Boosting

- But:
  - combiner des classifieurs “faibles” $h^{(1)}, h^{(2)}, \ldots \in H$ qui sont à peine meilleurs qu'une décision aléatoire
  - d'une façon itérative
- Exemples de classifieurs de base:
  - petits arbres de décisions
  - arbres d'une feuille: decision stumps
  - réseaux de neurones avec quelques unités cachées

Boosting

- Idée 1: pondération $w_1, \ldots, w_n$ des points d'entraînement
  - attribuer les plus grands poids aux points “difficiles”
  - si $h^{(t)}(\mathbf{x}_j) = y_j$ alors $w_j \downarrow$
  - si $h^{(t)}(\mathbf{x}_j) \neq y_j$ alors $w_j \uparrow$
- Idée 2: pondération $\alpha^{(1)}, \ldots, \alpha^{(t)}$ des classifieurs de base
  - le poids de $h^{(t)}$ est monotone décroissant en l'erreur de $h^{(t)}$

Boosting

- Démarche
  - entraîner $h^{(t)}$ sur l'ensemble d'entraînement pondéré par $\textsc{Base}(D_n, \mathbf{w})$
  - calculer $\alpha^{(t)}$ (le poids de $h^{(t)}$)
  - re-pondérer les points (calculer $w_1, \ldots, w_n$)
- Si $h^{(t)}$ ne peut pas gérer les points pondérés:
  - re-échantillonner selon la distribution $w_1, \ldots, w_n$

Boosting

$\textsc{AdaBoost}(D_n, \textsc{Base}(D_n, \mathbf{w}), T)$

$\mathbf{w}^{(1)} \leftarrow (1/n, \ldots, 1/n)$ ▷ *poids initiaux*
pour $t \leftarrow 1$ **à** T
  $h^{(t)} \leftarrow \textsc{Base}(D_n, \mathbf{w})$ ▷ *hypothèse de base*
  $\varepsilon^{(t)} \leftarrow \sum_{h^{(t)}(\mathbf{x}_i) \neq y_i} w_i$ ▷ *erreur pondérée*
  si $\varepsilon^{(t)} \geq 1/2$ **alors**
    retourner $f_{t-1}(\cdot) = \sum_{j=1}^{t-1} \alpha^{(j)} h^{(j)}(\cdot)$
  $\alpha^{(t)} \leftarrow \frac{1}{2} \ln \left(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}} \right)$ ▷ *poids de* $h^{(t)}$
  pour $i \leftarrow 1$ **à** n ▷ *re-pondération des points*
    si $h^{(t)}(\mathbf{x}_i) \neq y_i$ **alors**
      $w_i^{(t+1)} \leftarrow \frac{w_i^{(t)}}{2\varepsilon^{(t)}}$
    sinon
      $w_i^{(t+1)} \leftarrow \frac{w_i^{(t)}}{2(1-\varepsilon^{(t)})}$
retourner $f^{(T)}(\cdot) = \sum_{t=1}^{T} \alpha^{(t)} h^{(t)}(\cdot)$

- Analyse:
  - normalisation: $\sum_{i=1}^{n} w_i = 1$
  - décorrelation de $h^{(t)}$ et l'erreur pondérée: $\sum_{h^{(t)}(\mathbf{x}_i)\neq y_i} w_i^{(t+1)} = \frac{1}{2}$
  - $w_i^{(t+1)} = \frac{w_i^{(t)} e^{-\alpha^{(t)} h^{(t)}(\mathbf{x}_i) y_i}}{2\sqrt{\varepsilon^{(t)}(1-\varepsilon^{(t)})}} = \ldots = \frac{1}{n} \frac{e^{-y_i \sum_{j=1}^{t} \alpha^{(j)} h^{(j)}(\mathbf{x}_i)}}{\prod_{j=1}^{t} 2\sqrt{\varepsilon^{(j)}(1-\varepsilon^{(j)})}} = \frac{e^{-f(\mathbf{x}_i) y_i}}{\sum_{j=1}^{n} e^{-f(\mathbf{x}_j) y_j}}$
  - marge (non-normalisée, fonctionnelle): $\rho_i = f(\mathbf{x}_i) y_i$
  - $\rho_i < 0 \Rightarrow \mathbf{x}_i$ est mal classifié
  - ρ_i est grand $\Rightarrow \mathbf{x}_i$ est bien classifié avec confiance

