Nouvelles Approches Itératives avec Garanties Théoriques pour l'Adaptation de Domaine Non Supervisée

Jean-Philippe Peyrache

Rapporteurs : Antoine Cornuéjols et Jean-Christophe Janodet
Examinatrices : Elisa Fromont et Mikaela Keller
Directeurs : Marc Sebban et Amaury Habrard

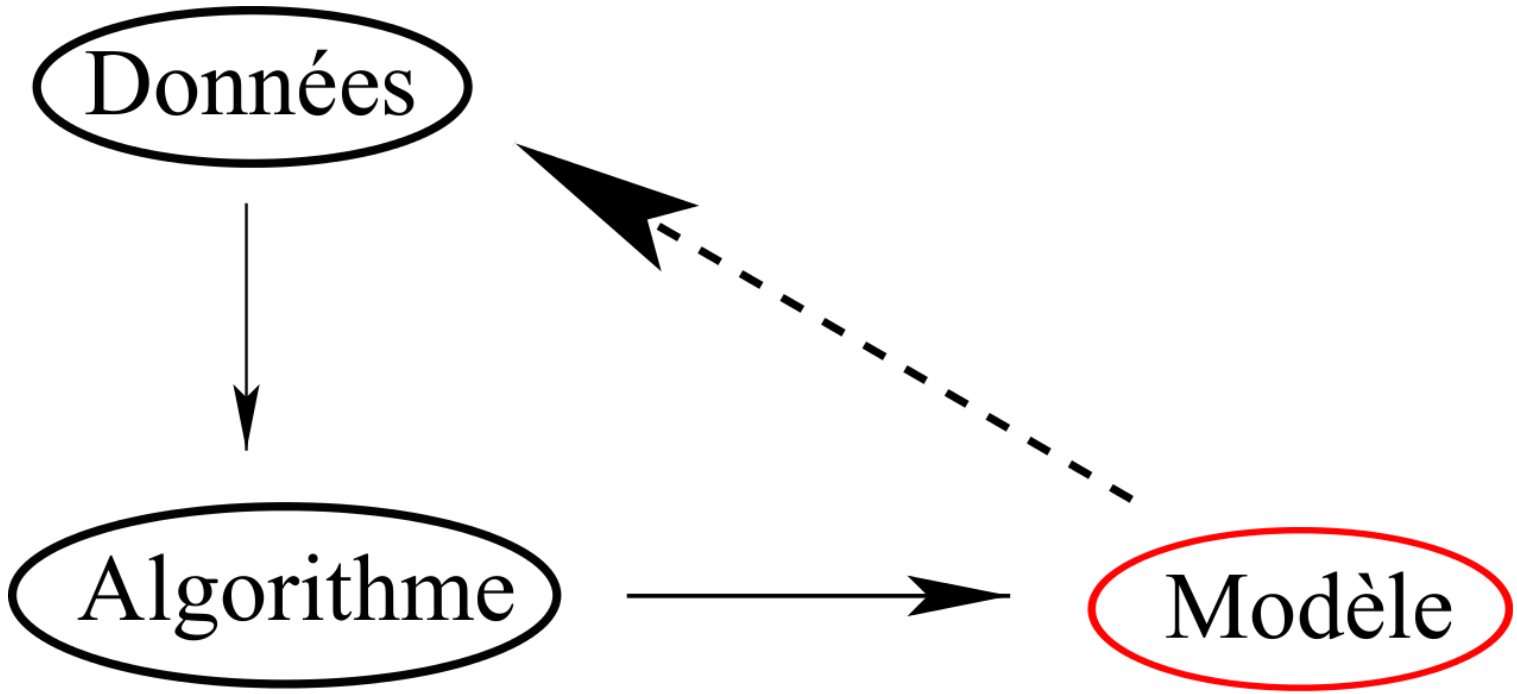
Laboratoire Hubert Curien, Université de Saint-Etienne, FRANCE

Soutenance de thèse : 11 juillet 2014

Apprentissage automatique

Définition

L'apprentissage automatique *définit un cadre permettant la conception d'algorithmes dont l'objectif est de réaliser des modèles automatisant certaines tâches : reconnaissance de caractères manuscrits, détection de spams. . .*



