

Méthodes d'ensemble :

boosting, bagging, random forests

...



Méthodes d'ensemble



Antoine Cornuéjols

AgroParisTech

Boosting 2

Motivation

■ « *The wisdom of crowd* »

[James Surowiecki, 2004]

– Estimation du poids d'un panier dans un marché

787 participants

- Le meilleur = plus d'un centième d'erreur
- Moyenne : moins d'un millième d'erreur

[Francis Galton¹, 1906 (85 ans)]



Illustration

Example: "How May I Help You?"

[Gorin et al.]

- **goal:** automatically categorize type of call requested by phone customer (*Collect*, *CallingCard*, *PersonToPerson*, etc.)
 - yes I'd like to place a collect call long distance please (*Collect*)
 - operator I need to make a call but I need to bill it to my office (*ThirdNumber*)
 - yes I'd like to place a call on my master card please (*CallingCard*)
 - I just called a number in sioux city and I musta rang the wrong number because I got the wrong party and I would like to have that taken off of my bill (*BillingCredit*)

¹ anthropologue, explorateur, géographe, inventeur, météorologue, proto-généticien, psychométricien et statisticien

Illustration

Example: "How May I Help You?"

[Gorin et al.]

- goal: automatically categorize type of call requested by phone customer (`Collect`, `CallingCard`, `PersonToPerson`, etc.)
 - yes I'd like to place a collect call long distance please (`Collect`)
 - operator I need to make a call but I need to bill it to my office (`ThirdNumber`)
 - yes I'd like to place a call on my master card please (`CallingCard`)
 - I just called a number in sioux city and I musta rang the wrong number because I got the wrong party and I would like to have that taken off of my bill (`BillingCredit`)
- observation:
 - easy to find "rules of thumb" that are "often" correct
 - e.g.: "IF 'card' occurs in utterance
THEN predict 'CallingCard'"
 - hard to find single highly accurate prediction rule

Boosting 5

Types d'experts

- Un seul expert sur l'ensemble de X
- Un expert par sous-régions de X (e.g. arbres de décisions)
- Plusieurs experts, tous sur l'ensemble de X
- Plusieurs experts spécialisés sur des sous-régions de X

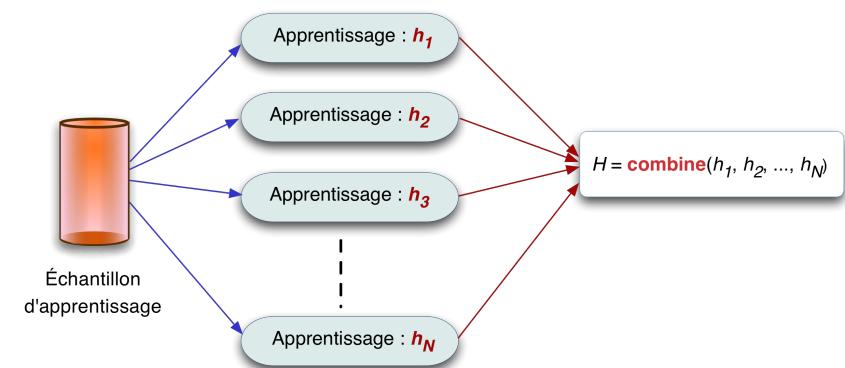
Boosting 6

Types d'experts

- Un seul expert sur l'ensemble de X
- Un expert par sous-régions de X (e.g. arbres de décisions)
- ■ Plusieurs experts, tous sur l'ensemble de X
- Plusieurs experts spécialisés sur des sous-régions de X

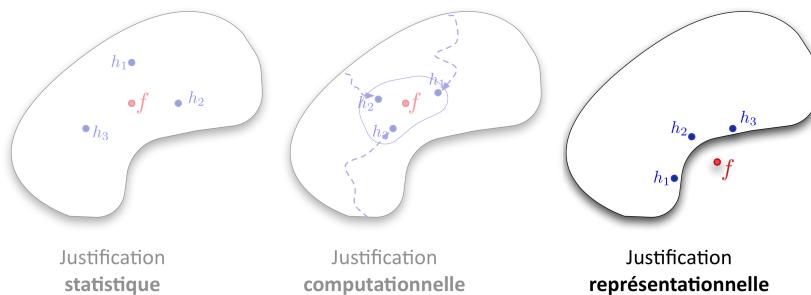
Boosting 7

Schéma général : apprentissage



Boosting 8

Justifications intuitives



- Dietterich T. (2000) « Ensemble Methods in Machine Learning ». Proc. 1st Int. Workshop on Multiple Classifier Systems, Sardinia, Italy, 2000.

Boosting 9

Succès applicatifs

KDD-Cup

- Network intrusion detection (1999) ; molecular bioactivity & protein locale prediction (2001) ; (...) pulmonary embolisme detection (2006) ; customer relationship management (2009) ; educational data mining (2010) ; music recommandation (2011) ; ...
- Tous les 1^{er} prix et 2^{ème} prix pour les méthodes d'ensemble (2009-2011 - ???)

Netflix prize

- Improve accuracy about how much someone is going to enjoy a movie based on their preference
- 1 000 000 \$ for a new algorithm improving on Netflix's one by more than 10%
- Winner in 2009 for an ensemble method

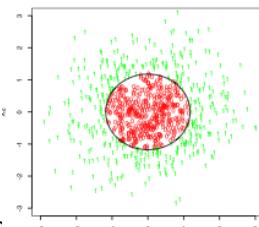
Boosting 10

Illustration

- Soit \mathcal{X} un espace d'entrée à **10 dimensions**
- Les attributs sont indépendants et de distribution gaussienne
- Le concept cible** est défini par :

$$u = \begin{cases} 1 & \text{si } \sum_{j=1,10} x_j^2 > \chi^2_{10}(0,5) \\ -1 & \text{sinon} \end{cases}$$

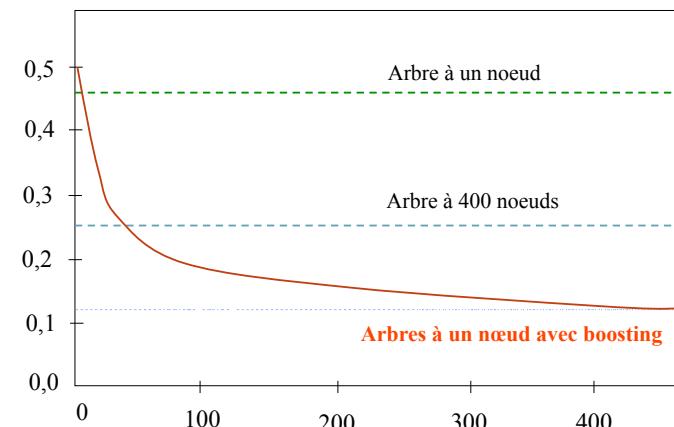
avec : $\chi^2_{10}(0,5) = 9,34$



- 2000 exemples d'apprentissage (1000+;1C)
- 10000 exemples de test
- Apprentissage d'arbres de décision

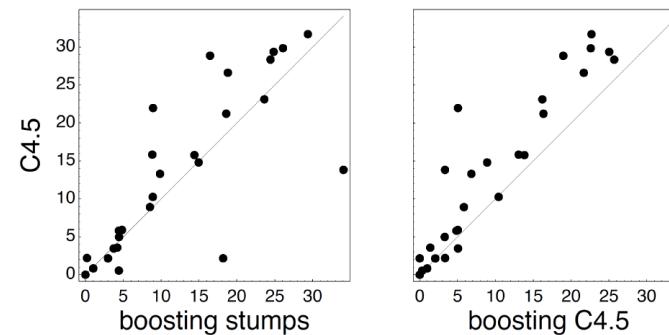
Boosting 11

Illustration



Boosting 12

Performances du boosting



Boosting 13

Exemple simple



- Quel est le meilleur séparateur linéaire ?

