

Étude de l'Apprentissage Actif, Application à la Conduite d'Expériences

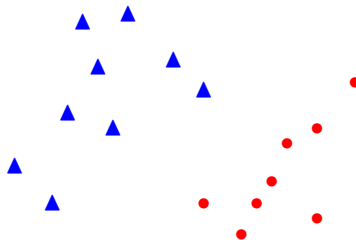
JÉRÉMIE MARY

Sous la direction de
ANTOINE CORNUÉJOLS et MICHÈLE SEBAG

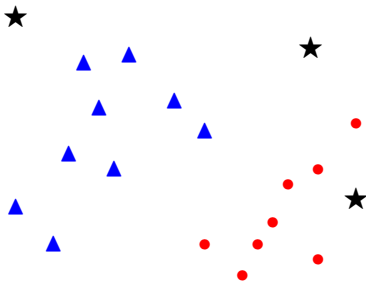
12 décembre 2005



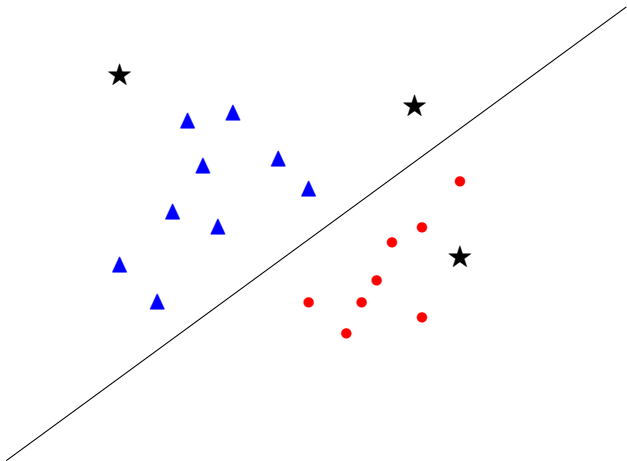
Apprentissage Supervis 



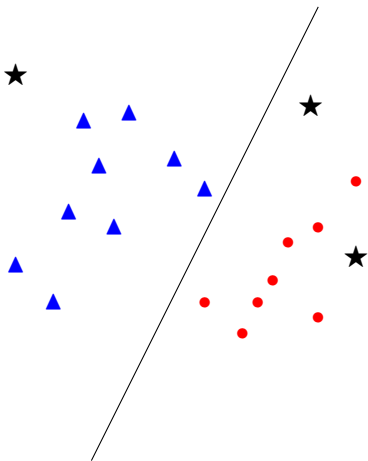
Apprentissage Supervisé



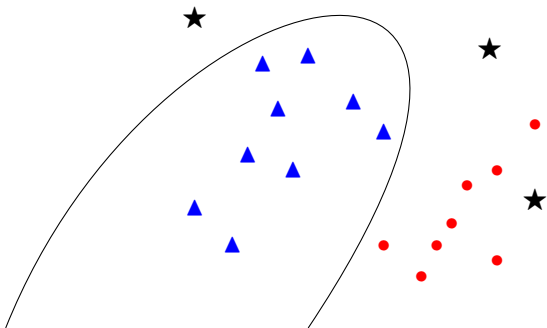
Apprentissage Supervis 



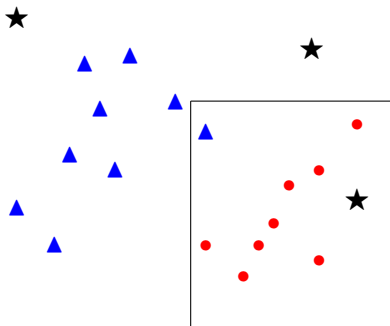
Apprentissage Supervis 



Apprentissage Supervis 



Apprentissage Supervis 



Apprentissage

Principe

- Il faut choisir une hypoth se $f \in F$ en regardant des exemples (x_1, \dots, x_n) ;
- Le but est d'avoir une **erreur en g n ralisation la plus faible** possible ;
- La **base d'exemples est cruciale**.

SI LA BASE D'APPRENTISSAGE N'EST PAS ADAPT E, LE CHOIX DE LA BONNE HYPOTH SE PEUT SE R V LER IMPOSSIBLE !



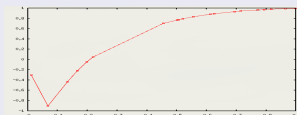
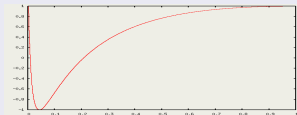
Apprentissage Actif

On peut influencer sur la base d'exemples

- *a priori*
 - En donnant les points au fur et à mesure (*on-line*)
 - Si l'espace est **divisé en strates** et que l'on peut demander des points dans une des strates. Les strates peuvent :
 - Coïncider avec les classes ;
 - Couvrir l'espace de façon équilibrée ;

Exemple

Approximer $\cos(\log(x))$ par une fonction linéaire par morceaux



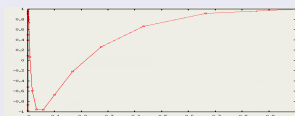
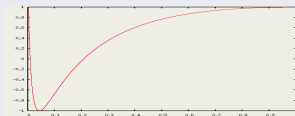
Apprentissage Actif

On peut influencer sur la base d'exemples

- *a priori*
- **En donnant les points au fur et a mesure (*on-line*)**
- Si l'espace est **divisé en strates** et que l'on peut demander des points dans une des strates. Les strates peuvent :
 - Coïncider avec les classes ;
 - Être prédéfinies ;
 - Être choisies par un processus d'optimisation.

Exemple

Approximer $\cos(\log(x))$ par une fonction linéaire par morceaux



Apprentissage Actif

On peut influencer sur la base d'exemples

- *a priori*
- En donnant les points au fur et a mesure (*on-line*)
- Si l'espace est **divisé en strates** et que l'on peut demander des points dans une des strates. Les strates peuvent :
 - **Coïncider avec les classes** ;
 - Être prédéfinies ;
 - Être elles mêmes définies pendant l'apprentissage.

Exemple

Étude biologique sur l'irradiation à faible doses de levures,

- Deux classes : irradié et non-irradié
- Il peut être intéressant d'avoir **davantage d'échantillons irradiés** pour mieux évaluer l'impact des changements induits.

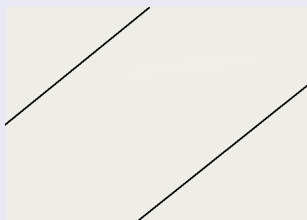


Apprentissage Actif

On peut influencer sur la base d'exemples

- *a priori*
- En donnant les points au fur et a mesure (*on-line*)
- Si l'espace est **divisé en strates** et que l'on peut demander des points dans une des strates. Les strates peuvent :
 - Coïncider avec les classes ;
 - **Être prédéfinies ;**
 - Être elles mêmes définies pendant l'apprentissage.