- Minimisation de la perte exponentielle sur la marge:

$$L_e\big((\mathbf{x},y),f\big) = e^{-f(\mathbf{x})y} = e^{-\rho}$$

  - risque exponentiel empirique:

$$\widehat{R}_e(f) = \frac{1}{n}\sum_{i=1}^{n} L_e\big((\mathbf{x}_i,y_i),f\big) = \frac{1}{n}\sum_{i=1}^{n} e^{-f(\mathbf{x}_i)y_i}$$

  - minimisation gloutonne:

$$h^{(t)}(\cdot), \alpha^{(t)} \leftarrow \underset{h(\cdot),\alpha}{\arg\min}\, \widehat{R}_e\big(f^{(t-1)}(\cdot) + \alpha h(\cdot)\big)$$

- Minimisation de la perte exponentielle sur la marge:
  - étant donné h:

$$\alpha^{(t)} \leftarrow \underset{\alpha}{\arg\min}\, \widehat{R}_e\big(f^{(t-1)}(\cdot) + \alpha h(\cdot)\big) = \frac{1}{2}\ln\left(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}}\right)$$

  - si $\alpha = \frac{1}{2}\ln\left(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}}\right)$:

$$\widehat{R}_e\big(f^{(t-1)}(\cdot) + \alpha h(\cdot)\big) = 2\sqrt{\varepsilon^{(t)}(1-\varepsilon^{(t)})} \prod_{j=1}^{t-1}\left[2\sqrt{\varepsilon^{(j)}(1-\varepsilon^{(j)})}\right]$$

  - $h^{(t)}$ doit minimiser $\varepsilon^{(t)}(1-\varepsilon^{(t)})$ avec $\varepsilon^{(t)} < 1/2 \equiv h^{(t)}$ doit minimiser $\varepsilon^{(t)}$

- Théorème de convergence

$$\begin{aligned}\widehat{R}(f) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{f(\mathbf{x}_i)y_i < 0\} \\ &\leq \widehat{R}_e(f) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i \sum_{t=1}^{T} \alpha^{(t)} h^{(t)}(\mathbf{x}_i)\right) \\ &= \prod_{t=1}^{T}\left[2\sqrt{\varepsilon^{(t)}(1-\varepsilon^{(t)})}\right]\end{aligned}$$

  - soit $\varepsilon^{(t)} \leq \frac{1}{2} - \delta$: $h^{(t)}$ est un peu meilleur qu'une décision aléatoire

$$\widehat{R}(f) \leq e^{-2T\delta^2}$$

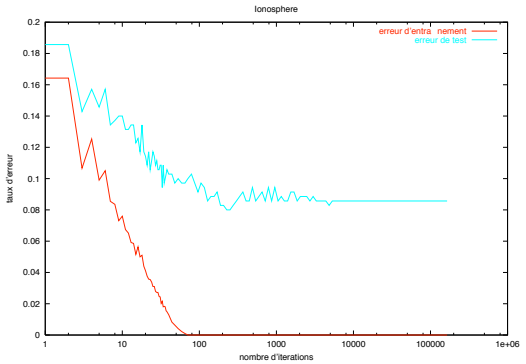
  - $\widehat{R}(f) = 0$ après $\left\lceil \frac{\ln n}{2\delta^2} \right\rceil + 1$ itérations

Boosting

- Expériences
  - decision stumps, 4 données de UCI, test croisé en 10 blocs
  - l'erreur de test diminue même après que l'erreur d'entraînement devienne 0