Boosting 14

Exemple simple



- Taux d'erreur = 5/20 = 0.25

Exemple simple



- Taux d'erreur (h_1) = 5/20 = 0.25

Et si je pouvais combiner avec un autre séparateur linéaire ?

Ou même plusieurs autres !

Boosting 15

Boosting 16

Exemple simple



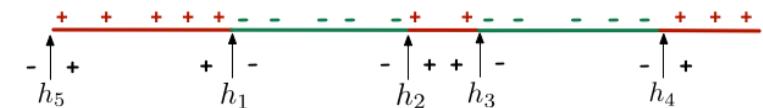
Et si je pouvais combiner avec un autre séparateur linéaire ? Ou même plusieurs autres !

Par exemple en utilisant un **vote pondéré** :

$$H(\mathbf{x}) = \text{sign} \left\{ \sum_{i=1}^l \alpha_i h_i(\mathbf{x}) \right\}$$

Boosting 17

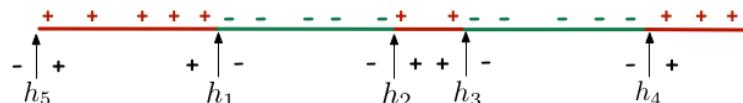
Exemple simple



$$H(\mathbf{x}) = \text{sign} \{ 0.549 h_1(\mathbf{x}) + 0.347 h_2(\mathbf{x}) + 0.310 h_3(\mathbf{x}) + 0.406 h_4(\mathbf{x}) + 0.503 h_5(\mathbf{x}) \}$$

Boosting 18

Exemple simple



$$H(\mathbf{x}) = \text{sign} \{ 0.549 h_1(\mathbf{x}) + 0.347 h_2(\mathbf{x}) + 0.310 h_3(\mathbf{x}) + 0.406 h_4(\mathbf{x}) + 0.503 h_5(\mathbf{x}) \}$$

Le boosting

- Comment arriver à ce genre de combinaison ?

Algorithme du boosting

Boosting 19

Boosting 20

Une question théorique

■ Apprentissage « **fort** » (PAC learning)

- Une classe de fonctions \mathcal{F} est **apprenable** (au sens **fort**) si il existe un algorithme d'apprentissage \mathcal{A} qui pour toute distribution \mathcal{D}_X sur X , et pour toute fonction f est tel que :

$$\forall \varepsilon, \delta : \exists m(\varepsilon, \delta) \text{ st. } \text{Prob}[R(h_{\mathcal{S}}) > \varepsilon] \leq \delta$$

■ Apprentissage « **faible** »

- Une classe de fonctions \mathcal{F} est **apprenable** (au sens **faible**) si, pour $\gamma > 0$, il existe un algorithme d'apprentissage \mathcal{A} qui pour toute distribution \mathcal{D}_X sur X , et pour toute fonction f est tel que :

$$\forall \delta : \exists m(\delta) \text{ st. } \text{Prob}[R(h_{\mathcal{S}}) > 1/2 - \gamma] \leq \delta$$

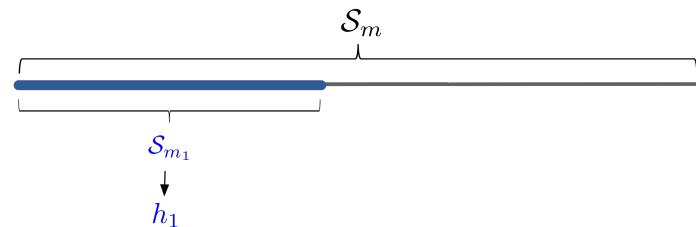
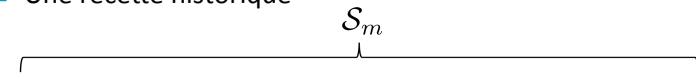
■ Sont-ils de nature différente ?

Boosting 21

Boosting 22

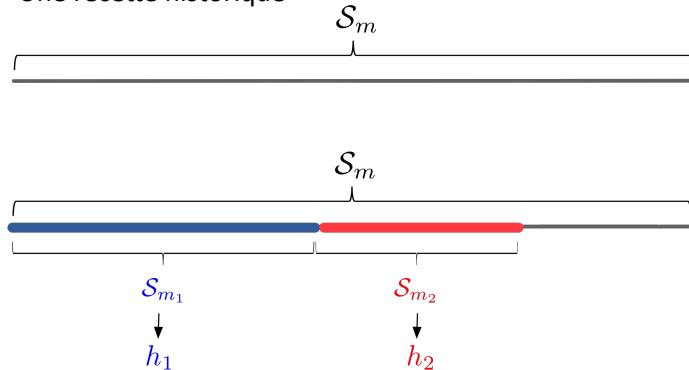
Comment engendrer les apprenants

■ Une recette historique



Comment engendrer les apprenants

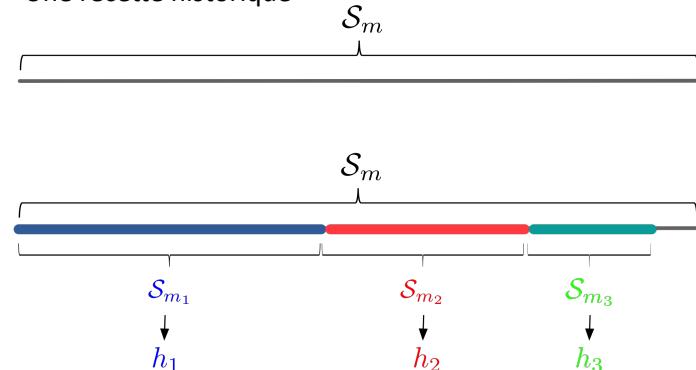
■ Une recette historique



Boosting 23

Comment engendrer les apprenants

■ Une recette historique



$$H(\mathbf{x}) = \text{sign}\left(h_1(\mathbf{x}) + h_2(\mathbf{x}) + h_3(\mathbf{x}) \right)$$

Boosting 24

Questions

- Comment engendrer des **apprenants faibles décorrélés** ?
- Comment **combiner** leurs prédictions ?

Comment engendrer les apprenants (suite)

- Modifier l'échantillon d'apprentissage à chaque étape
 - En **diminuant** l'importance des exemples **bien classés**
 - En **augmentant** ----- **mal** -----
 - De combien ?

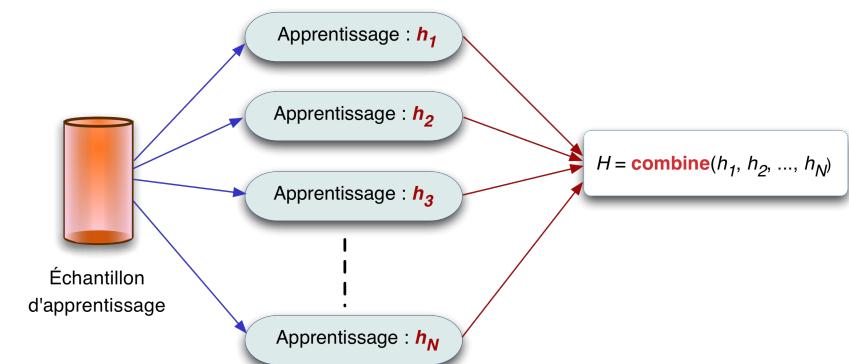
Boosting 25

Boosting 26

Boosting

- **boosting** = méthode générale pour convertir des règles de prédiction peu performantes en une règle de prédiction (très) performante
- Plus précisément :
 - Étant donné un algorithme d'apprentissage “faible” qui peut toujours retourner une hypothèse de taux d'erreur $\leq 1/2 - \gamma$
 - Un algorithme de boosting peut construire (de manière prouvée) une règle de décision (hypothèse) de taux d'erreur $\leq \epsilon$