Exemple

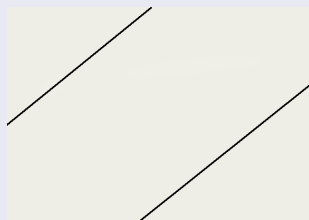


Apprentissage Actif

On peut influencer sur la base d'exemples

- *a priori*
- En donnant les points au fur et a mesure (*on-line*)
- Si l'espace est **divisé en strates** et que l'on peut demander des points dans une des strates. Les strates peuvent :
 - Coïncider avec les classes ;
 - Être prédéfinies ;
 - **Être elles mêmes définies pendant l'apprentissage.**

Exemple

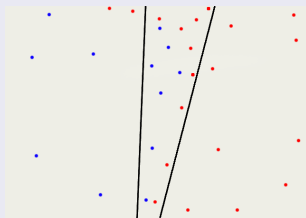


Apprentissage Actif

On peut influencer sur la base d'exemples

- *a priori*
- En donnant les points au fur et a mesure (*on-line*)
- Si l'espace est **divisé en strates** et que l'on peut demander des points dans une des strates. Les strates peuvent :
 - Coïncider avec les classes ;
 - Être prédéfinies ;
 - **Être elles mêmes définies pendant l'apprentissage.**

Exemple



Minimisation du Risque Empirique

Fonction de Coût

- On définit le coût L associé au choix d'une fonction $f \in F$. Par exemple $L(f) = E(|f - g|)$.
- Un des meilleur choix possible est $f^* \in \underset{f \in F}{\operatorname{argmin}} L(f)$
- Problème : **on ne dispose pas de l'opérateur d'espérance.**

Solution

- On utilise un **estimateur basé sur les exemples**. Dans le cas précédent,
$$\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|$$
- On choisit alors $\hat{f}_n \in \underset{f \in F}{\operatorname{argmin}} (\hat{L}_n(f))$



Minimisation du Risque Empirique

Fonction de Coût

- On d finit le coût L associ  au choix d'une fonction $f \in F$. Par exemple $L(f) = E(|f - g|)$.
- Un des meilleur choix possible est $f^* \in \underset{f \in F}{\operatorname{argmin}} L(f)$
- Probl me : **on ne dispose pas de l'op rateur d'esprance.**

Solution

- On utilise un **estimateur bas  sur les exemples**. Dans le cas pr c dent, $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i|$
- On choisit alors $\hat{f}_n \in \underset{f \in F}{\operatorname{argmin}} (\hat{L}_n(f))$



Le Cas de l'Erreur Nulle

Cas particulier

- Si $E(L(f^*)) = 0$, alors l'erreur en g n ralisation **converge vers 0** en $O\left(\frac{\log(n)}{n}\right)$

Mais...

- Cette hypoth se est **fausse si le concept cible n'est pas dans F**
- Ou si le concept comporte une **part stochastique** (ou du bruit).

ON SOUHAITE S'AFFRANCHIR DE CETTE HYPOTH SE.



Le Cas de l'Erreur Nulle

Cas particulier

- Si $E(L(f^*)) = 0$, alors l'erreur en g n ralisation **converge vers 0** en $O\left(\frac{\log(n)}{n}\right)$

Mais...

- Cette hypoth se est **fausse si le concept cible n'est pas dans F**
- Ou si le concept comporte une **part stochastique** (ou du bruit).

ON SOUHAITE S'AFFRANCHIR DE CETTE HYPOTH SE.



Dimension de Vapnik-Chervonenkis

Id e

- Mesurer la taille de l'espace d'hypoth ses F .
- On d finit une mesure sur F , appel e VC-dim d pendant du nombre de fa ons dont F peut s parer un ensemble de n points.

R sultat

- 1 Dans les cas ou la VC-dim de F est finie (hypoth se forte), la convergence vers l'erreur optimale est en $O\left(\frac{\log(n)}{\sqrt{n}}\right)$

LA D CROISSANCE EST MOINS RAPIDE QUE DANS LE CAS PR C DENT.
EN G N RAL, IL FAUDRA D'AVANTAGE D'EXEMPLES POUR ATTEINDRE
LE M ME TAUX D'ERREUR.



Plan

- 1 Introduction
- 2 Exemples a priori : Discrépance**
- 3 Stratification de la base d'apprentissage
- 4 Sélection d'Attributs
- 5 Conclusion, Perspectives

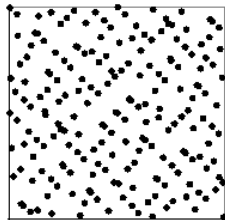
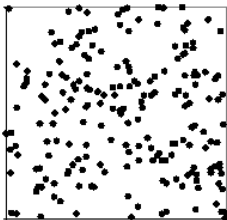


Définition

Discrépance

La star-discrépance de la suite (x_i) de $[0, 1]^d$ est définie par :

$$D_n^*(X) = \sup_{a \in [0;1]^d} \left| \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{x_k \in [0;a]\}} - \prod_{i=1}^d a_i \right|$$

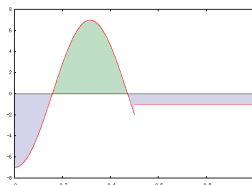
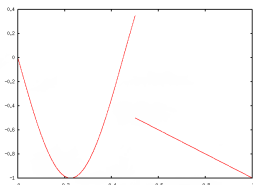


In galit  de Koksma-Hlawka

Th or me

Si f est une application de $[0, 1]^d$ dans \mathbb{R} , de variation totale au sens de Hardy et Krause major e par $V_{HK}(f)$, et si x_i est une suite dans $[0, 1]^d$, alors

$$\left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \int f \right| \leq V_{HK}(f) D_n^*(x_1, \dots, x_n)$$

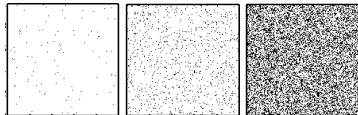


Faible Discrédance

Candidat

- 1 Points tirés uniformément ; discrédance de l'ordre de

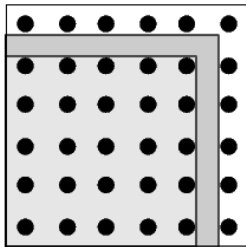
$$O\left(\frac{\sqrt{\ln \ln(n)}}{\sqrt{2n}}\right)$$