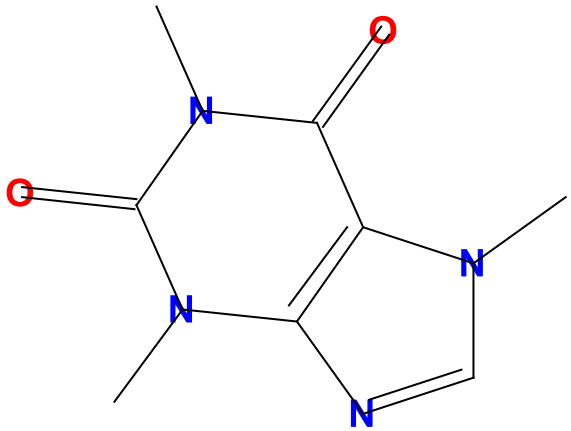
Apprentissage automatique

Représentation des données

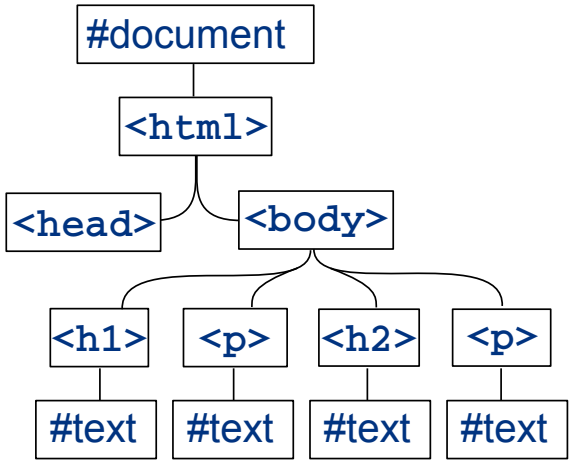
Les données peuvent prendre diverses formes : vecteurs numériques ou données structurées.

$$\begin{pmatrix} \text{Âge} & 25 \\ \text{Poids} & 76 \\ \vdots & \vdots \\ \text{Tension} & 14 \end{pmatrix}$$