Schéma général : apprentissage



Boosting 27

Boosting 28

Questions

- Comment choisir les courses à chaque étape?
 - ▶ Se concentrer sur les courses les plus “difficiles”
(celles sur lesquelles les heuristiques précédentes sont les moins performantes)

- Comment combiner les heuristiques (règles de prédiction) en une seule règle de prédiction ?
 - ▶ Prendre une vote (pondéré) majoritaire de ces règles

Boosting 29

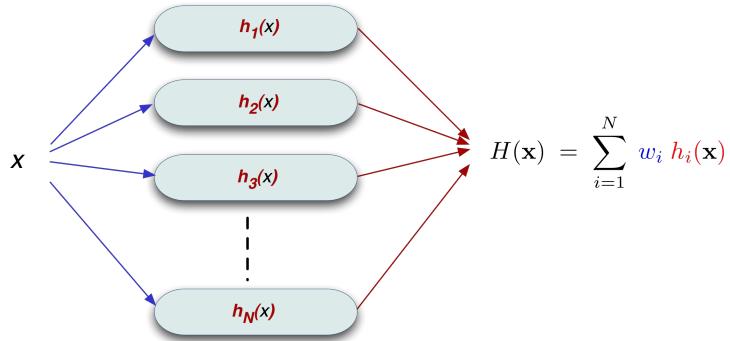
Boosting : vue formelle

- Étant donné l'échantillon d'apprentissage $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$
 - $y_i \in \{-1, +1\}$ étiquette de l'exemple $x_i \in S$

 - Pour $t = 1, \dots, T$:
 - Construire la distribution D_t sur $\{1, \dots, m\}$
 - Trouver l'hypothèse faible (“heuristique”)
 - $h_t : S \rightarrow \{-1, +1\}$
 - avec erreur petite ε_t sur D_t :
- $$\varepsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$$
- Retourner l'hypothèse finale h_{final}

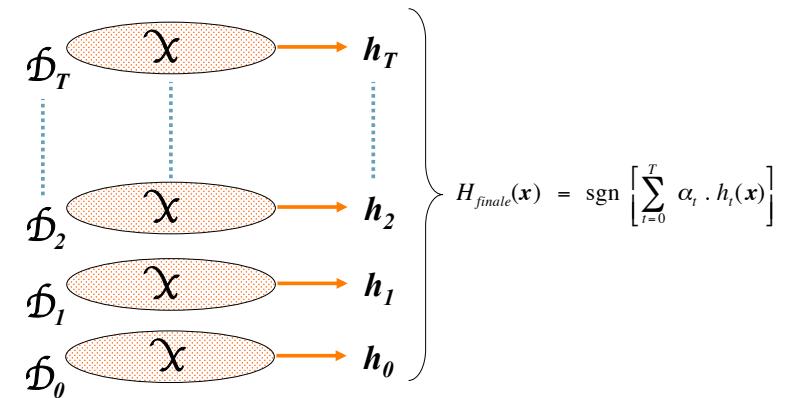
Boosting 30

Schéma général : *prédiction*



Boosting 31

Le principe général



- Comment passer de \mathcal{D}_t à \mathcal{D}_{t+1} ?
- Comment calculer la pondération α_t ?

Boosting 32

AdaBoost [Freund&Schapire '97]

- construire D_t : $D_t(i) = \frac{1}{m}$

Étant donnée D_t et h_t :

$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$= \frac{D_t}{Z_t} \cdot \exp(-\alpha_t \cdot y_i \cdot h_t(x_i))$$

où: Z_t = constante de normalisation

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right) > 0$$

- Hypothèse finale :

$$H_{\text{final}}(x) = \operatorname{sgn} \left(\sum_t \alpha_t h_t(x) \right)$$

AdaBoost en plus gros

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right) > 0$$

$$D_{t+1} = \frac{D_t}{Z_t} \cdot \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$H_{\text{final}}(x) = \operatorname{sgn} \left(\sum_t \alpha_t h_t(x) \right)$$

Boosting 33

Boosting 34

Exemple simple



- Taux d'erreur = $5/20 = 0.25$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon}{\varepsilon} = \frac{1}{2} \ln \frac{0.75}{0.25} = 0.549$$

Boosting 35

Exemple simple

- Nouvelle pondération des exemples d'apprentissage



$$\text{Exemples bien classés} \quad p_b(x) = \frac{e^{-\alpha}}{Z} = \frac{e^{-0.549}}{Z} = \frac{0.577}{Z}$$

$$\text{Exemples mal classés} \quad p_m(x) = \frac{e^{\alpha}}{Z} = \frac{e^{0.549}}{Z} = \frac{1.732}{Z}$$

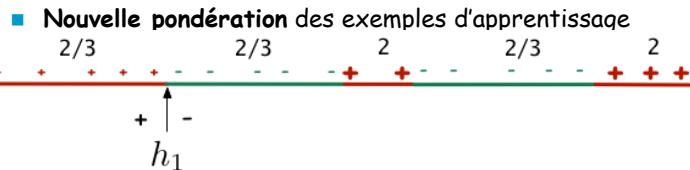
$$Z = \frac{(15 \times 0.577) + (5 \times 1.732)}{20} = \frac{8.660 + 8.660}{20} = \frac{17.32}{20} = 0.866$$

$$p_b(x) = 0.666 = 2/3$$

$$p_m(x) = 2$$

Boosting 36

Exemple simple



- Exemples bien classés**

$$p_b(x) = \frac{1}{2(1-\varepsilon)} = \frac{1}{2 \times 0.75} = \frac{1}{1.5} = \frac{2}{3}$$

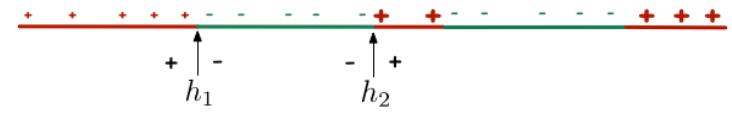
- Exemples mal classés**

$$p_m(x) = \frac{1}{2\varepsilon} = \frac{1}{2 \times 0.25} = \frac{1}{0.5} = 2$$

$$Z = 2\varepsilon^{1/2} (1-\varepsilon)^{1/2}$$

Boosting 37

Exemple simple



- Taux d'erreur :** $\varepsilon_2 = \frac{10 \times 2/3}{20} = \frac{1}{3}$

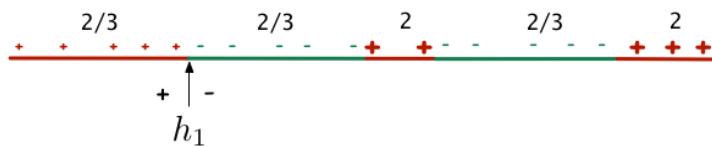
$$\alpha_2 = \frac{1}{2} \ln \frac{1 - \varepsilon_2}{\varepsilon_2} = \frac{1}{2} \ln \frac{2/3}{1/3} = 0.347$$

- Sous-pondération des bien classés :** $p_b(x) = \frac{1}{2(1-\varepsilon)} = \frac{1}{2 \times 2/3} = \frac{3}{4} = 0.75$

- Sur-pondération des mal classés :** $p_m(x) = \frac{1}{2\varepsilon} = \frac{1}{2 \times 1/3} = \frac{3}{2} = 1.5$

Boosting 38

Exemple simple

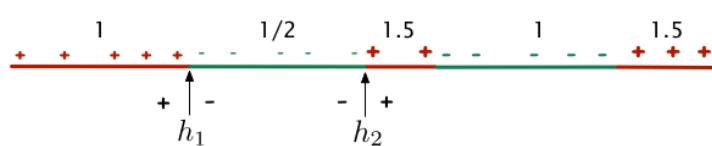


- Sous-pondération des bien classés :**

$$p_b(x) = \frac{1}{2(1-\varepsilon)} = \frac{1}{2 \times 2/3} = \frac{3}{4} = 0.75$$

