Boosting pour la Classification Binaire Supervisée

Travaux Pratiques

- AdaBoost -

A partir d'un échantillon d'apprentissage $\mathcal{D}_n = \{(X_i, Y_i) : \ 1 \leq i \leq n\}$, l'algorithme AdaBoost vise à minimiser une version convexifiée du risque empirique correspondant à la fonction de perte

$$l_e(y, f) = \exp(-yf).$$

1. Justifier l'utilisation de la perte exponentielle en calculant $f^* = \arg_f \mathbb{E}[l_e(Y, f(X))]$. Discuter son intérêt.
2. Montrer que si $L_e(f_n) = \mathbb{E}[l_e(Y, f_n(X))] \to L_e(f^*)$, alors $L(f_n) = \mathbb{P}\{Yf(X) < 0\} \to L^* = \inf_f L(f)$. Que peut-on dire de la vitesse ?
3. En vous inspirant de la démonstration de l'inégalité de Vapnik-Chervonenkis, montrer comment on peut contrôler l'excès de risque convexifié du minimiseur du risque empirique convexifié.
4. La minimisation directe du risque empirique convexifié

$$\widehat{L}_e(f) = \frac{1}{n} \sum_{i=1}^{n} l_e(Y_i f(X_i))$$

est une tâche difficile en pratique. On peut chercher à s'approcher pas à pas d'une solution en ajoutant à une fonction de décision courante $f_{m-1}(x)$ un terme $\beta_m \cdot C_m(x)$, où $\beta_m \in \mathbb{R}$ et C_m est une règle de classification (simple dans la cas du "slow learning", typiquement un arbre de décision de faible profondeur). Démontrer que cette approche itérative conduit exactement à l'algorithme AdaBoost.

- Application -

5. Ecrire une routine mettant en oeuvre AdaBoost avec des arbres de profondeur 1 ("stumps"), puis 2 (sans élagage).
6. Appliquer AdaBoost sur les données relatif au cancer du sein de l'UCI data repository (http ://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29), découpées en deux échantillons : apprentissage et test. Tracer les erreurs d'apprentissage et de test en fonction du nombre d'itérations du boosting (pour le risque non convexifié). Que remarquez vous ? Que se passe-t-il si la profondeur des arbres de classiification est beaucoup plus grande ?
7. Ecrire une routine mettant en oeuvre Random Forest avec des arbres de profondeur 2 (sans élagage), en n'utilisant que 20% des variables explicatives à chaque scission.
8. Comparer les performances sur les données "Breast Cancer" de l'UCI.

- Généralisation -

9. Dans le cas où l'on considère la fonction de perte "logit"

$$l_{logit}(y, f) = \log\left(\frac{1}{1 + e^{-2Yf}}\right),$$

calculer le minimiseur du risque théorique ainsi convexifié. Que devient la procédure itérative décrite précédemment dans ce cas ?