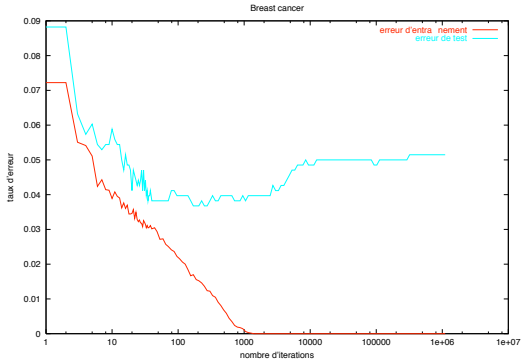
Boosting

- Pas beaucoup de sur-apprentissage



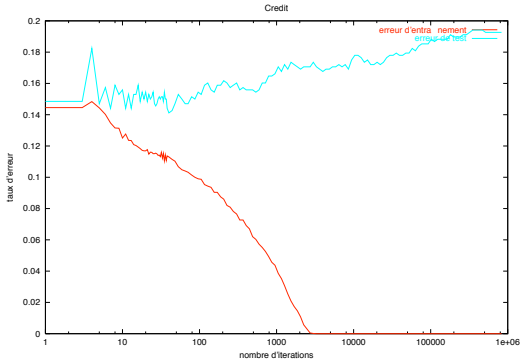
Boosting

- Plus de sur-apprentissage

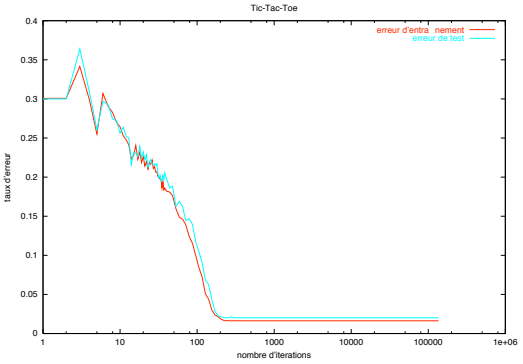


Boosting

- Beaucoup de sur-apprentissage



- L'erreur d'entraînement asymptotique n'est pas 0



- Marge normalisée
  - coefficients normalisés: $\widetilde{\alpha}^{(j)} = \dfrac{\alpha^{(j)}}{\sum_{\ell=1}^{t} \alpha^{(\ell)}}$
  - vote normalisé: $\widetilde{f}(\cdot) = \sum_{j=1}^{t} \widetilde{\alpha}^{(j)} h^{(j)}(\cdot) = \dfrac{\sum_{j=1}^{t} \alpha^{(j)} h^{(j)}(\cdot)}{\sum_{j=1}^{t} \alpha^{(j)}} = \dfrac{f(\cdot)}{\|\boldsymbol{\alpha}\|_1}$
  - marge normalisée (géométrique en L_1): $\widetilde{\rho}_i(\boldsymbol{\alpha}) = y_i \widetilde{f}(\mathbf{x}_i) = \dfrac{y_i f(\mathbf{x}_i)}{\|\boldsymbol{\alpha}\|_1}$
  - marge minimale: $\widetilde{\rho}^*(\boldsymbol{\alpha}) = \min_{i=1,\ldots,n} \widetilde{\rho}_i(\boldsymbol{\alpha})$
  - maximum possible marge minimale: $\widetilde{\rho}^* = \max_{\boldsymbol{\alpha}} \widetilde{\rho}^*(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{i=1,\ldots,n} \widetilde{\rho}_i(\boldsymbol{\alpha})$

- Avantage (edge): $\gamma(\mathbf{w}) = 1 - 2\varepsilon = \sum_{i=1}^{n} w_i h(\mathbf{x}_i) y_i$
  - juste une transformation linéaire de l'erreur pondérée
  - but de $h^{(t)}$: maximiser l'avantage
  - avantage maximal: $\gamma^*(\mathbf{w}^{(t)}) = \max_{h \in H} \gamma(\mathbf{w}^{(t)})$
  - minimum possible avantage maximal: $\gamma^* = \min_{\mathbf{w}} \gamma^*(\mathbf{w}) = \min_{\mathbf{w}} \max_{h \in H} \gamma(\mathbf{w})$
  - $\gamma^*(\mathbf{w}^{(t)}) > 2\delta \;\Rightarrow \widehat{R}(f^{(t)}) = 0$ après $t = \left\lceil \dfrac{\ln n}{2\delta^2} \right\rceil + 1$ itérations