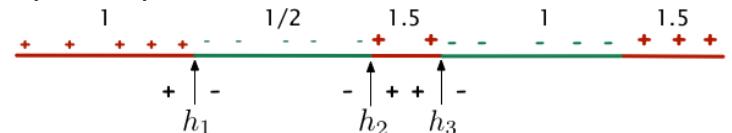
- Sur-pondération des mal classés :**

$$p_m(x) = \frac{1}{2\varepsilon} = \frac{1}{2 \times 1/3} = \frac{3}{2} = 1.5$$



Boosting 39

Exemple simple



- Taux d'erreur :** $\varepsilon_3 = \frac{(5 \times 1/2) + (3 \times 1.5)}{20} = \frac{7}{20} = 0.35$

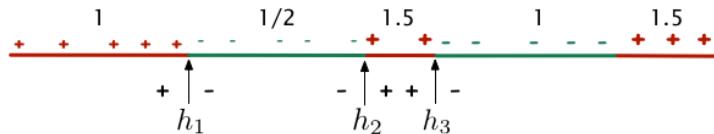
$$\alpha_3 = \frac{1}{2} \ln \frac{1 - \varepsilon_2}{\varepsilon_2} = \frac{1}{2} \ln \frac{0.65}{0.35} = 0.310$$

- Sous-pondération des bien classés :** $p_b(x) = \frac{1}{2(1-\varepsilon)} = \frac{1}{2 \times 0.65} = \frac{1}{1.3} = 0.769$

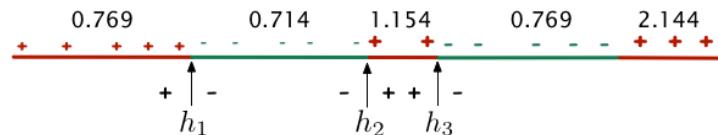
- Sur-pondération des mal classés :** $p_m(x) = \frac{1}{2\varepsilon} = \frac{1}{2 \times 0.35} = \frac{1}{0.7} = 1.429$

Boosting 40

Exemple simple

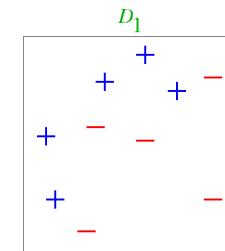


- Sous-pondération des bien classés : $p_b(x) = \frac{1}{2(1-\varepsilon)} = \frac{1}{2 \times 0.65} = \frac{1}{1.3} = 0.769$
- Sur-pondération des mal classés : $p_m(x) = \frac{1}{2\varepsilon} = \frac{1}{2 \times 0.35} = \frac{1}{0.7} = 1.429$



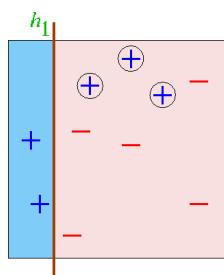
Boosting 41

Exemple jouet

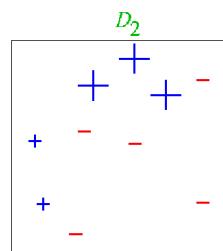


Boosting 42

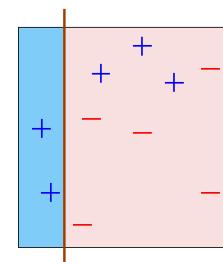
Étape 1



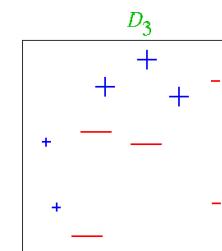
$$\varepsilon_1=0.30 \\ \alpha_1=0.42$$



Étape 2



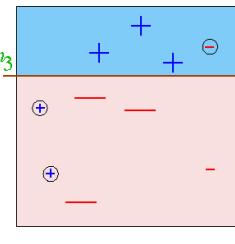
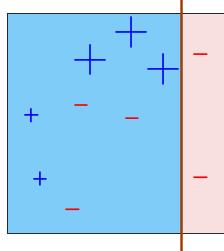
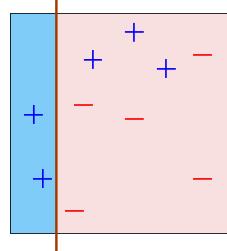
$$\varepsilon_2=0.21 \\ \alpha_2=0.65$$



Boosting 43

Boosting 44

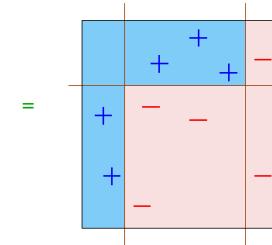
Étape 3



Boosting 45

Hypothèse finale

$$H_{\text{final}} = \text{sign} \left(0.42 + 0.65 + 0.92 \right)$$



Boosting 46

Boosting : Des demos en ligne

<http://www.research.att.com/~yoav/adaboost/index.html>

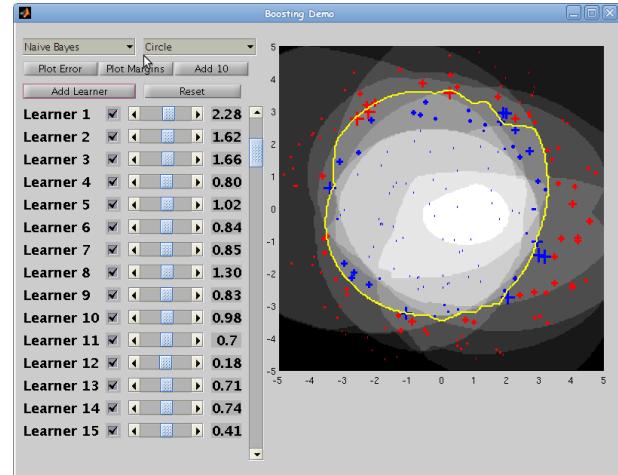
[http://www.mathworks.com/matlabcentral/fx_files/29245/1/
boosting_demo.png](http://www.mathworks.com/matlabcentral/fx_files/29245/1/boosting_demo.png)

Exercice

Boosting 47

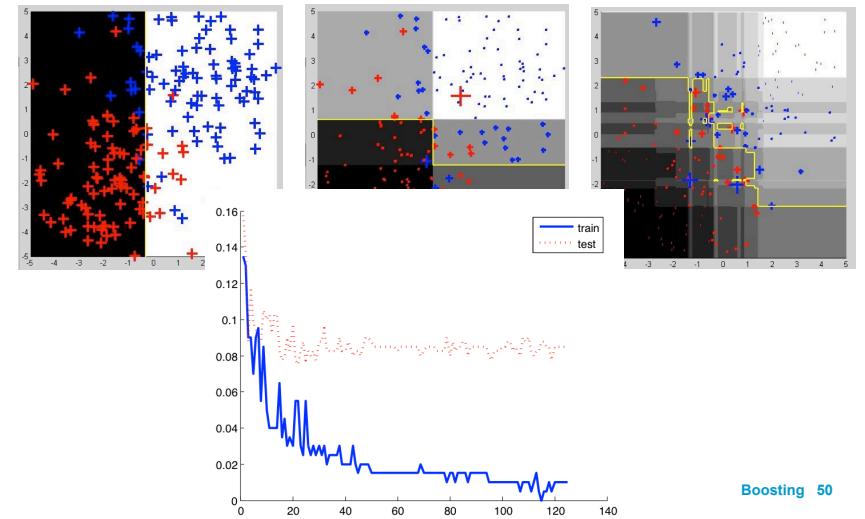
Boosting 48

Boostingdemo de Richard Stapenhurst



Boosting 49

Boostingdemo de Richard Stapenhurst



Boosting 50

Analyse théorique du boosting

Boosting 51

Dérivation de l'algorithme du boosting

Boosting 52

Dérivation de l'algorithme du boosting

■ Re-déivation du boosting

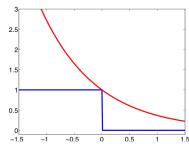
- En choisissant une **fonction de perte surrogée** de forme exponentielle

