Boosting

- Extensions
  - autres fonctions de perte
  - fonctions de base non-binaires
  - multiclasse, régression
- Applications
  - arbres de décisions alternants
  - apprentissage des règles logiques
  - traitement d'images
  - traitement de langue

Boosting

- 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:

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

Boosting

- 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 $\varepsilon^{(t)}$

Boosting

- Minimisation d'une perte arbitraire sur la marge:

$$L\big((\mathbf{x},y),f\big) = L\big(f(\mathbf{x})y\big) = L(\rho)$$

  - risque empirique:

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

  - minimisation gloutonne:

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

  - la minimisation analytique n'est pas possible en général

Boosting

- Descente de gradient dans l'espace fonctionnel

$$\begin{aligned} h^{(t)}(\cdot) &= \arg\max_h -\left.\frac{\partial \widehat{R}\left(f^{(t-1)}(\cdot)+\alpha h(\cdot)\right)}{\partial \alpha}\right|_{\alpha=0} \\ &= \arg\max_h -\frac{1}{n}\sum_{i=1}^{n} L'\left(f^{(t-1)}(\mathbf{x}_i)y_i\right)\cdot h(\mathbf{x}_i)y_i \end{aligned}$$

- $w_i^{(t)} = \dfrac{-L'\left(f^{(t-1)}(\mathbf{x}_i)y_i\right)}{-\sum_{j=1}^{n} L'\left(f^{(t-1)}(\mathbf{x}_j)y_j\right)} \Longrightarrow h^{(t)}(\cdot) = \arg\max_h \sum_{i=1}^{n} w_i^{(t)} h(\mathbf{x}_i)y_i$
- minimisation scalaire:

$$\alpha^{(t)} = \arg\min_\alpha \widehat{R}\left(f^{(t-1)}(\cdot)+\alpha h^{(t)}(\cdot)\right) = \arg\min_\alpha \sum_{i=1}^{n} L\left(f^{(t-1)}(\mathbf{x}_i)y_i + \alpha h^{(t)}(\mathbf{x}_i)y_i\right)$$

Boosting

- Exemples

Algorithme	fonction de perte $L(\rho)$	$\alpha^{(t)}$
AdaBoost	$e^{-\rho}$	line search (analytique)
ConfBoost	$e^{-\rho}$	line search
ARC-X4	$(1-\rho)^5$	$1/t$
LogitBoost	$\ln(1+e^{-\rho})$	Newton-Raphson

- LogitBoost
  - interprétation statistique de boosting, connection à la régression logistique
  - en pratique: plus de tolérance de bruit

Boosting

- AdaBoost avec abstention
  - $h : \mathbb{R}^d \mapsto \{-1, 0, 1\}$
  - $h^{(t)}(\cdot) = \arg\min_{h:\varepsilon_+ > \varepsilon_-} \left(\varepsilon_0 + 2\sqrt{\varepsilon_+ \varepsilon_-}\right)$
  - $\alpha^{(t)} = \dfrac{1}{2}\ln\dfrac{\varepsilon_+^{(t)}}{\varepsilon_-^{(t)}}$
- Problème si $\varepsilon_-^{(t)} = 0$
  - solution 1: $\alpha^{(t)} = \dfrac{1}{2}\ln\dfrac{\varepsilon_+^{(t)}+\delta}{\varepsilon_-^{(t)}+\delta}$
  - solution 2: utilise $\theta > 0$ dans AdaBoost$_\theta$

Boosting

- ConfBoost
  - $h : \mathbb{R}^d \mapsto \mathbb{R}$
  - $h^{(t)}(\cdot) = \arg\max_h \sum_{i=1}^{n} w_i^{(t)} h(\mathbf{x}_i)y_i$
  - pas de solution analytique pour $\alpha^{(t)}$: line search

ConfBoost$(D_n, \text{Base}(D_n, \mathbf{w}), T)$

$\mathbf{w}^{(1)} \leftarrow (1/n, \ldots, 1/n)$ ▷ *poids initiaux*

pour $t \leftarrow 1$ **à** T

  $h^{(t)} \leftarrow \text{Base}(D_n, \mathbf{w})$ ▷ $\leftarrow \arg\max_h \sum_{i=1}^n w_i^{(t)} h(\mathbf{x}_i) y_i$