Faible Discr ance

Candidat

- 1 Grille uniforme, discr ance de l'ordre de $\frac{1}{\sqrt{n}}$

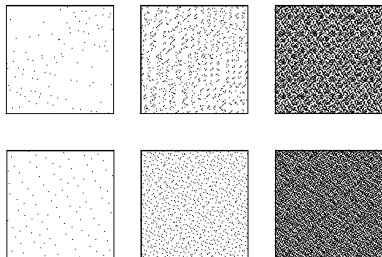


Faible Discr panance

Candidat

1 Constructions arithm tiques

$$O\left(\frac{(\ln(n))^d}{n}\right)$$

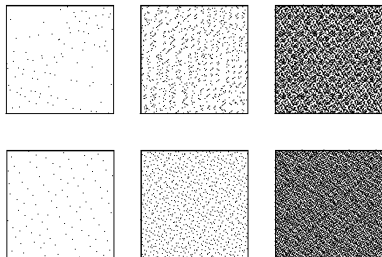


Faible Discr pance

Candidat

1 Constructions arithm tiques

$$O\left(\frac{(\ln(n))^d}{n}\right)$$



A d FIX , IMPOSSIBLE DE FAIRE MIEUX QUE $O\left(\frac{(\ln(n))^{\frac{d}{2}}}{n}\right)$



Intérêt pour l'Apprentissage

Application directe

Si l'on suppose que pour tout $f \in F$ la variation de $|f - g|^p$ est bornée par M alors :

$$L(\hat{f}) \leq \inf_f L(f) + 2MD_n^*(x_1, \dots, x_n)$$

Améliorations

- $E|\hat{f} - g| \leq \hat{E}|\hat{f} - g| + \left(V_{HK}(\hat{f}) + V_{HK}(g) \right) D_n^*$
- Et si tous les $\hat{f} - g$ sont dans un espace de fonction borné pour la norme de Hölder de degré d

$$E|\hat{f} - g| \leq \underbrace{\hat{E}|\hat{f} - g|}_{\text{erreur empirique}} + \left(2V_{HK}(\hat{f}) + o(1) \right) D_n^*$$



Intérêt pour l'Apprentissage

Application directe

Si l'on suppose que pour tout $f \in F$ la variation de $|f - g|^p$ est bornée par M alors :

$$L(\hat{f}) \leq \inf_f L(f) + 2MD_n^*(x_1, \dots, x_n)$$

Améliorations

- 1 $E|\hat{f} - g| \leq \hat{E}|\hat{f} - g| + \left(V_{HK}(\hat{f}) + V_{HK}(g) \right) D_n^*$
- 2 Et si tous les $\hat{f} - g$ sont dans un espace de fonction borné pour la norme de Hölder de degré d

$$E|\hat{f} - g| \leq \underbrace{\hat{E}|\hat{f} - g|}_{\text{erreur empirique}} + \left(2V_{HK}(\hat{f}) + o(1) \right) D_n^*$$



Comparaisons th oriques

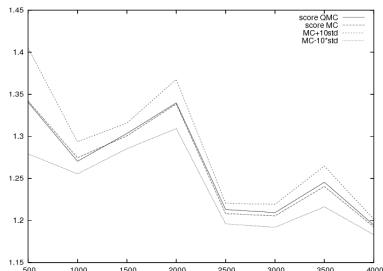
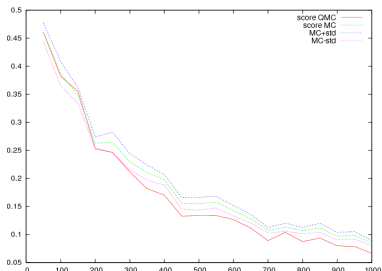
	Suites � faible discr�pance	Monte-Carlo avec VC-dimension finie	Monte-Carlo avec classes de Donsker
Norme de H�lder born�e de degr� α	$\alpha > d, O(\log(n)^d/n)$	-	$\alpha > d/2, O(1/\sqrt{n})$
VC-dim finie et $L^* = 0$	- -	$O(\log(n)/\sqrt{n})$ $O(\log(n)/n)$	$O(1/\sqrt{n})$ $O(1/\sqrt{n})$
Variation totale born�e	$O(\log(n)^d/n)$	-	-
Minimisation du risque structurel	pour une variation totale finie, $O(\log(n)^d/n)$	Consistance universelle et $O(\log(n)/\sqrt{n})$ dans une famille de VC-dimension ∞ d'approximation de fonctions	- -



Validation Exp rimentale

Protocole

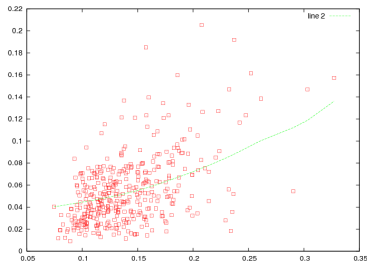
- La fonction cible d'un point (x_1, \dots, x_d) est la somme des d dimensions : $\sum_{i=1}^d x_i$.
- L'apprentissage est effectu  par un RBF avec $\sigma = 0.1$



Validation Expérimentale

Protocole

- La fonction cible d'un point (x_1, \dots, x_d) est la somme des d dimensions : $\sum_{i=1}^d x_i$.
- L'apprentissage est effectué par un RBF avec $\sigma = 0.1$



Plan

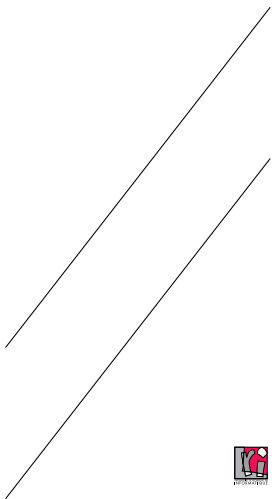
- 1 Introduction
- 2 Exemples a priori : Discrépance
- 3 Stratification de la base d'apprentissage**
- 4 Sélection d'Attributs
- 5 Conclusion, Perspectives



Donn es Naturellement Stratifi es

Algorithme

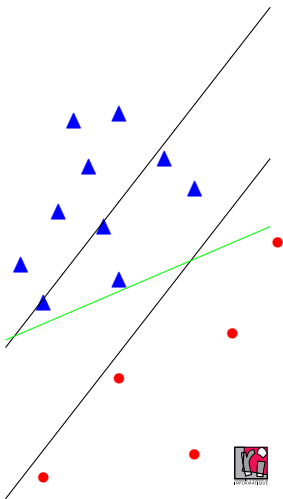
- 1 Tirer $k = o(n)$ exemples ;
Soit \hat{L}' le c t empirique associ  ; soit ε' la pr cision telle que pour tous f optimaux avec une pr cision $\varepsilon'/2$ pour \hat{L}' sont aussi optimaux pour L avec pr cision ε' au niveau de confiance $1 - \delta(\varepsilon', k) \geq 1 - \delta_1$.
- 2 Phase d'apprentissage : Trouver $h = \operatorname{argmin} \hat{L}'$;