Patient



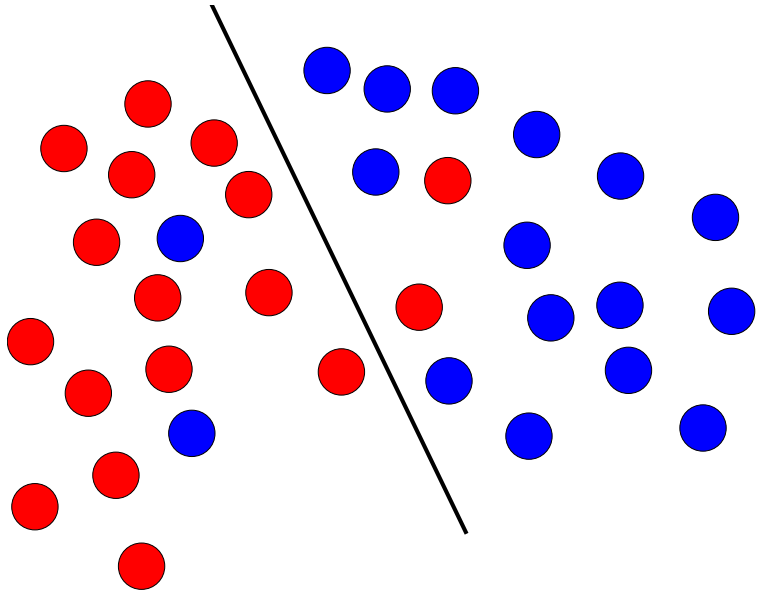
Molécule



Document HTML

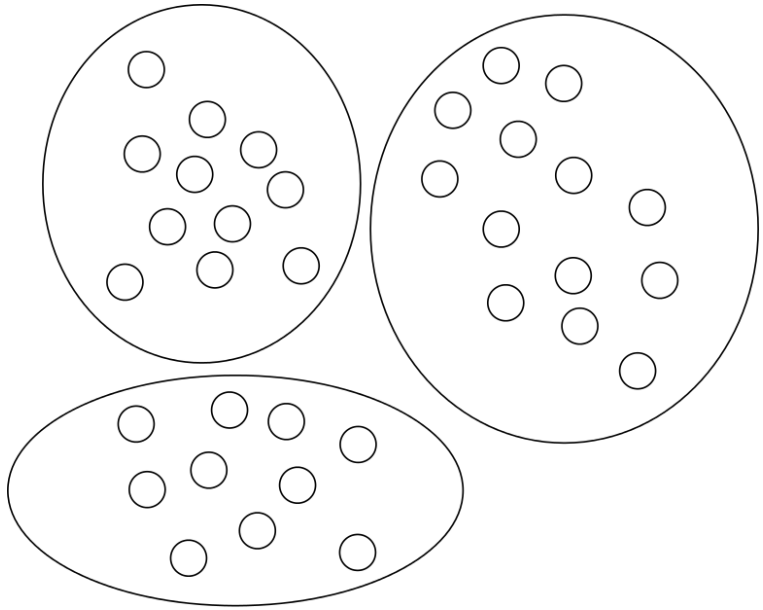
Apprentissage automatique

Apprentissage supervisé



Tâche de classification

Apprentissage non supervisé

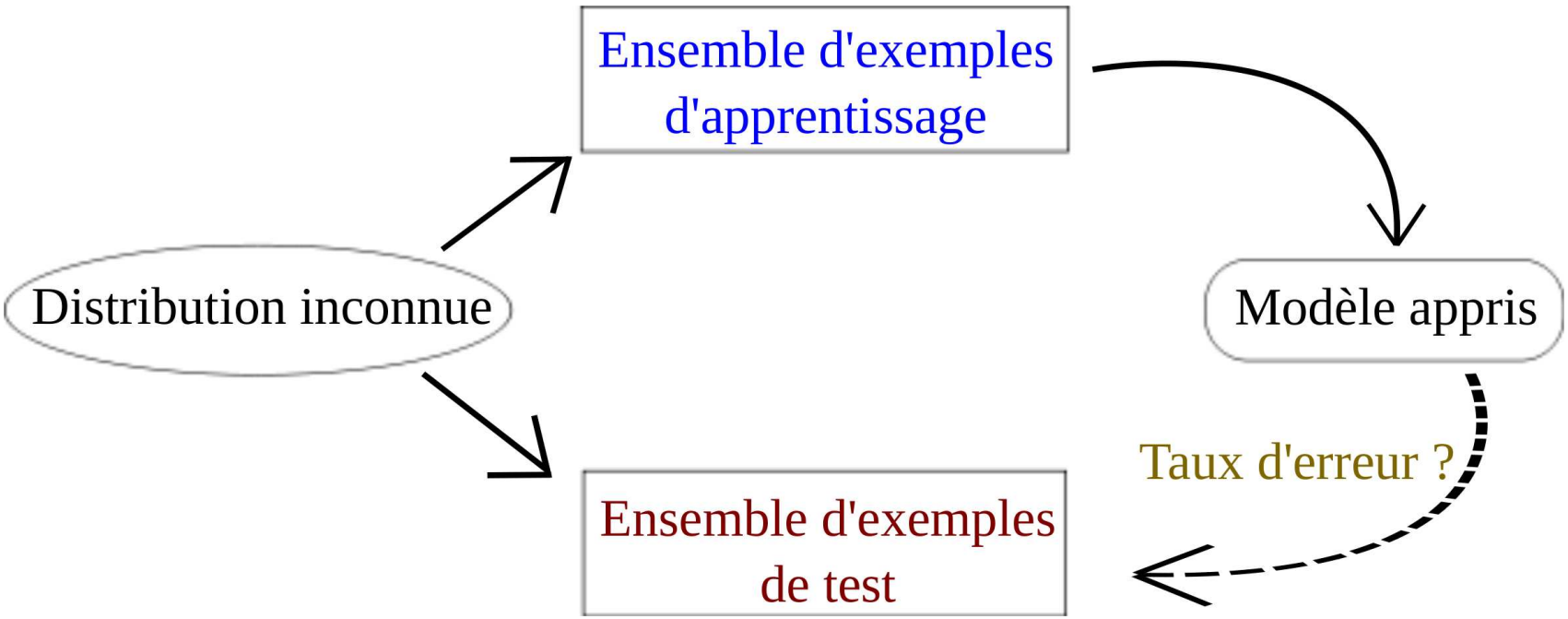


Tâche de *clustering*

Cadre mixte

L'apprentissage semi-supervisé consiste à apprendre un modèle depuis un ensemble contenant à la fois des exemples étiquetés (en faible quantité) et des exemples non étiquetés (en grand nombre).