  $\alpha^{(t)} \leftarrow \arg\min_\alpha \sum_{i=1}^n w_i^{(t)} e^{-\alpha h^{(t)}(\mathbf{x}_i) y_i}$ ▷ *poids de* $h^{(t)}$

  si $\alpha^{(t)} = \infty$ **alors** ▷ *"overfitting":* $h(\mathbf{x}_i) y_i > 0 \ \forall i = 1, \ldots, n$

    retourner $h^{(t)}(\cdot)$

  si $\alpha^{(t)} < 0$ **alors** ▷ *"underfitting":* $\gamma^{(t)} = \sum_{i=1}^n w_i^{(t)} h^{(t)}(\mathbf{x}_i) y_i < 0$

    retourner $f^{(t-1)}(\cdot) = \sum_{j=1}^{t-1} \alpha^{(j)} h^{(j)}(\cdot)$

  pour $i \leftarrow 1$ **à** n ▷ *re-pondération des points*

    $w_i^{(t+1)} \leftarrow w_i^{(t)} \dfrac{e^{-\alpha^{(t)} h^{(t)}(\mathbf{x}_i) y_i}}{\sum_{j=1}^n w_j^{(t)} e^{-\alpha^{(t)} h^{(t)}(\mathbf{x}_j) y_j}}$

retourner $f^{(T)}(\cdot) = \sum_{t=1}^T \alpha^{(t)} h^{(t)}(\cdot)$

Boosting

- Notations
  - avantages (marges) de base: $\gamma_i = h(\mathbf{x}_i) y_i$
  - avantage: $\gamma = \sum_{i=1}^n w_i \gamma_i$
  - paramètre de relaxation/régularisation: θ
  - $E^{(t)}(\alpha) = e^{\theta\alpha} \sum_{i=1}^n w_i^{(t)} e^{-\alpha \gamma_i^{(t)}} = \sum_{i=1}^n w_i^{(t)} e^{-\alpha(\gamma_i^{(t)} - \theta)}$
- Théorème de convergence de l'erreur marginale

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

ConfBoost$_\theta(D_n, \text{Base}(D_n, \mathbf{w}), T, \theta)$

$\mathbf{w}^{(1)} \leftarrow (1/n, \ldots, 1/n)$ ▷ *poids initiaux*

pour $t \leftarrow 1$ **à** T

  $h^{(t)} \leftarrow \text{Base}(D_n, \mathbf{w})$ ▷ $\leftarrow \arg\max_h \sum_{i=1}^n w_i^{(t)} h(\mathbf{x}_i) y_i$

  pour $i \leftarrow 1$ **à** n ▷ *avantages de base*

    $\gamma_i^{(t)} \leftarrow h^{(t)}(\mathbf{x}_i) y_i$

  $\alpha^{(t)} \leftarrow \arg\min_\alpha e^{\theta\alpha} \sum_{i=1}^n w_i^{(t)} e^{-\alpha \gamma_i^{(t)}} = \arg\min_\alpha E^{(t)}(\alpha)$ ▷ *poids de* $h^{(t)}$

  si $\alpha^{(t)} = \infty$ **alors** ▷ *"overfitting":* $\gamma_i^{(t)} > \theta \ \forall i = 1, \ldots, n$

    retourner $h^{(t)}(\cdot)$

  si $\alpha^{(t)} < 0$ **alors** ▷ *"underfitting":* $\gamma^{(t)} < \theta$

    retourner $f^{(t-1)}(\cdot) = \sum_{j=1}^{t-1} \alpha^{(j)} h^{(j)}(\cdot)$

  pour $i \leftarrow 1$ **à** n ▷ *re-pondération des points*

    $w_i^{(t+1)} \leftarrow w_i^{(t)} \dfrac{e^{-\alpha^{(t)} \gamma_i^{(t)}}}{\sum_{j=1}^n w_j^{(t)} e^{-\alpha^{(t)} \gamma_i^{(t)}}}$

retourner $f^{(T)}(\cdot) = \sum_{t=1}^T \alpha^{(t)} h^{(t)}(\cdot)$

Boosting

- Multiclasse (K classes)
  - étiquettes K-dimensionnelles
  - en général, une observation peut appartenir à plusieurs classes (classification "multiétiquettes")
  - cas spécial: "uniétiquettes"

$$y_\ell = \begin{cases} 1 & \text{si } \ell \text{ est la classe correcte } \ell(\mathbf{x}), \\ -1 & \text{sinon} \end{cases}$$