Donn es Naturellement Stratifi es

Algorithme

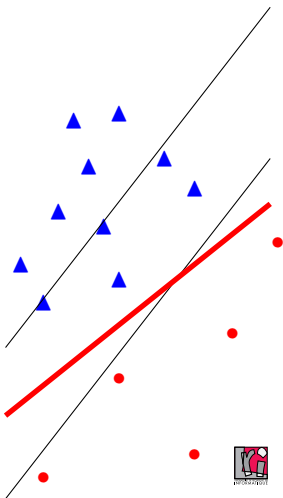
- 1 Tirer $k = o(n)$ exemples ;
Soit \hat{L}' le c t empirique associ  ; soit ε' la pr cision telle que pour tous les f optimaux avec une pr cision $\varepsilon'/2$ pour \hat{L}' sont aussi optimaux pour L avec pr cision ε' au niveau de confiance $1 - \delta(\varepsilon', k) \geq 1 - \delta_1$.
- 2 Phase d'apprentissage : Trouver $h = \operatorname{argmin} \hat{L}'$;



Donn es Naturellement Stratifi es

Algorithme

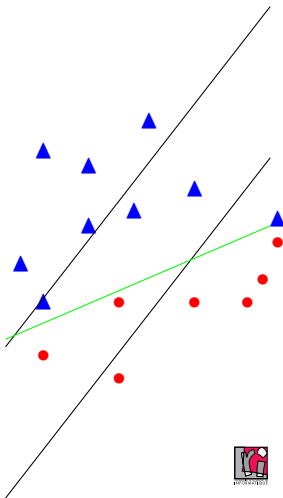
- 1 Tirer $k = o(n)$ exemples ;
Soit \hat{L}' le co t empirique associ  ; soit ε' la pr cision telle que pour tous les f optimaux avec une pr cision $\varepsilon'/2$ pour \hat{L}' sont aussi optimaux pour L avec pr cision ε' au niveau de confiance $1 - \delta(\varepsilon', k) \geq 1 - \delta_1$.
- 2 Phase d'apprentissage : Trouver $h = \operatorname{argmin} \hat{L}'$;



Donn es Naturellement Stratifi es

Algorithme

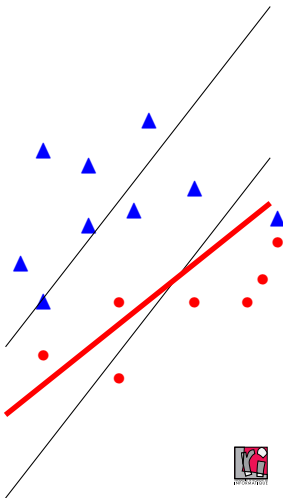
- 1 Tirer $k' = o(n)$ exemples (i.i.d.)
- 2 Estimer $\sigma_i(h)$ qui est la racine de la variance du co t de h dans la classe i (avec $h = \operatorname{argmin} \hat{L}'$);



Donn es Naturellement Stratifi es

Algorithme

- 1 Tirer $k' = o(n)$ exemples (i.i.d.)
- 2 Estimer $\sigma_i(h)$ qui est la racine de la variance du co t de h dans la classe i (avec $h = \operatorname{argmin} \hat{L}'$);



Donn es Naturellement Stratifi es

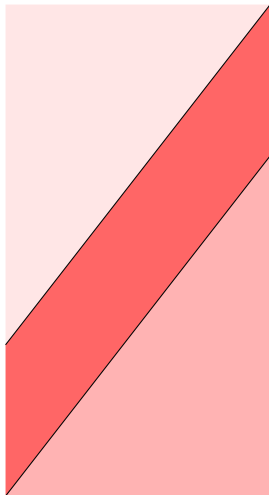
Algorithme

- Calculer, gr ce aux estimations de $\sigma_i(h)$, un vecteur $(\sigma'_i)_{i \in \llbracket 1, K \rrbracket}$ tel que

$$\forall f; \hat{L}'(f) \leq \hat{L}'(h) + \varepsilon'/2 \Rightarrow \forall i; \sigma_i(f) \leq \sigma'(i)$$

au niveau de confiance $1 - \delta_2$.

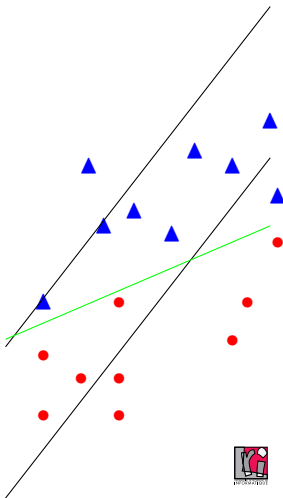
- D duire de cette borne une **formule** $\Delta'(n_1, \dots, n_K)$ telle que pour un co t empirique \hat{L} associ    n_K exemples dans la classe K :
 $P(\exists f; L(f) \leq \inf_F \hat{L} + \varepsilon'/2 \text{ et } \text{Var} \hat{L}(f) > \Delta'(n_1, \dots, n_K)) \leq \delta_2$



Donn es Naturellement Stratifi es

Algorithme

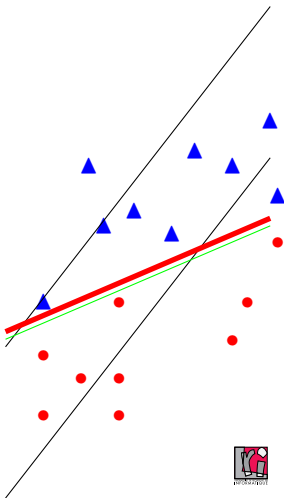
- 1 Avec n_1, \dots, n_k optimisant $\Delta'(n_1, \dots, n_k)$ sous la contrainte $\sum_i n_i = n$, tirer n exemples avec n_i exemples dans la i^{e} classe.
- 2 Soit ε tel que $|F_{\varepsilon'}^e| \delta'(\varepsilon, \Delta') \leq \delta_3$,
Apprendre $\hat{f} \in \operatorname{argmin}_{F_{\varepsilon'}^e} \hat{L}$.



Donn es Naturellement Stratifi es

Algorithme

- 1 Avec n_1, \dots, n_k optimisant $\Delta'(n_1, \dots, n_k)$ sous la contrainte $\sum_i n_i = n$, tirer n exemples avec n_i exemples dans la i^{e} classe.
- 2 Soit ε tel que $|F_{\varepsilon'}^e| \delta'(\varepsilon, \Delta') \leq \delta_3$,
Apprendre $\hat{f} \in \operatorname{argmin}_{F_{\varepsilon'}^e} \hat{L}$.