Soit : $H_{T-1} = \alpha_1 h_1(\mathbf{x}) + \alpha_2 h_2(\mathbf{x}) + \dots + \alpha_{T-1} h_{T-1}(\mathbf{x})$

On veut ajouter : $\alpha_T h_T(\mathbf{x})$

$$\begin{aligned} R_{\text{Emp}}(H_T) &= \sum_{i=1}^m e^{-y_i [H_{T-1}(\mathbf{x}_i) + \alpha_T h_T(\mathbf{x}_i)]} \\ &= \sum_{i=1}^m e^{-y_i H_{T-1}(\mathbf{x}_i)} \cdot e^{-\alpha_T y_i h_T(\mathbf{x}_i)} \\ &= \sum_{i=1}^m W_{T-1}(\mathbf{x}_i) \cdot e^{-\alpha_T y_i h_T(\mathbf{x}_i)} \end{aligned}$$

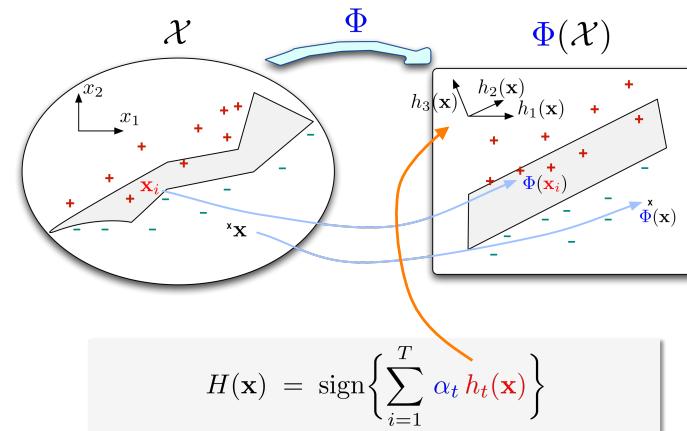
$$\frac{\partial R_{\text{Emp}}(H_T)}{\partial \alpha} \propto e^{-\alpha} \underbrace{(1 - \varepsilon_T)}_{\substack{\text{poids des exemples} \\ \text{correctement prédits}}} + e^{\alpha} \underbrace{\varepsilon_T}_{\substack{\text{poids des exemples} \\ \text{incorrectement prédits}}} \rightarrow \alpha_T = \frac{1}{2} \log \frac{1 - \varepsilon_T}{\varepsilon_T}$$



$$l(h(\mathbf{x}), y) = e^{-y \cdot h(\mathbf{x})}$$

Boosting 53

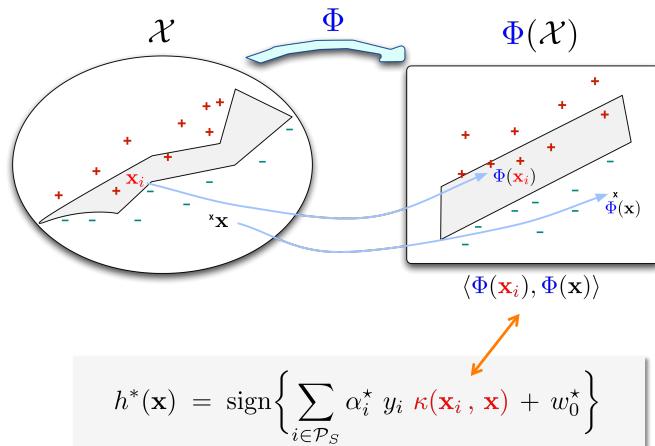
Boosting et redescription



■ Construction itérative de l'espace de redescription

Boosting 54

SVM et méthodes à noyaux



Boosting 55

Bornes en apprentissage
et
en généralisation

Boosting 56

Bornes sur l'erreur en apprentissage

- Theorem:

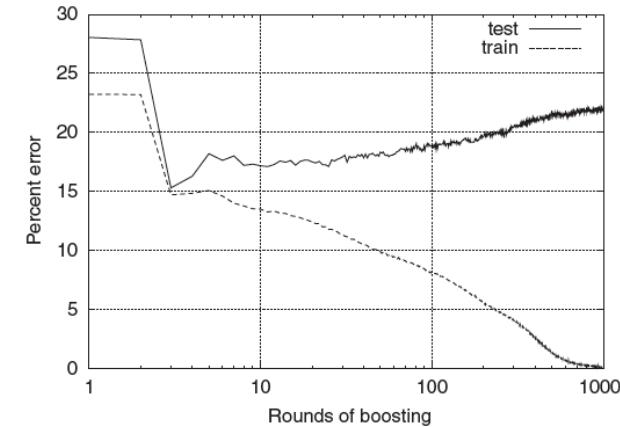
- write ϵ_t as $1/2 - \gamma_t$ [γ_t = "edge"]
- then

$$\begin{aligned} \text{training error}(H_{\text{final}}) &\leq \prod_t \left[2\sqrt{\epsilon_t(1-\epsilon_t)} \right] \\ &= \prod_t \sqrt{1-4\gamma_t^2} \\ &\leq \exp \left(-2 \sum_t \gamma_t^2 \right) \end{aligned}$$

- so: if $\forall t : \gamma_t \geq \gamma > 0$
then training error(H_{final}) $\leq e^{-2\gamma^2 T}$
- AdaBoost is adaptive:
 - does not need to know γ or T a priori
 - can exploit $\gamma_t \gg \gamma$

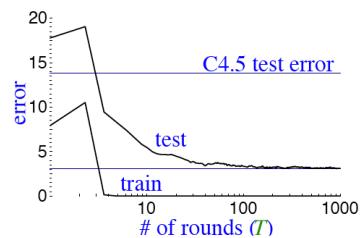
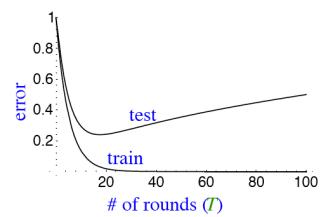
Boosting 57

Evolution des courbes d'erreur (apprentissage & test)



Boosting 58

Comportement de l'erreur en généralisation ?



Arguments pour expliquer
les propriétés du boosting

(the unreasonable power of boosting)

- L'erreur en test **n'augmente pas**, même après 1000 étapes
(2.10^6 noeuds test !!)
 - Boosting de C4.5 sur la base de données « letter »

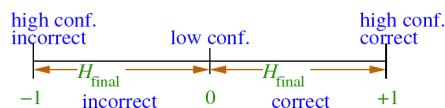
Boosting 59

Boosting 60

Une explication par la « marge »

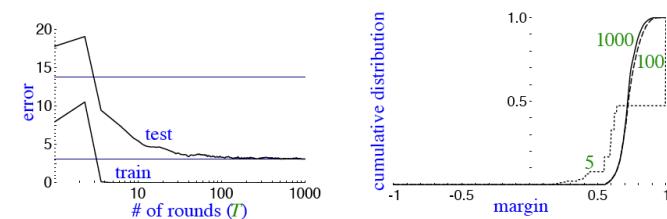
■ Idée :

- L'erreur en apprentissage ne mesure que le taux d'erreur en prédiction
- Il faut aussi **tenir compte de la confiance de la prédiction**
- On peut estimer cette confiance par la **marge**
 - = poids des classificateurs ayant voté correctement
 - poids des classificateurs en erreur



Boosting 61

Étude de la distribution des marges des x_i



	# rounds	5	100	1000
train error	0.0	0.0	0.0	
test error	8.4	3.3	3.1	
% margins ≤ 0.5	7.7	0.0	0.0	
minimum margin	0.14	0.52	0.55	