Boosting

- Distribution des poids $w_{i,\ell}$
  - $w_{i,\ell}$ exprime les difficultés de classifier $\mathbf{x}_i$ à sa classe correcte ($\ell = \ell(\mathbf{x}_i)$) ou contre ses classes incorrectes ($\ell \neq \ell(\mathbf{x}_i)$)
  - peuvent être initialisés à

$$w_{i,\ell}^{(1)} = \begin{cases} \frac{1}{2n} & \text{si } \ell \text{ est la classe correcte (si } y_{i,\ell} = 1), \\ \frac{1}{2n(K-1)} & \text{sinon (si } y_{i,\ell} = -1) \end{cases}$$

Boosting

- Classifieurs faibles:

$$\mathbf{h} : X \to \mathbb{R}^K;\ h_\ell : X \to \mathbb{R}$$

  - peuvent voter aux plusieurs classes

- Classifieur final:

$$\mathbf{f}^{(T)}(\mathbf{x}) = \sum_{t=1}^{T} \alpha^{(t)} \mathbf{h}^{(t)}(\mathbf{x})$$

  - la meilleure classe (cas “uniétiquettes”):

$$\widehat{\ell}(\mathbf{x}) = \arg\max_{\ell} f_\ell^{(T)}(\mathbf{x})$$

Boosting

- Avantages de base
  - nombre finis de classifieurs faibles $\mathbf{h}_1, \ldots, \mathbf{h}_N$

$$\gamma_{j,i,\ell} = h_{j,\ell}(\mathbf{x}_i) y_{i,\ell}$$

  - par exemple, si $\mathbf{h} : X \to \{-1, 0, +1\}^K$:

$$\gamma_{j,i,\ell} = \begin{cases} 1 & \text{si } h_{j,\ell}(\mathbf{x}_i) = y_{i,\ell} \\ -1 & \text{si } h_{j,\ell}(\mathbf{x}_i) \neq y_{i,\ell} \\ 0 & \text{si } h_{j,\ell}(\mathbf{x}_i) = 0 \end{cases}$$

Boosting

- Fonction d’énergie:

$$E^{(t)}(\mathbf{h}_j, \alpha) = e^{\theta\alpha} \sum_{i=1}^{n} \sum_{\ell=1}^{K} w_{i,\ell}^{(t)} \exp\left(-\alpha \gamma_{j,i,\ell}\right)$$

  - dans chaque itération, sélectionne $\mathbf{h}^{(t)} = \mathbf{h}_{j^{(t)}}$ qui maximise

$$\gamma_j^{(t)} = \sum_{i=1}^{n} \sum_{\ell=1}^{K} w_{i,\ell}^{(t)} \gamma_{j,i,\ell} = \theta + \left. \frac{-\partial E^{(t)}(\mathbf{h}_j, \alpha)}{\partial \alpha} \right|_{\alpha=0}$$

  - coefficient de $\mathbf{h}^{(t)}$:

$$\alpha^{(t)} = \arg\min_{\alpha} E^{(t)}\left(\mathbf{h}^{(t)}, \alpha\right)$$

- Mise à jour des poids:

$$w_{i,\ell}^{(t+1)} \leftarrow w_{i,\ell}^{(t)} \frac{\exp\left(-\alpha^{(t)}\gamma_{j^{(t)},i,\ell}\right)}{Z^{(t)}},$$

ou

$$Z^{(t)} = \sum_{i=1}^{n}\sum_{\ell=1}^{K} w_{i,\ell}^{(t)} \exp\left(-\alpha^{(t)}\gamma_{j^{(t)},i,\ell}\right)$$

- Marges par classe:

$$\rho_{i,\ell}^{(t)} = \sum_{\tau=1}^{t} \alpha^{(\tau)}\gamma_{j^{(\tau)},i,\ell} = f_\ell^{(t)}(\mathbf{x}_i) y_{i,\ell}$$

  - connection entre les poids et les marges

$$w_{i,\ell}^{(t+1)} = w_{i,\ell}^{(1)} \prod_{\tau=1}^{t} \frac{\exp\left(-\alpha^{(\tau)}\gamma_{j^{(\tau)},i,\ell}\right)}{Z^{(\tau)}} = w_{i,\ell}^{(1)} \frac{\exp\left(-\rho_{i,\ell}^{(t)}\right)}{\prod_{\tau=1}^{t} Z^{(\tau)}}$$