Apprentissage automatique



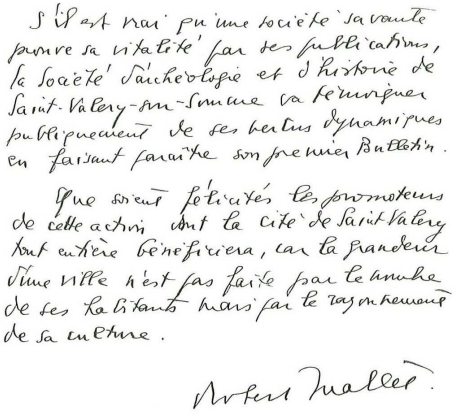
Adaptation de domaine

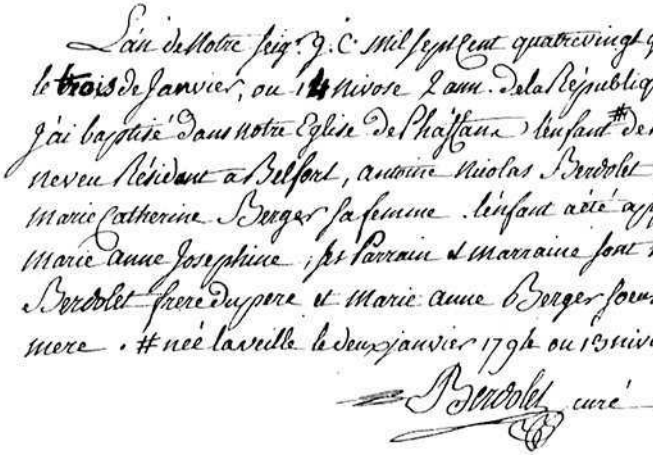
Problème

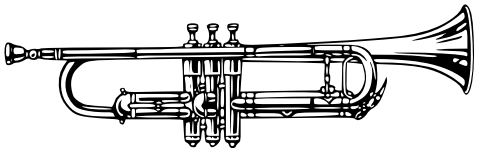
Le modèle est performant si les données *d'apprentissage* et *de test* sont issues de **la même distribution**, ce qui n'est pas vrai dans de nombreux des cas réels.
Ces situations nécessitent un nouveau cadre :

l'Adaptation de Domaine (AD)

Adaptation de domaine







Données d'apprentissage **Données de test**

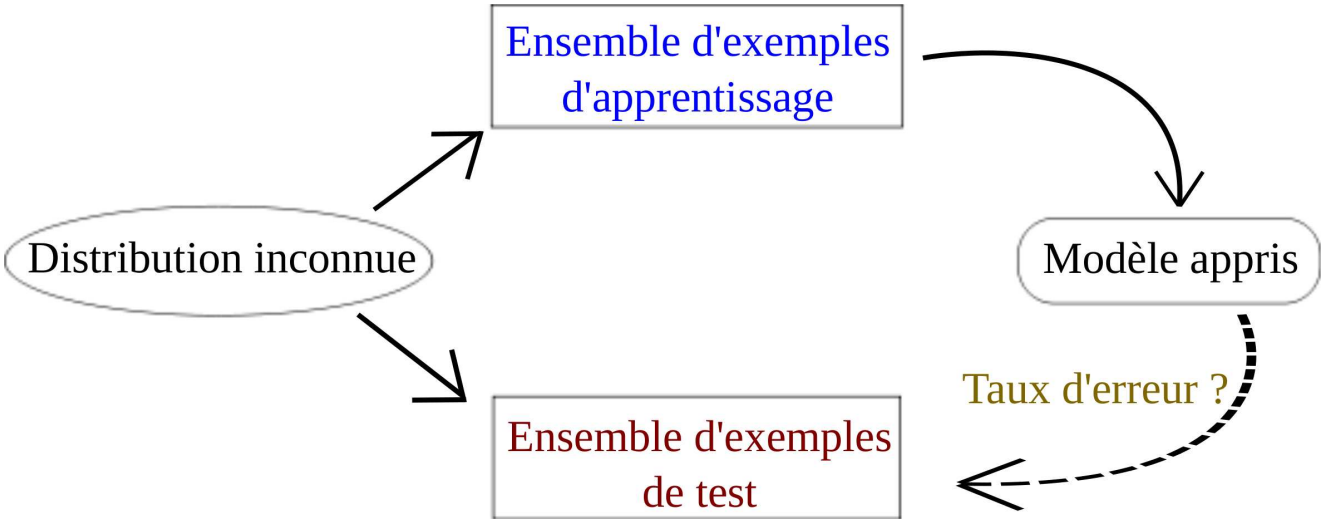
Sommaire

Notations et Pré-requis

Apprentissage supervisé

Définition (Ensemble d'apprentissage)