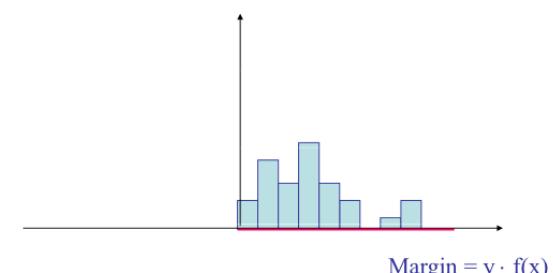
Boosting 62

Argument de la maximisation de la marge

- À chaque étape, AdaBoost placerait le **plus de poids sur les exemples x_i de plus faible marge** tout en continuant à améliorer la marge sur les autres
- L'hypothèse finale serait **complexe mais proche d'une hypothèse simple** d'erreur d'apprentissage proche (et donc d'erreur en généralisation faible aussi)

Boosting 63

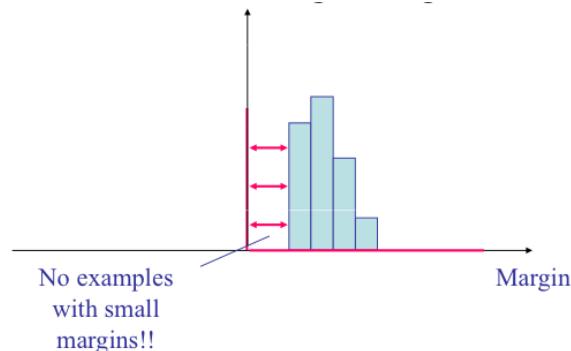
Analyse théorique : maximisation de la marge



Histogram of functional margin for ensemble just after achieving zero training error

Boosting 64

Analyse théorique : maximisation de la marge



Even after zero training error the margin of examples increases.
This is one reason that the generalization error may continue decreasing.

Boosting 65

Plus techniquement

- with high probability, $\forall \theta > 0$:

$$\text{generalization error} \leq \hat{\Pr}[\text{margin} \leq \theta] + \tilde{O}\left(\frac{\sqrt{d/m}}{\theta}\right)$$

$(\hat{\Pr}[\cdot] = \text{empirical probability})$

- bound depends on
 - $m = \#$ training examples
 - $d =$ “complexity” of weak classifiers
 - entire distribution of margins of training examples
- $\hat{\Pr}[\text{margin} \leq \theta] \rightarrow 0$ exponentially fast (in T) if
(error of h_t on D_t) $< 1/2 - \theta$ ($\forall t$)
 - so: if weak learning assumption holds, then all examples will quickly have “large” margins

Boosting 66

Bornes en généralisation

$$R_{\text{Réel}} \leq R_{\text{Emp}} + \mathcal{O}\left(\sqrt{\frac{T \cdot d_{\mathcal{H}}}{m}}\right)$$

Boosting 67

Autres perspectives pour expliquer les propriétés du boosting

Boosting 68

Autres perspectives

- La théorie des jeux
 - Inspiration originale pour développer le boosting
- Minimisation de la perte (surrogée)
 - Vue plus haut (et plus loin)
- Point de vue géométrie de l'information

Bilan

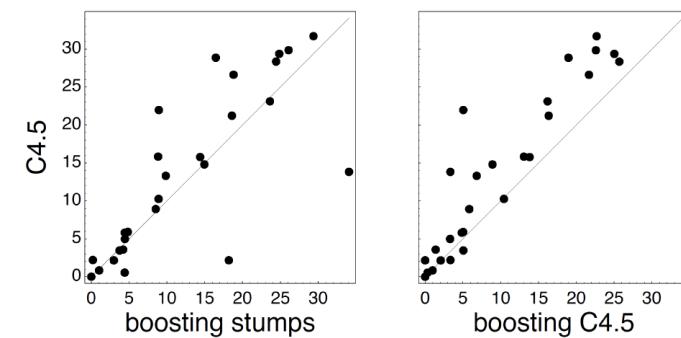
Boosting 69

Boosting 70

Avantages pratiques de AdaBoost

- (très) rapide
- simple + facile à programmer
- Un seul paramètre à régler : le nombre d'étapes de boosting (T)
- Applicable à de nombreux domaines par un bon choix de classifieur faible (neuro net, C4.5, ...)
- Pas de sur-spécialisation par la maximisation des marges
- Peut être adapté au cas où $h_t : \mathcal{X} \rightarrow \mathcal{R}$; la classe est définie par le signe de $h_t(x)$; la confiance est donnée par $|h_t(x)|$
- Peut être adapté aux problèmes multi-classes où $y_i \in \{1, \dots, c\}$ et aux problèmes multi-étiquettes
- Permet de trouver les exemples aberrants (outliers)

Performances du boosting



Test error rate on 27 benchmark problems
x-axis: boosting; y-axis: base-line (C4.5)

Boosting 71

Boosting 72

Quand est-ce que ça ne marche pas

- Rappel : No-free-lunch-theorem
- Boosting inadapté quand
 - Pas assez de données
 - Comité (d'apprenants faibles) trop restreint
 - Apprenants faibles trop stables
 - Apprenants faibles trop forts !
 - Peuvent faire du surapprentissage par eux-mêmes
 - Données bruitées

Aspects pratiques

Avantages	Difficultés
<ul style="list-style-type: none">• Un meta-algorithme d'apprentissage : utiliser n'importe quel algorithme d'apprentissage faible• En principe, un seul paramètre à régler (le nombre T d'itérations)• Facile et aisément programmable• Performances théoriques garanties	<ul style="list-style-type: none">• Difficile d'incorporer des connaissances a priori• Difficile de savoir comment régulariser• Le meilleur choix d'un apprenant faible n'est pas évident• Les frontières de décision en utilisant des méthodes parallèles aux axes sont souvent très irrégulières (non interprétable)

Boosting 73

Boosting 74

Boosting : résumé

- La prédiction finale est issue d'une combinaison (vote pondéré) de plusieurs prédictions
- Méthode :
 - Itérative
 - Chaque classifieur dépend des précédents
(les classificateurs ne sont donc pas indépendants comme dans d'autres méthodes de vote)
 - Les exemples sont pondérés différemment
 - Le poids des exemples reflète la difficulté des classificateurs précédents à les apprendre

Autres algorithmes de boosting

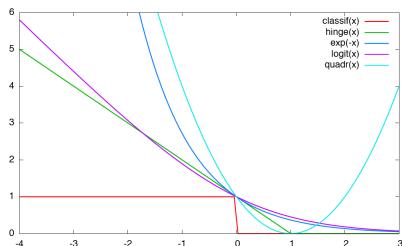
Boosting 75

Boosting 76

Fonctions « surrogées »

Fonctions de la forme $\ell(x, y) = \varphi(-xy)$:

Nom	$\varphi(x)$
Exponentiel	$\exp(-x)$
Logit	$\log_2(1 + \exp(-x))$
Quadratique	$(1 - x)^2$
Quadratique tronqué	$(1 - x)_+^2$
Hinge (SVM)	$(1 - x)_+$



Boosting 77

LogitBoost

■ Utilisation de la fonction de perte logloss

- Pas de poids exorbitant sur les points aberrants
- Possibilité de calculer des probabilités associées aux étiquettes

$$p(y|\mathbf{x}) = \frac{e^{h(\mathbf{x})}}{e^{-h(\mathbf{x})} + e^{h(\mathbf{x})}} = \frac{1}{1 + e^{-2h(\mathbf{x})}}$$

Boosting 78

Applications

Applications du boosting

Text classification

Schapire and Singer - Used stumps with normalized term frequency and multi-class encoding

OCR

Schwenk and Bengio (neural networks)

Natural language Processing

Collins; Haruno, Shirai and Ooyama

Image retrieval

Thieu and Viola

Medical diagnosis

Merle *et al.*

Fraud Detection

Rätsch & Müller 2001

Drug Discovery

Rätsch, Demiriz, Bennett 2002

Elect. Power Monitoring

Onoda, Rätsch & Müller 2000

Fuller list: Schapire's 2002, Meir & Rätsch 2003 review

Boosting 79

Boosting 80

Face detection idea

- 1) in Adaboost use parallel-axis (tree decision) classifier
- 2) in Viola Jones, the weak classifier is the specially designed classifier described in the paper.

