow \text{sgn}(w_t \cdot x_t)$
  Receive y_t
  if $\widehat{y}_t = y_t$ **then**
    $w_{t+1} \leftarrow w_t$
  else
    $w_{t+1} \leftarrow f^{-1}(f(w_t) + \eta y_t x_t)$