Ensemble $S = \{z_i = (x_i, y_i)\}_{i=1}^{m}$ *de* m *exemples i.i.d. selon une distribution* $\mathcal{D}_S$ *sur* $Z = X \times Y$. *Dans ce travail,* $Y = \{-1, +1\}$.



Définition (Hypothèse)

La fonction apprise $h \in \mathcal{H}$, *depuis un ensemble d'apprentissage* S, *cherche à prédire au mieux* $y \in Y$ *à partir de* $x \in X$, *pour tout* $(x, y) \sim \mathcal{D}_S$.

Notion d'erreur

Définition (Erreur réelle)

L'erreur réelle $\epsilon^{\ell}_{\mathcal{D}_S}$ d'une hypothèse h selon une fonction de perte ℓ sur une distribution $\mathcal{D}_S$ correspond à :

$$\epsilon^{\ell}_{\mathcal{D}_S}(h) = \mathbb{E}_{z\sim\mathcal{D}_S}[\ell(h,z)].$$

Définition (Erreur empirique)

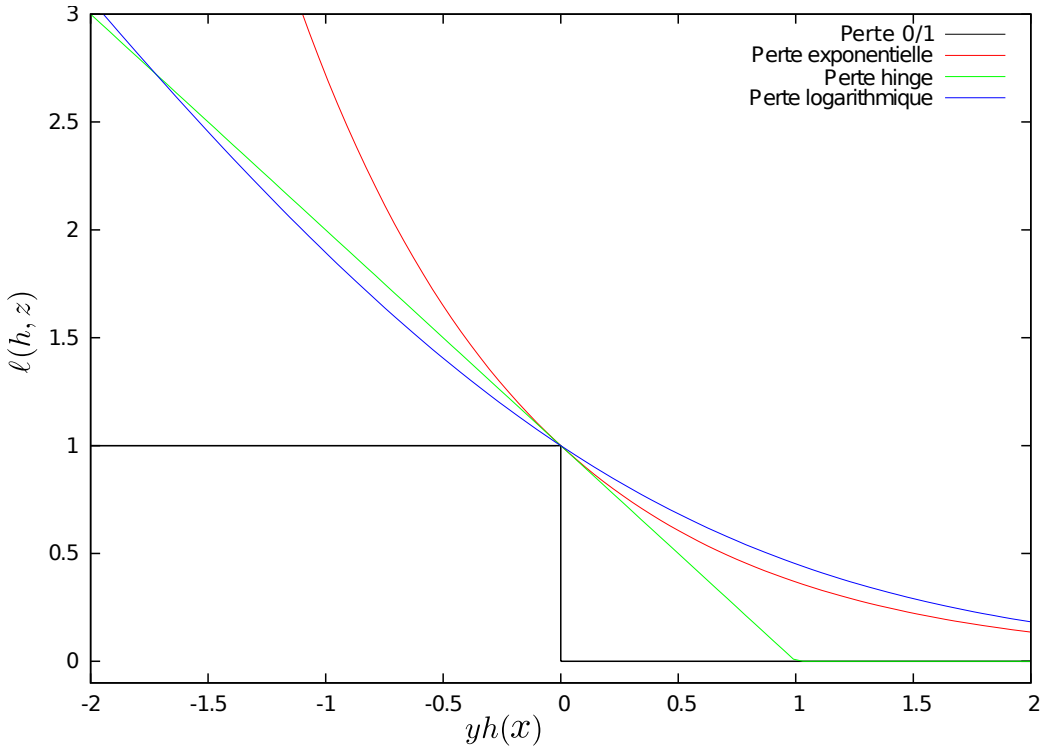
L'erreur empirique ϵ^{ℓ}_{S} d'une hypothèse h selon une fonction de perte ℓ sur un ensemble d'apprentissage $S = \{z_i\}_{i=1}^{m}$ correspond à :

$$\epsilon^{\ell}_{S}(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i).$$

Fonctions de perte

Définition ($\ell_{0/1}$)

$$\ell_{0/1}(h, z) = \begin{cases} 1 & \text{si } yh(x) < 0 \\ 0 & \text{sinon.} \end{cases}$$



