TME 7 - Bagging, Boosting

Forêts aléatoires

En utilisant les fonctions du module `sklearn.ensemble`, expérimenter les forêts aléatoires sur les données artificielles 2D habituelles. Vous étudierez comment évoluent les frontières de décision, l'erreur en apprentissage et en test en fonction de la profondeur des arbres, du nombre d'arbres et des autres paramètres disponibles. Expérimenter également sur les données IMDB du premier TME.

Boosting : AdaBoost

Rappel de l'algorithme pour un ensemble d'apprentissage $E = \{x^i, y^i\}_{i=0}^N$

- Initialiser la distribution sur les exemples $D_0(i) = \frac{1}{N}$
- Répéter :
  1. Apprendre h_t sur la distribution D_t
  2. Calculer l'erreur $\epsilon_t = \sum_i D_t(i)\mathbf{1}_{h_t(x^i)\neq y^i}$
  3. Fixer $\alpha_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
  4. Fixer $D_{t+1}(i) = \frac{1}{Z_t} D_t(i) e^{-\alpha_t y_i h_t(x^i)}$, avec Z_t normalisation de la distribution.
- Le classifieur est $F(x) = \sum_t \alpha_t h_t(x)$

Programmer l'algorithme du boosting. Expérimenter sur les données artificielles et réelles. Regarder en particulier l'évolution des poids et l'évolution de $Z = \prod Z_t = \frac{1}{N} \sum_{i=1}^N e^{-y^i F(x^i)}$.