Robust real-time face detection [Viola & Jones, 2004]

- Images 384 x 288 (niveaux de gris)
- Déetecter des visages à toute échelle
- En temps réel (15 images / s) sur un smartphone !!

Adaboost v.2c

81 Boosting 81

Boosting 82

Les descripteurs

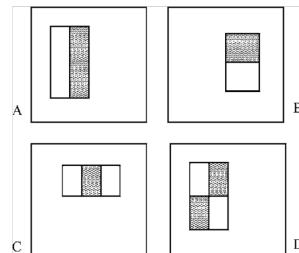
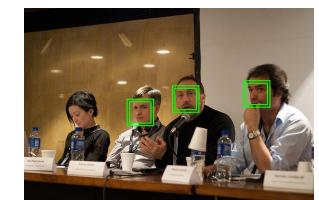
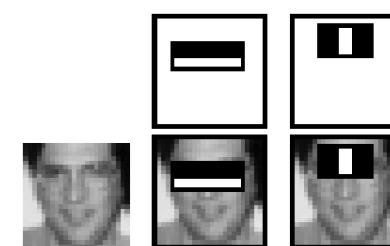
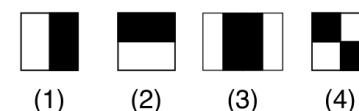


Figure 1. Example rectangle features shown relative to the enclosing detection window. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. Two-rectangle features are shown in (A) and (B). Figure (C) shows a three-rectangle feature, and (D) a four-rectangle feature.

- Plus de 3 000 000 000 !
 - Toutes échelles
 - Seuil à définir

Boosting 83

Useful Features Learned by Boosting



Adaboost v.2c

Boosting 84

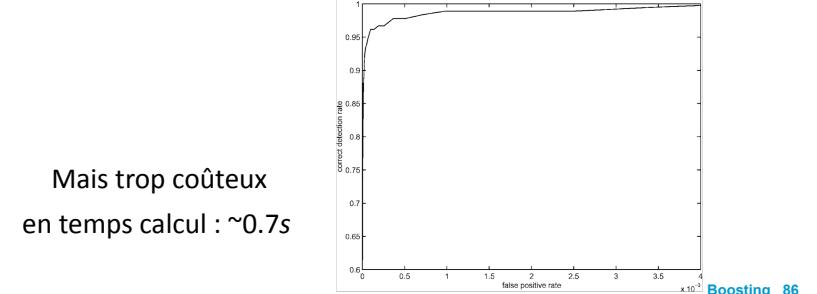
Example of the training set



Sélection des descripteurs utiles

■ En utilisant AdaBoost

- Les descripteurs sont associés à des « decision stumps »
- Le boosting sélectionne une séquence de descripteurs
 - 200 dans cette étude



Organisation des détecteurs en cascade

- Éliminer le plus tôt possible les négatifs

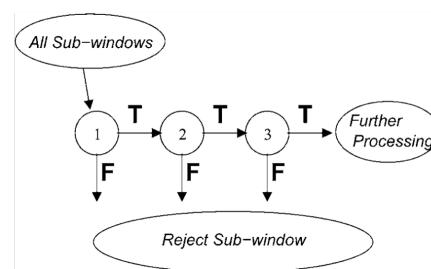
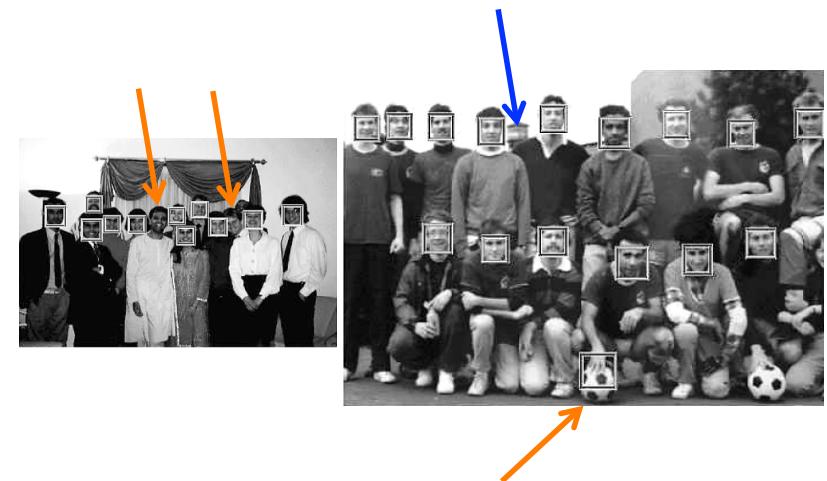


Figure 6. Schematic depiction of a the detection cascade. A series of classifiers are applied to every sub-window. The initial classifier eliminates a large number of negative examples with very little processing. Subsequent layers eliminate additional negatives but require additional computation. After several stages of processing the number of sub-windows have been reduced radically. Further processing can take any form such as additional stages of the cascade (as in our detection system) or an alternative detection system.

Boosting 87

Face detection using boosting



Applications

- Reconnaissance du contour du rein



Le bagging

Boosting 90

Bagging

[Breiman, 96]

- Génération de k échantillons « indépendants » par tirage avec remise dans l'échantillon S_m
- Pour chaque échantillon, apprentissage d'un classifieur en utilisant le même algorithme d'apprentissage
- La **prédiction finale** pour un nouvel exemple est obtenue par vote (simple) des classifieurs

Boosting 91

Bagging (suite)

- Il est souvent dit que :
 - Le bagging fonctionne en réduisant la variance en laissant le biais inchangé
- Mais, encore incomplètement compris
 - Voir [Yves Grandvalet : « Bagging equalizes influence », *Machine Learning*, 55(3), pages 251-270, 2004.]

Boosting 92

Les forêts aléatoires

Les forêts aléatoires (Random forests)

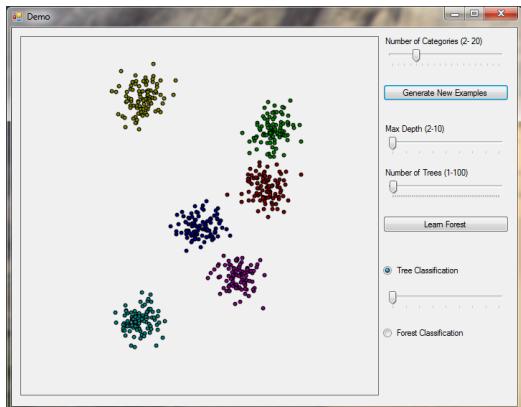
- Principe :
 - réduire la corrélation entre apprenants faibles

- Apprendre des arbres
 - Sur des jeux de données différents (comme le bagging)
 - Sur des sous-ensembles d'attributs tirés aléatoirement

Boosting 93

Boosting 94

Toy Forest Classification Demo

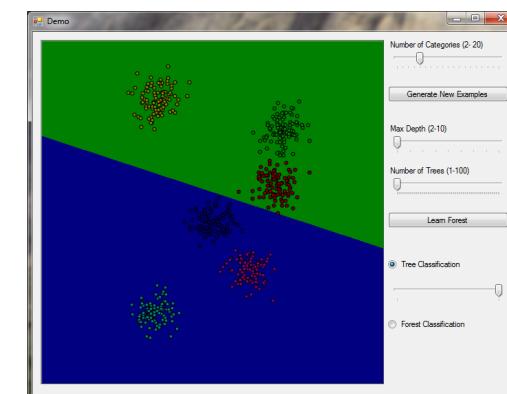


6 classes in a 2 dimensional feature space.

Split functions are lines in this space.