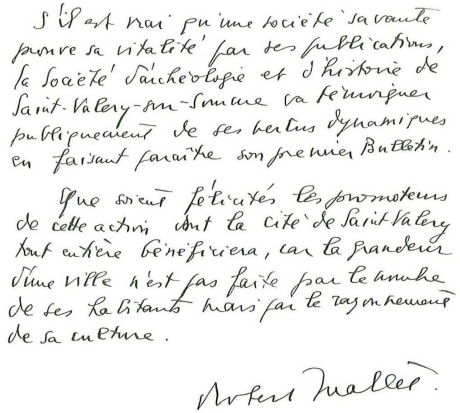
Le modèle PAC

Bornes PAC [Valiant, 1984]

Le cadre PAC permet la dérivation de bornes de la forme suivante :

$$Pr[|\epsilon^{\ell}_{\mathcal{D}_S}(h) - \epsilon^{\ell}_{S}(h)| < \mu] \geq 1 - \delta,$$

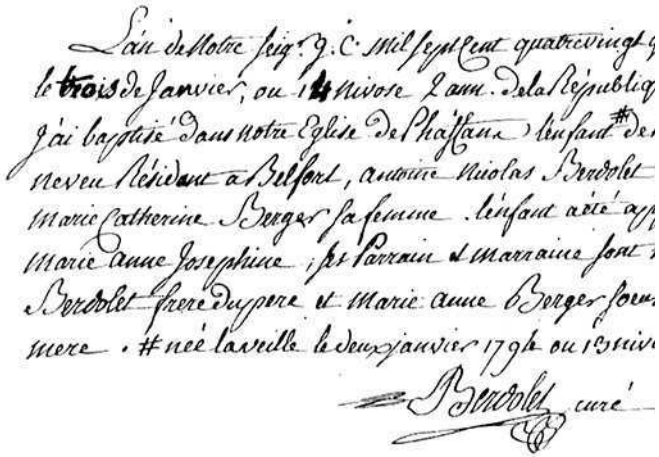
avec $\mu \geq 0$ et $\delta \in [0, 1]$, où les données sont i.i.d. selon $\mathcal{D}_S$.



S'il est vrai qu'une société savante puise sa vitalité par ses publications, la Société d'archéologie et d'histoire de Saint-Valery-sur-Somme va le montrer publiquement de ses beaux dynamismes en faisant paraître son premier Bulletin.

Que soient félicités les promoteurs de cette action dont la cité de Saint-Valery tout entière bénéficiera, car la grandeur d'une ville n'est pas faite par le nombre de ses habitants mais par le rayonnement de sa culture.

Robert Mallet.