Résultat

Théorème

En utilisant l'algorithme précédent on a :

$$P(L(\hat{f}) \geq \inf_f L(f) + 3\varepsilon) \leq |F_{\varepsilon'}^{\varepsilon}| 2e^{-\frac{\varepsilon^2}{4\Delta'(n_1, \dots, n_K)}} + \delta(\varepsilon', k) + \delta_2$$

Remarques

- 1 L'amélioration de la borne provient de l'optimisation finale ;
- 2 Si la racine de l'erreur est équidistribuée entre les classes, on retrouve les mêmes résultats que dans le cas i.i.d.
- 3 Regrouper les deux bases d'apprentissage ne modifie pas beaucoup la borne ;
- 4 On peut réaliser plusieurs passes.



Variantes

Partitionnement automatique

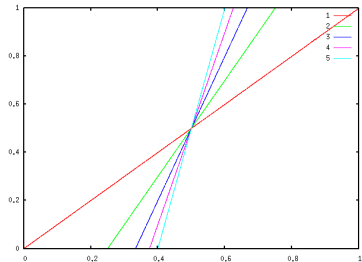
- 1 Cas non-asymptotique, La borne est meilleure que celle de l'apprentissage passif si la **moyenne des variances locales est plus faible que la variance globale**.
- 2 Cas asymptotique, on parvient a **supprimer la variance inter-strates**, conservant seulement la variance intra-strates. On est **équivalent au à** $\sup_f |\hat{L}(f) - L(f)|$.



Validation Exp rimentale

Protocole

- Soit $p \in \llbracket 1, 5 \rrbracket$. La t che objectif est une classification pour $x \in [0, 1]$. La classe de x est 1 avec une probabilit  $\frac{1}{2} + p(x - \frac{1}{2})$ (tronqu e entre 0 et 1 si n cessaire) et 0 sinon.
- On apprend avec $200 + 5000$ exemples par moindres carr s, stratification en 3 sous-espaces connexes.



Validation Exp rimentale

Protocole

- Soit $p \in \llbracket 1, 5 \rrbracket$. La t che objectif est une classification pour $x \in [0, 1]$. La classe de x est 1 avec une probabilit  $\frac{1}{2} + p(x - \frac{1}{2})$ (tronqu e entre 0 et 1 si n cessaire) et 0 sinon.
- On apprend avec 200 + 5000 exemples par moindres carr s, stratification en 3 sous-espaces connexes.

R sultats

p	1	2	3	4	5
% stratification	44	50.5	52	44	46.5
% <i>ex aequo</i>	11	12.5	23.5	29.5	30
% na�ve	45	37	24.5	26.5	23.5

param�tre p	% erreur m�diane meilleure avec stratification
$p = 1$	51.25 %
$p = 2$	63 %
$p = 3$	53.5 %
$p = 4$	66.5 %
$p = 5$	76.75 %

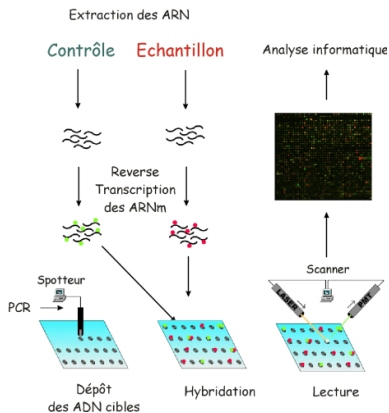


Plan

- 1 Introduction
- 2 Exemples a priori : Discrépance
- 3 Stratification de la base d'apprentissage
- 4 Sélection d'Attributs**
- 5 Conclusion, Perspectives



Principe des puces   ADN



Caract ristiques des donn es

- 1   Tr s **grande dimension** > 6000
- 2   Peu d'exemples (prix  lev )
- 3   Peu fiables
- 4   Concept **cible ne faisant pas intervenir** tous les  l ments

[NAR-04] G. Mercier, N. Berthault, J. Mary, A. Antoniadis, J-P. Comet, A. Cornu jols, Ch. Froidevaux and M. Dutreix.
Comparing and combining feature estimation methods for the analysis of microarray data.
Nucleic Acids Research (NAR), vol.32, No.1, 1-8 (2004).



S lection d'attributs

X_1	X_2	...	X_N	Classe
$x_{1,1}$	$x_{1,2}$...	$x_{1,N}$	●
$x_{2,1}$	$x_{2,2}$...	$x_{2,N}$	▲
⋮	⋮		⋮	⋮
$x_{K,1}$	$x_{K,2}$...	$x_{K,N}$	●



S lection d'attributs

X_1	X_2	...	X_N	Classe
$x_{1,1}$	$x_{1,2}$...	$x_{1,N}$	●
$x_{2,1}$	$x_{2,2}$...	$x_{2,N}$	▲
⋮	⋮		⋮	⋮
$x_{K,1}$	$x_{K,2}$...	$x_{K,N}$	●

Obtention d'un ordre sur les attributs :

$X_{10} X_3 X_1 X_{60} X_{40} X_7 X_5 \dots X_{45} X_{11} X_{32} X_4$



Sélection d'attributs

X_1	X_2	...	X_N	Classe
$x_{1,1}$	$x_{1,2}$...	$x_{1,N}$	●
$x_{2,1}$	$x_{2,2}$...	$x_{2,N}$	▲
⋮	⋮		⋮	⋮
$x_{K,1}$	$x_{K,2}$...	$x_{K,N}$	●

Choix d'un seuil :

$$\underbrace{X_{10} \ X_3 \ X_1 \ X_{60} \ X_{40} \ X_7}_{\text{pertinents}(P)} \quad \underbrace{X_5 \ \dots \ X_{45} \ X_{11} \ X_{32} \ X_4}_{\text{non pertinents}(N)}$$

$$P = VP + FP$$

$$N = VN + FN$$



Sélection d'attributs

X_1	X_2	...	X_N	Classe
$x_{1,1}$	$x_{1,2}$...	$x_{1,N}$	●
$x_{2,1}$	$x_{2,2}$...	$x_{2,N}$	▲
⋮	⋮		⋮	⋮
$x_{K,1}$	$x_{K,2}$...	$x_{K,N}$	●

Choix d'un seuil :

$X_{10} X_3 X_1 X_{60} X_{40} X_7$ $X_5 \dots X_{45} X_{11} X_{32} X_4$
 pertinents(P) non pertinents(N)