- Théorème min-max de von Neumann [1928]:
  $\gamma^* \geq \widetilde{\rho}^* \quad (\gamma^* = \widetilde{\rho}^* \text{ si } H \text{ est fini})$
  - théorème principal de la théorie des jeux
  - $\gamma^*(\mathbf{w}^{(t)}) \geq \widetilde{\rho}^*$
- Si $\widetilde{\rho}^* > 0$
  - D_n est séparable
  - $\widetilde{\rho}^*$ est la marge L_1 de séparation
  - $2\delta = \widetilde{\rho}^* \;\Rightarrow \widehat{R}(f^{(t)}) = 0$ après $t = \left\lceil \dfrac{2 \ln n}{\widetilde{\rho}^{*2}} \right\rceil + 1$ itérations

Boosting

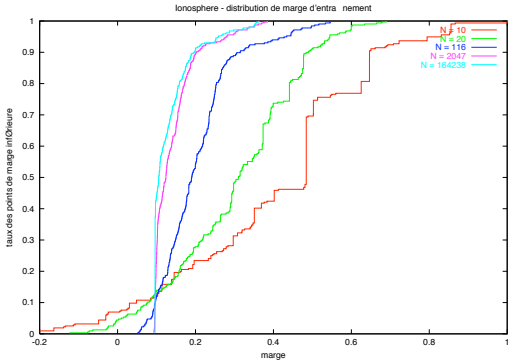
- Comportement asymptotique
  - séparabilité $\Rightarrow$ $\|\alpha^{(t)}\|_1 \to \infty$ $\Rightarrow$ sur-apprentissage?
- Théorème de convergence de l'erreur marginale

$$\widehat{R}^{(\theta)}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{\widetilde{f}(\mathbf{x}_i) y_i < \theta\right\} \leq \widehat{R}_e^{(\theta)}(f) = 2^T \prod_{t=1}^{T} \sqrt{\varepsilon^{(t)\,1-\theta}(1-\varepsilon^{(t)})^{1+\theta}}$$

  - $\varepsilon^{(t)} \leq \frac{1}{2} - \theta - \delta$ $\Rightarrow$ convergence en temps fini
  - $\theta < \frac{\widetilde{\rho}^*}{2}$ $\Rightarrow$ convergence asymptotique
  - la marge minimale asymptotique est $\lim_{t\to\infty} \widetilde{\rho}^*(\boldsymbol{\alpha}^{(t)}) \geq \frac{\widetilde{\rho}^*}{2}$