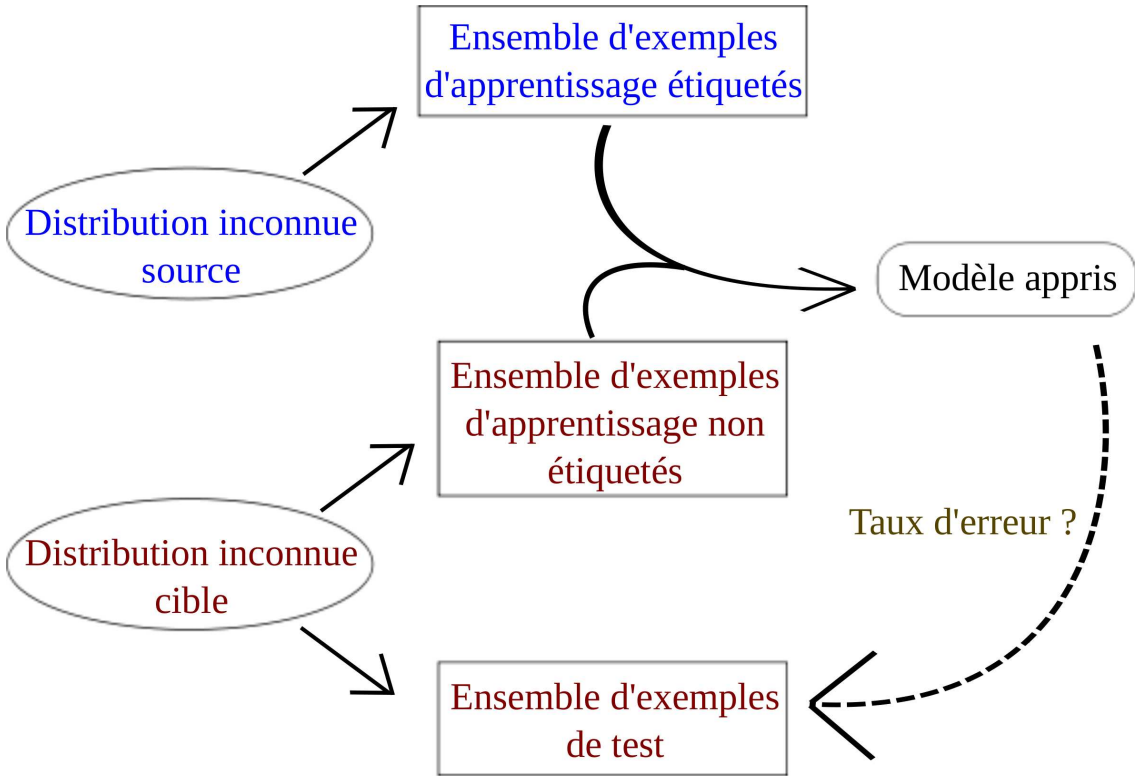


L'an de notre Seig.r J.C. Mil sept cent quatre-vingt
le trois de Janvier, ou 14 Nivose 2 ann. de la Républiq
j'ai baptisé dans notre Eglise de Chaffaux Jean fils de
neveu Résidant à Belfort, autorisé Nicolas Berdolet
Marie Catherine Berger sa femme. L'enfant né le
Marie Anne Josephine ; le Parrain et Marraine sont
Berdolet frere du pere et Marie Anne Berger sœur
mere. # née la veille le deux janvier 1794 ou 13 Nivo
Berdolet curé

Adaptation de domaine

Définition (Ensemble d'apprentissage en AD)

Un ensemble d'apprentissage de taille m dans le cas de l'AD est un ensemble $S \cup T$ composé de deux sous-ensembles $S = \{z_i = (x_i, y_i)\}_{i=1}^{|S|}$ i.i.d. selon $\mathcal{D}_S$ et $T = \{x_j\}_{j=1}^{|T|}$ i.i.d. selon $\mathcal{D}_T^X$, tels que $|S| + |T| = m$.



Adaptation de domaine

Bornes en généralisation [Ben-David et al., 2010]

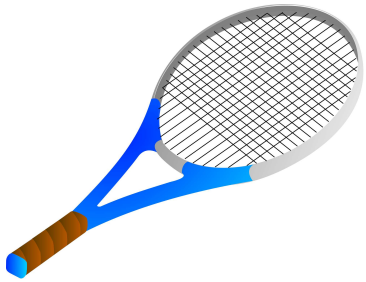
Généralisation du cadre PAC, en intégrant une mesure de divergence :

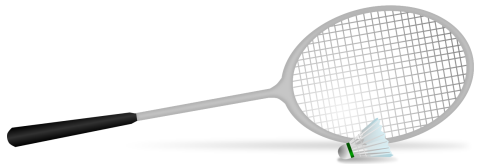
$$\epsilon_{\mathcal{D}_T}(h) \leq \epsilon_{\mathcal{D}_S}(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S,T) + 4\sqrt{\frac{2d\log(2m) + \log\frac{2}{\delta}}{m}} + \lambda,$$

où $\lambda = \epsilon_{\mathcal{D}_S}(h^\star) + \epsilon_{\mathcal{D}_T}(h^\star)$.

Terme incompressible λ, apportant une information théorique importante :

- λ trop élevé $\rightarrow$ Adaptation impossible.
- λ trop faible $\rightarrow$ Adaptation non nécessaire.