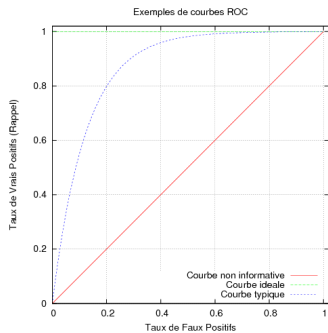
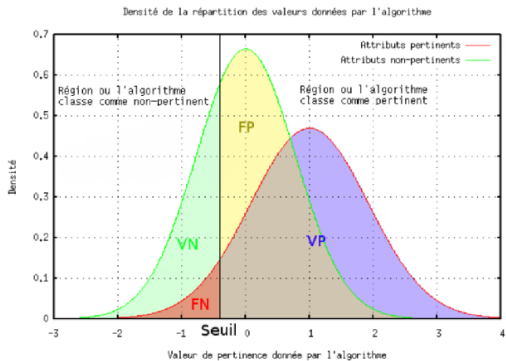
$$P = VP + FP$$

$$N = VN + FN$$

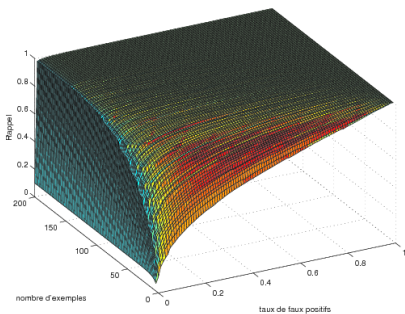
PLUS ON A D'EXEMPLES PLUS LE CLASSEMENT EST BON MAIS NOTRE EXPERT SOUHAITE FAIRE LE MOINS DE MESURES POSSIBLE.



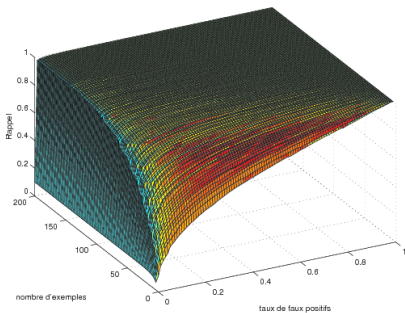
R partition des attributs



Courbes ROC



Courbes ROC



Formules

$$\underline{VP}_s = \Delta \cdot \prod_{i=1}^s r_i$$

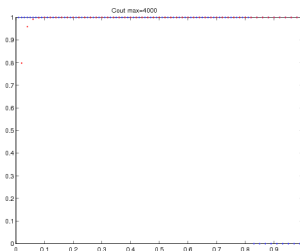
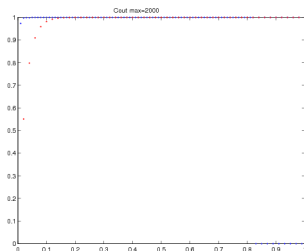
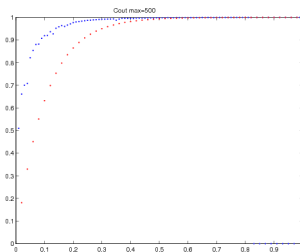
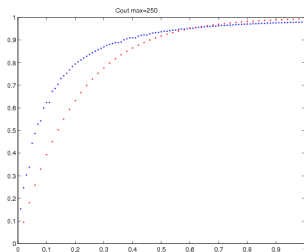
$$\underline{VN}_s = \sum_{i=1}^s (1 - \tau_i) \left(\prod_{j=1}^{i-1} \tau_j \right) (N_0 - \Delta)$$

$$\underline{FP}_s = \left(\prod_{i=1}^s \tau_i \right) (N_0 - \Delta)$$

$$\underline{FN}_s = \sum_{i=1}^s \left(\prod_{j=1}^i (1 - r_j) \right) \Delta$$



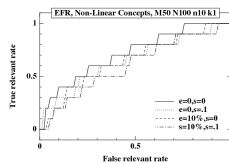
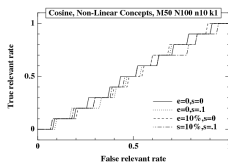
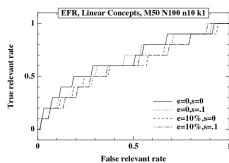
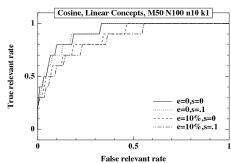
Optimisation



Ensemble Feature Ranking

Principe

- Probl me d fini par des param tres d'ordre :
 - Le nombre d'exemples n ;
 - Le nombre total d'attributs d ;
 - Le nombre d'attributs pertinents r .
 - La redondance ou non k
 - Le bruit σ
- Strat gie  volutionnaire pour obtenir **plusieurs classeurs**.



Plan

- 1 Introduction
- 2 Exemples a priori : Discrépance
- 3 Stratification de la base d'apprentissage
- 4 Sélection d'Attributs
- 5 Conclusion, Perspectives**



Bilan

Points donnés a priori

- **Discrépance induit une borne sup sur l'erreur. De plus on observe une corrélation avec l'erreur.**
- En optimisant la vitesse de convergence est quasi-identique au cas où le concept cible est dans F (intéressant pour les applications réelles).

Stratification

- Approches à étapes basées sur la variance de l'erreur;

• Méthodes de régression linéaire.

Sélection d'attributs



Bilan

Points donn s a priori

- Discr ance induit une borne sup sur l'erreur. De plus on observe une corr lation avec l'erreur.
- **En optimisant la vitesse de convergence est quasi-identique au cas o  le concept cible est dans F (int ressant pour les applications r elles).**

Stratification

- Approches    tapes bas es sur la variance de l'erreur ;
- Vitesses de convergences am lior es .

S lection d'attributs



Bilan

Points donnés a priori

- Discrépance induit une borne sup sur l'erreur. De plus on observe une corrélation avec l'erreur.
- En optimisant la vitesse de convergence est quasi-identique au cas où le concept cible est dans F (intéressant pour les applications réelles).

Stratification

- **Approches à étapes basées sur la variance de l'erreur ;**
- Vitesses de convergences améliorées .

Sélection d'attributs

- Optimisation de la sélection en réalisant plusieurs passes ;



Bilan

Points donnés a priori

- Discrédance induit une borne sup sur l'erreur. De plus on observe une corrélation avec l'erreur.
- En optimisant la vitesse de convergence est quasi-identique au cas où le concept cible est dans F (intéressant pour les applications réelles).

Stratification

- Approches à étapes basées sur la variance de l'erreur ;
- **Vitesses de convergences améliorées .**

Sélection d'attributs

- Optimisation de la sélection en réalisant plusieurs passes ;
- Bagging d'ordres.



Bilan

Points donn s a priori

- Discr pance induit une borne sup sur l'erreur. De plus on observe une corr lation avec l'erreur.
- En optimisant la vitesse de convergence est quasi-identique au cas o  le concept cible est dans F (int ressant pour les applications r elles).