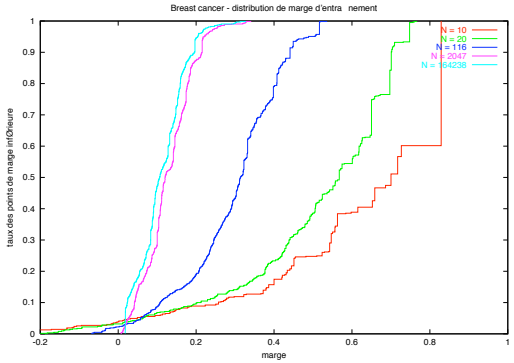
Boosting

- Distribution de marge d'entraînement



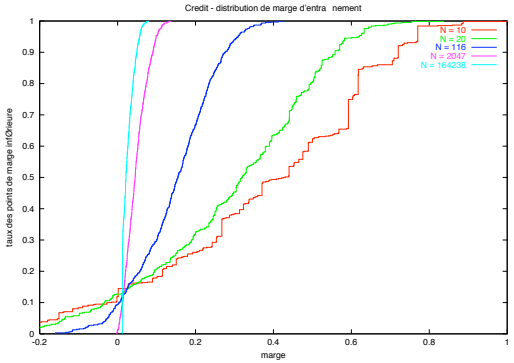
Boosting

- Distribution de marge d'entraînement

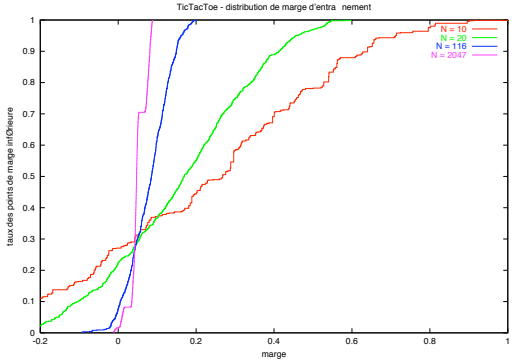


Boosting

- Distribution de marge d'entraînement



- Distribution de marge d'entraînement



- Maximisation de la marge: $\lim_{t\to\infty} \widetilde{\rho}^*(\boldsymbol{\alpha}^{(t)}) = \widetilde{\rho}^*$?
  - $h^{(1)}, \ldots, h^{(t)}$ donnés: programmation linéaire
  - boosting marginal: minimiser l'erreur marginale $\widehat{R}_e^{(\theta)}(f)$
  - choisir $\theta = \theta^{(t)}$ adaptivement

ADABOOST$_\theta(D_n, \text{BASE}(D_n, \mathbf{w}), T, \theta)$

1 $\mathbf{w}^{(1)} \leftarrow (1/n, \ldots, 1/n)$ ▷ *poids initiaux*
2 **pour** $t \leftarrow 1$ **à** T
3 $\quad h^{(t)} \leftarrow \text{BASE}(D_n, \mathbf{w})$ ▷ *hypothèse de base*
4 $\quad \varepsilon^{(t)} \leftarrow \sum_{h^{(t)}(\mathbf{x}_i)\neq y_i} w_i$ ▷ *erreur pondérée*
5 $\quad$ **si** $\varepsilon^{(t)} \geq 1/2 - \theta/2$ **alors**
6 $\quad\quad$ **retourner** $f_{t-1}(\cdot) = \sum_{j=1}^{t-1} \alpha^{(j)} h^{(j)}(\cdot)$
7 $\quad \alpha^{(t)} \leftarrow \frac{1}{2}\ln\left(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}}\right) - \frac{1}{2}\ln\left(\frac{1+\theta}{1-\theta}\right)$ ▷ *poids de* $h^{(t)}$
8 $\quad$ **pour** $i \leftarrow 1$ **à** n ▷ *re-pondération des points*
9 $\quad\quad$ **si** $h^{(t)}(\mathbf{x}_i) \neq y_i$ **alors**
10 $\quad\quad\quad w_i^{(t+1)} \leftarrow w_i^{(t)} \frac{1-\theta}{2\varepsilon^{(t)}}$
11 $\quad\quad$ **sinon**
12 $\quad\quad\quad w_i^{(t+1)} \leftarrow w_i^{(t)} \frac{1+\theta}{2(1-\varepsilon^{(t)})}$
13 **retourner** $f^{(T)}(\cdot) = \sum_{t=1}^{T} \alpha^{(t)} h^{(t)}(\cdot)$

- θ constant
  - weight decay: $\alpha^{(t)}$ est plus petit qu'avant
  - minimisation de la perte exponentielle pénalisée (L_1):

$$L_e^{(\theta)}\big((\mathbf{x}, y), f\big) = e^{-f(\mathbf{x})y + \theta\|\alpha\|_1}$$

  - en pratique, la validation de θ (avec $T \to \infty$) et la validation de T (avec $\theta = 0$) produisent les résultats similaires

Boosting

- θ adaptatif
  - arc-gv: $\theta^{(t)} = \widetilde{\rho}^*(\boldsymbol{\alpha}^{(t)})$: convergence asymptotique $\lim_{t\to\infty} \widetilde{\rho}^*(\boldsymbol{\alpha}^{(t)}) = \widetilde{\rho}^*$
  - AdaBoost$^*(\nu)$: $\theta^{(t)} = \min_{j=1,\ldots,t} \gamma^{(j)} - \nu$: convergence en temps fini:

    $$\rho^*(\boldsymbol{\alpha}^{(t)}) > \widetilde{\rho}^* - \nu$$

    après $\left\lceil \frac{2\ln n}{\nu^2} \right\rceil + 1$ itérations

- Observation en pratique
  - la maximisation agressive de la marge minimale empire la généralisation!!!
  - pas d'explication théorique