  - marges normalisées:

$$\widetilde{\rho}_{i,\ell}^{(t)} = \sum_{\tau=1}^{t} \widetilde{\alpha}^{(\tau)}\gamma_{j^{(\tau)},i,\ell} = \widetilde{f}_\ell^{(t)}(\mathbf{x}_i) y_{i,\ell}.$$

- Erreur marginale:

$$\widehat{R}_n^{(\theta)}\left(\mathbf{f}^{(T)}, \mathbf{w}^{(1)}\right) = \sum_{i=1}^{n}\sum_{\ell=1}^{K} w_{i,\ell}^{(1)} \mathbb{I}\left\{\widetilde{\rho}_{i,\ell}^{(T)} \leq \theta\right\}$$

  - si $\forall t : E^{(t)}\left(\mathbf{h}^{(t)}, \alpha^{(t)}\right) < 1-\delta$, alors $\widehat{R}_n^{(\theta)}\left(\mathbf{f}^{(T)}, \mathbf{w}^{(1)}\right) = 0$ après

$$T^* = \left\lceil \left(\delta \ln \left(\min_{i,\ell: w_{i,\ell}^{(1)} \neq 0} w_{i,\ell}^{(1)} \right) \right)^{-1} \right\rceil + 1$$

  - en particulier, si $w_{i,\ell}^{(1)}$ sont initialisés comme ci-dessus:

$$T^* = \left\lceil \frac{\ln\left(2n(K-1)\right)}{\delta} \right\rceil + 1.$$

- Erreur de classification:

$$\widehat{R}\left(\mathbf{f}^{(T)}\right) = \sum_{i=1}^{n} \mathbb{I}\left\{\widehat{\ell}(\mathbf{x}_i) \neq \ell(\mathbf{x}_i)\right\}$$

- $\widehat{R}_n^{(\theta)}\left(\mathbf{f}^{(T)}, \mathbf{w}^{(1)}\right) = 0 \quad \Longrightarrow \quad \widehat{R}\left(\mathbf{f}^{(T)}\right) = 0$:

$$\forall \ell \neq \ell(\mathbf{x}_i) : \widetilde{f}_{i,\ell}^{(T)}(\mathbf{x}_i) = -\widetilde{\rho}_{i,\ell}^{(T)} < -\theta \leq 0 \leq \theta < \widetilde{\rho}_{i,\ell(\mathbf{x}_i)}^{(T)} = \widetilde{f}_{i,\ell(\mathbf{x}_i)}^{(T)}(\mathbf{x}_i).$$

- $\widehat{R}\left(\mathbf{f}^{(T)}\right) = 0 \quad \not\Longrightarrow \quad \widehat{R}_n^{(\theta)}\left(\mathbf{f}^{(T)}, \mathbf{w}^{(1)}\right) = 0$

- Classifieurs faibles scalaires + vecteurs de votes

$$\mathbf{h} = \varphi\mathbf{v}$$

  - classifieurs faibles $\varphi : X \to \{-1, 1\}$
  - vecteurs de votes: $\mathbf{v} : X \to \{-1, 0, 1\}^K$

- Stratégies
  - "one-against-all": $\mathbf{v} = \{-1, -1, \ldots, -1, 1, -1, \ldots, -1, -1\}$
  - "one-against-one": $\mathbf{v} = \{0, 0, \ldots, 0, -1, 0, \ldots, 0, 1, 0, \ldots, 0, 0\}$
  - ECOC
  - optimiser $\mathbf{v}^{(t)}$ (?)

- Arbres de décisions alternants
  - noeuds de prédiction et noeuds de découpage
  - chaque noeud de prédiction contient un nombre $\in [-1, 1]$
  - la prédiction finale est la signe de la somme des prédictions au long des chemins valides
  - utiliser ADABOOST avec abstention

- Apprentissage des règles logiques: SLIPPER
  - classifieurs de base: des conjonctions positives et une règle de défaut négative
  - construire une conjonction d'une façon gloutonne en maximisant $\sqrt{\varepsilon_+} - \sqrt{\varepsilon_-}$ sur un ensemble GROWSET
  - élaguer les règles sur un autre ensemble PRUNESET
  - utiliser ADABOOST avec abstention

- Traitement d'images: détection d'objets
  - classifieurs de base: des masques rectangulaires simples
  - l'astuce de l'image intégral