Stratification

- Approches    tapes bas es sur la variance de l'erreur ;
- Vitesses de convergences am lior es .

S lection d'attributs

- **Optimisation de la s lection en r alisant plusieurs passes ;**
- Bagging d'ordres.



Bilan

Points donn s a priori

- Discr pance induit une borne sup sur l'erreur. De plus on observe une corr lation avec l'erreur.
- En optimisant la vitesse de convergence est quasi-identique au cas o  le concept cible est dans F (int ressant pour les applications r elles).

Stratification

- Approches    tapes bas es sur la variance de l'erreur ;
- Vitesses de convergences am lior es .

S lection d'attributs

- Optimisation de la s lection en r alisant plusieurs passes ;
- **Bagging d'ordres.**



Perspectives

Liens avec le boosting

- **Faire de l'actif comme du boosting.**
- Essayer de perturber la distribution optimale conseill e par l' tude de la variance
- Voir s'il existe un lien entre l'am lioration (ou la variance de l'am lioration) fournie par le boosting et la discr ance de la base initiale.

Challenge Pascal

- Essayer de borner $P(\text{Card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > C)$ au lieu de $P(\text{card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > 0)$ qui est le classique :
 $P(\exists f \in F; |\hat{L}(f) - L(f)| > \varepsilon)$
- Utile pour estimer la probabilit  pour que **au moins C g nes soient sur-exprim s de ε** au niveau de leur diff rence d'expression entre les individus sains et les individus malades.



Perspectives

Liens avec le boosting

- Faire de l'actif comme du boosting.
- **Essayer de perturber la distribution optimale conseill e par l' tude de la variance**
- Voir s'il existe un lien entre l'am lioration (ou la variance de l'am lioration) fournie par le boosting et la discr panance de la base initiale.

Challenge Pascal

- Essayer de borner $P(\text{Card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > C)$ au lieu de $P(\text{card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > 0)$ qui est le classique :
 $P(\exists f \in F; |\hat{L}(f) - L(f)| > \varepsilon)$
- Utile pour estimer la probabilit  pour que **au moins C g nes soient sur-exprim s de ε** au niveau de leur diff rence d'expression entre les individus sains et les individus malades.



Perspectives

Liens avec le boosting

- Faire de l'actif comme du boosting.
- Essayer de perturber la distribution optimale conseill e par l' tude de la variance
- **Voir s'il existe un lien entre l'am lioration (ou la variance de l'am lioration) fournie par le boosting et la discr panance de la base initiale.**

Challenge Pascal

- Essayer de borner $P(\text{Card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > C)$ au lieu de $P(\text{card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > 0)$ qui est le classique :
 $P(\exists f \in F; |\hat{L}(f) - L(f)| > \varepsilon)$
- Utile pour estimer la probabilit  pour que **au moins C g nes soient sur-exprim s de ε** au niveau de leur diff rence d'expression entre les individus sains et les individus malades.



Perspectives

Liens avec le boosting

- Faire de l'actif comme du boosting.
- Essayer de perturber la distribution optimale conseill e par l' tude de la variance
- Voir s'il existe un lien entre l'am lioration (ou la variance de l'am lioration) fournie par le boosting et la discr pance de la base initiale.

Challenge Pascal

- Essayer de borner $P(\text{Card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > C)$ au lieu de $P(\text{card}\{f \in F; |\hat{L}(f) - L(f)| > \varepsilon\} > 0)$ qui est le classique :
 $P(\exists f \in F; |\hat{L}(f) - L(f)| > \varepsilon)$
- Utile pour estimer la probabilit  pour que **au moins C g nes soient sur-exprim s de ε** au niveau de leur diff rence d'expression entre les individus sains et les individus malades.



Questions



Publications

Chapter / Book

- Mary J., Mercier G., Comet J-P., Cornuéjols A., Froidevaux Ch. & Dutreix M. An attribute estimation technique for the analysis of microarray data. In Proc. of the Dieppe Spring school on Modelling and simulation of biological processes in the context of genomics (eds. P. Amar, F. Képès, V. Norris and P. Tracqui) Publisher Frontier group, ISBN : 2 91 4601 09 3, 2003 .

International Journal

- G. Mercier, N. Berthault, J. Mary, J. Peyre, A. Antoniadis, J.-P. Comet , A. Cornuéjols, Ch. Froidevaux & M. Dutreix, Biological detection of low radiation by combining results of two microarray analysis methods, Nucleic Acids Research, 32(1) :e12, 2004.

International Conference

- Jong K., Mary J., Cornuéjols A., Marchiori E. & Sebag M. Ensemble feature ranking. Proc. of the Conf. "Current trends in drug discovery research" (ECML/PKDD-2004), Pisa, Italy, Sept. 20-24, 2004.
- Claude M. Dion, Eric Cancès, & Jérémie Mary, Efficient algorithms for the time-dependent Gross-Pitaevskii equation with harmonic potentials , Cold Molecules 2002 : Ultracold Molecules and Bose-Einstein Condensation conference, Les Houches, France, 4-8 March 2002



Publications

National Conference

- Jérémie Mary, Géraldine Mercier, Jean-Paul Comet, Antoine Cornuéjols, Christine Froidevaux et Marie Dutreix Utilisation d'une méthode de sélection d'attributs pour l'analyse du transcriptome de cellules de levure exposées à de faibles doses de radiation in Informatique pour l'analyse du transcriptome, troisièmes journées de Post-Génomique de la DOUA, Lyon, 14 Mai 2003.
- Sylvain Gelly, Jérémie Mary, Olivier Teytaud : Taylor-based pseudo-metrics for random process fitting in dynamic programming : expected loss minimisation and risk management. Processus décisionnels de Markov et Intelligence Artificielle (PDM et IA), Lille juin 2005.

Workshop / Poster

- Cornuéjols A., Froidevaux Ch. and Mary J. : Comparing and combining feature estimation methods for the analysis of microarray data. JOBIM-05.
- Cornuéjols A., Sebag M., Mary, J. : "Classification d'images à l'aide d'un codage par motifs fréquents" (Image classification using a frequent item sets coding). Workshop sur la fouille d'images (RFIA-04), Toulouse, janvier 2004.
- Sylvain Gelly, Jérémie Mary, Olivier Teytaud : Taylor-based pseudo-metrics for random process fitting in dynamic programming : expected loss minimisation and risk management. CAP, Nice juin 2005.