Boosting 95

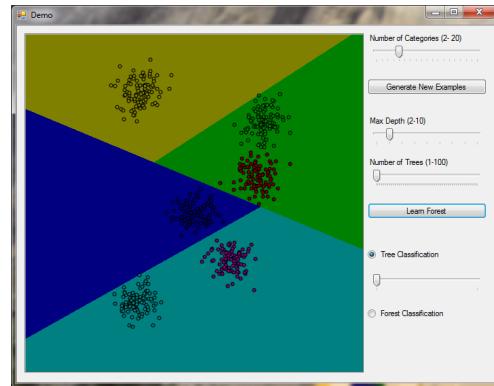
Toy Forest Classification Demo



With a **depth 2** tree, you can not separate all six classes.

Boosting 96

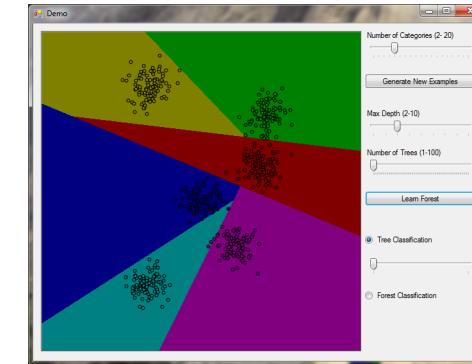
Toy Forest Classification Demo



With a **depth 3** tree, you are doing better,
but still cannot separate all six classes.

Boosting 97

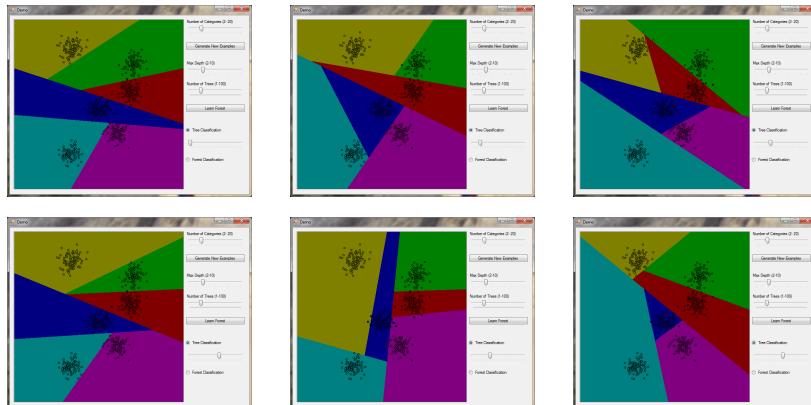
Toy Forest Classification Demo



With a **depth 4** tree, you now have at least as many leaf nodes
as classes, and so are able to classify most examples correctly.

Boosting 98

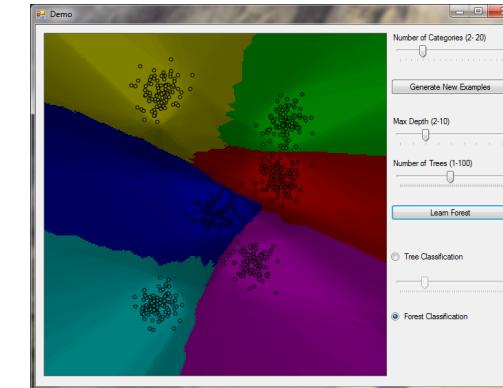
Toy Forest Classification Demo



Different trees within a forest can give rise to very different decision
boundaries, none of which is particularly good on its own.

Boosting 99

Toy Forest Classification Demo



But **averaging together many trees** in a forest can result in decision boundaries
that look very sensible, and are even quite close to the max margin classifier.
(Shading represents entropy – darker is higher entropy).

Boosting 100

Bilan sur les forêts aléatoires

- Souvent puissant
- Mais
 - Perte d'interprétabilité
 - Et plus coûteux en temps calcul

Error Correcting Output Codes (ECOC)

Boosting 101

Boosting 102

ECOC

- Apprendre à distinguer 26 lettres
- Codage
 - Sur 5 bits ?
 - Sur 10 bits !

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
A	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	1	0	1	0	1	1
C	0	0	0	1	0	1	0	1	0	1

4 4
6

ECOC

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
A	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	1	0	1	0	1	1
C	0	0	0	1	0	1	0	1	0	1

- On apprend 10 classificateurs binaires

$$h_5(\mathbf{x}) = \begin{cases} 1 & \text{si } y \in \{A, C, E, G, I, K, M, O, Q, S, U, W, Y\} \\ 0 & \text{sinon} \end{cases}$$

$$h_4(\mathbf{x}) = \begin{cases} 1 & \text{si } y \in \{A, B, E, F, I, J, M, N, Q, R, U, V, Y, Z\} \\ 0 & \text{sinon} \end{cases}$$

- On calcule $h_1(\mathbf{x}), h_2(\mathbf{x}), h_3(\mathbf{x}), h_4(\mathbf{x}), h_5(\mathbf{x}), h_6(\mathbf{x}), h_7(\mathbf{x}), h_8(\mathbf{x}), h_9(\mathbf{x}), h_{10}(\mathbf{x})$
- On décode $H(\mathbf{x})$ par plus proche voisin / au codage de A, B, C, ...

Boosting 103

Boosting 104

Références bibliographiques

- Bob Shapire and Yoav Freund
Boosting: Foundations and Algorithms
MIT Press, 2012
- Ron Meir and Gunnar Rätsch
An introduction to Boosting and Leveraging
In *Advanced Lectures on Machine Learning* (LNAI-2600), 2003
<http://www.boosting.org/papers/MeiRae03.pdf>
- Zhi-Hua Zhou
Ensemble Methods. Foundations and Algorithms
CRC Press, 2012
- Antoine Cornuéjols et Laurent Miclet
Apprentissage artificiel. Concepts et algorithmes
Eyrolles, 2010

Boosting 105

Des méthodes d'ensemble
en apprentissage *non supervisé* ?

Boosting 106

Quels ingrédients ?

1. Comment **sélectionner** des « experts » ?
 - Qu'est-ce qu'un **expert** ?
 - Qu'est-ce qu'un **bon panel d'experts** ?
2. Comment leur attribuer un **poids** (éventuellement) ?
3. Comment **combiner** leur avis ?

Références

Boosting 107

Boosting 108

Références bibliographiques

- Bob Shapire and Yoav Freund
Boosting: Foundations and Algorithms
MIT Press, 2012
- Ron Meir and Gunnar Rätsch
An introduction to Boosting and Leveraging
In *Advanced Lectures on Machine Learning* (LNAI-2600), 2003
<http://www.boosting.org/papers/MeiRae03.pdf>
- Zhi-Hua Zhou
Ensemble Methods. Foundations and Algorithms
CRC Press, 2012
- Antoine Cornuéjols et Laurent Miclet
Apprentissage artificiel. Concepts et algorithmes
Eyrolles, 2010

Etat de l'art (historique)

- [Valiant'84]
introduced theoretical PAC model for studying machine learning
- [Kearns & Valiant'88]
open problem of finding a boosting algorithm
- [Schapire'89], [Freund'90]
first polynomial-time boosting algorithms
- [Drucker, Schapire & Simard '92]
first experiments using boosting

Boosting 109

Boosting 110

Etat de l'art (suite)

- [Freund & Schapire '95]
 - introduced AdaBoost algorithm
 - strong practical advantages over previous boosting algorithms
- experiments using AdaBoost:

[Drucker & Cortes '95]	[Schapire & Singer '98]
[Jackson & Cravon '96]	[Maclin & Opitz '97]
[Freund & Schapire '96]	[Bauer & Kohavi '97]
[Quinlan '96]	[Schwenk & Bengio '98]
[Breiman '96]	[Dietterich'98]
- continuing development of theory & algorithms:

[Schapire, Freund, Bartlett & Lee '97]	[Schapire & Singer '98]
[Breiman '97]	[Mason, Bartlett & Baxter '98]
[Grove and Schuurmans'98] [Friedman, Hastie & Tibshirani '98]	

Boosting 111