Coefficient de Pulvérisation et VC dim

Coefficient de pulvérisation

Étant donné F , une famille de fonctions de $X \rightarrow Y = \{0, 1\}$, le $n^{\text{ème}}$ **coefficient de pulvérisation** de F , est :

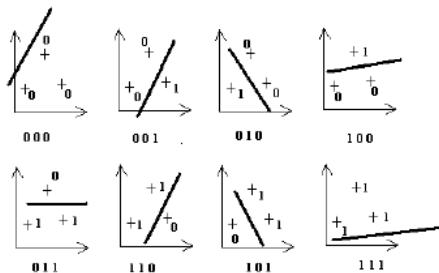
$$S(F, n) = \sup_{(x_1, \dots, x_n) \in X^n} \text{Card} \{(f(x_1), \dots, f(x_n)) \mid f \in F\}$$

VC-dim

Étant donné F une famille de fonctions de $X \rightarrow Y = \{0, 1\}$, la **VC-dimension** de F est, s'il existe, le plus grand entier n , tel que $S(F, n) = 2^n$. Sinon, la VC-dimension est infinie.



Coefficient de Pulvérisation et VC dim



Théorème de VC

Vapnik-Chervonenkis

Pour toute mesure μ et pour \mathcal{A} , $\forall \varepsilon > 0$ si $S(F, n)$ désigne le $n^{\text{ème}}$ coefficient de pulvérisation de F :

$$P(\sup_{f \in F} |\mu_n(f) - \mu(f)| > \varepsilon) \leq 8S(F, n)e^{-n\varepsilon^2/32}$$

Lemme de Sauer

Si la VC dim de F est majorée par V alors :

$$S(F, n) \leq \sum_{i=0}^V C_n^i < \left(\frac{en}{V}\right)^V ; \forall n \geq V \geq 1$$

Convergence en $\frac{1}{\sqrt{n}}$ si la VC-dim est finie



Exemple

Soit (x_1, \dots, x_n) un échantillon i.i.d. et $F = (f_1, \dots, f_d)$

- 1 Soit M le vecteur des moyennes empiriques des f_i .
 $M = (M_1, \dots, M_d)$ avec $M_i = (f_i(x_1) + \dots + f_i(x_n))/n$
- 2 Soit D le vecteur des écarts de la moyenne empirique à l'espérance
 $D = (M_1 - E(M_1), \dots, M_d - E(M_d))$
- 3 D tend ps vers 0 par la loi des grands nombres
- 4 Le TCL (s'il existe un moment d'ordre 2) donne la vitesse de convergence $(1/\sqrt{n})$



Classe de Glivenko-Cantelli

Classe de Glivenko-Cantelli

Étant donnée L_F une famille d'applications de $(X, Y) \rightarrow \mathbb{R}$, et \mathcal{P} une famille de distributions sur (X, Y) , On dit que L_F est faiblement (resp. fortement) **\mathcal{P} -Glivenko-Cantelli** pour la convergence faible (resp. pour la convergence p.s.) si $\forall P \in \mathcal{P}$:

$$\sup_{L_f \in L_F} \left| \frac{1}{n} \sum_{i=1}^n L_f(z_i) - E_P(L_f) \right| \xrightarrow{n \rightarrow \infty} 0$$



Donsker

Étant donné une famille L_F de fonctions de (X, Y) dans \mathbb{R} , on dit que L_F est **\mathcal{P} -Donsker** si :

- 1 $\forall P \in \mathcal{P}, \sup_{L_f \in L_F} |L_f(x) - E_P(L_f)| < \infty$
- 2 avec $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_i$ et $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, pour tout $P \in \mathcal{P}$, \mathbb{G}_n converge faiblement dans $l_\infty(L_F)$ vers un certain \mathbb{G} tendu mesurable.

Permet de traiter le cas de nombre de couvertures exponentiels



Démonstration

- On établit quelle est la variance associée à une allocation donnée des exemples aux différentes strates ;
- Une forme spécialisée de l'apprentissage PAC montre que nous pouvons nous restreindre à considérer des hypothèses "assez bonnes".
- On utilise l'inégalité de Chernoff pour tirer parti de la variance optimisée grâce à la stratification ;
- On établit grâce à Chernoff que l'algorithme améliore la vitesse de convergence par rapport à une base d'apprentissage i.i.d.
- Conclusion.



Bagging

Principe

- Améliore les performances d'un ensemble de **weak-learners** indépendants en les faisant voter ;
- Dans le cas où l'on a l'indépendance les résultats sont garantis par Chernoff (borne exponentielle) ;
- Sinon il est nécessaire que l'apprentissage soit **"suffisamment" instable**.



Bagging

Principe

- Améliore les performances d'un ensemble de **weak-learners** indépendants en les faisant voter ;
- Dans le cas où l'on a l'indépendance les résultats sont garantis par Chernoff (borne exponentielle) ;
- Sinon il est nécessaire que l'apprentissage soit **"suffisamment" instable**.



Bagging

Principe

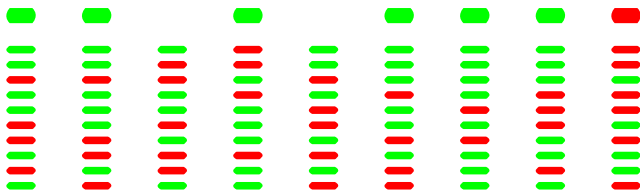
- Améliore les performances d'un ensemble de **weak-learners** indépendant en les faisant voter ;
- Dans le cas où l'on a l'indépendance les résultats sont garantis par Chernoff (borne exponentielle) ;
- Sinon il est nécessaire que l'apprentissage soit **"suffisamment" instable**.



Bagging

Principe

- Améliore les performances d'un ensemble de **weak-learners** indépendant en les faisant voter ;
- Dans le cas où l'on a l'indépendance les résultats sont garantis par Chernoff (borne exponentielle) ;
- Sinon il est nécessaire que l'apprentissage soit **"suffisamment" instable**.



Bagging

Principe

- Améliore les performances d'un ensemble de **weak-learners** indépendants en les faisant voter ;
- Dans le cas où l'on a l'indépendance les résultats sont garantis par Chernoff (borne exponentielle) ;
- Sinon il est nécessaire que l'apprentissage soit **"suffisamment" instable**.



Sélection d'attributs

Différentes techniques

- les approches **embarquées** c'est l'algorithme d'apprentissage lui même qui effectue la sélection.
- Les **filtres** (RELIEF, FOCUS...) : la sélection se fait indépendamment de l'apprentissage
- Les **wrappers**, la précision de l'apprentissage en validation croisée est utilisée pour évaluer la qualité de la sélection.



Comparasion au hasard

Interpréter les résultats

- Les algorithmes de sélection donnent des seuils
- On **mélange les étiquettes** et on regarde leur comportement.