Adaptation de domaine

$$\epsilon_{\mathcal{D}_T}(h) \leq \epsilon_{\mathcal{D}_S}(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(S,T) + 4\sqrt{\frac{2d\log(2m) + \log\frac{2}{\delta}}{m}} + \lambda.$$

L'idée

Pour obtenir une faible erreur en généralisation, la borne suggère :

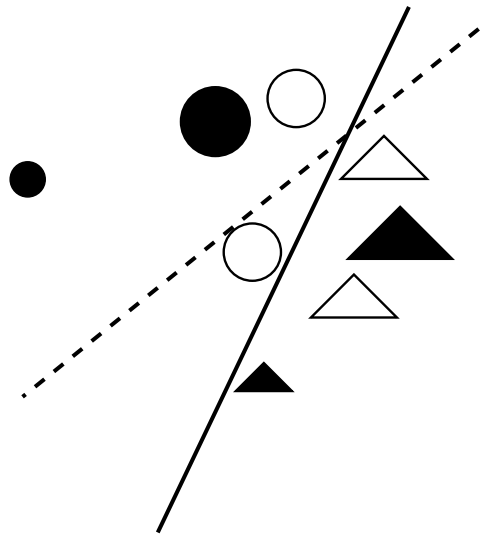
- d'obtenir une faible erreur sur la source
- de minimiser la divergence entre S et T

Adaptation de domaine

Approches algorithmiques

Trois principales catégories d'approches :

- Les méthodes de repondération.
-
-

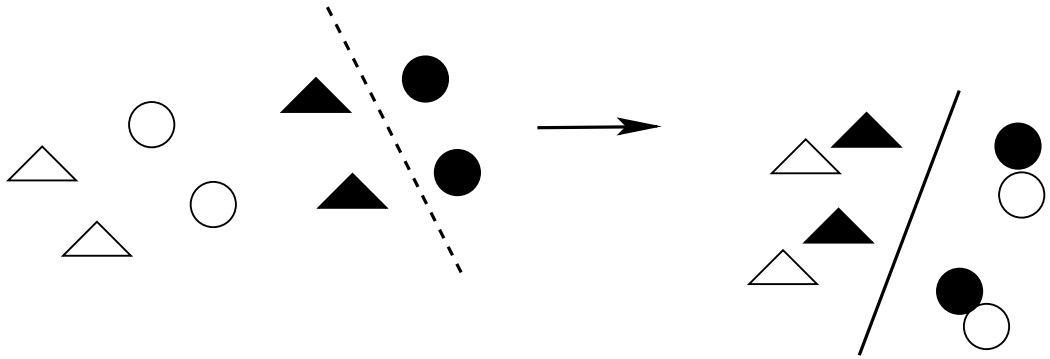


Adaptation de domaine

Approches algorithmiques

Trois principales catégories d'approches :

-
- Les méthodes de reprojection
-

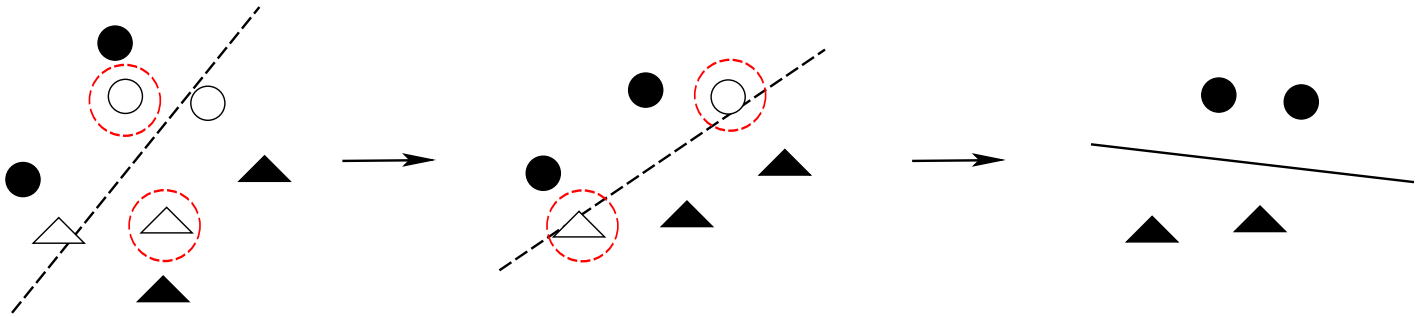


Adaptation de domaine

Approches algorithmiques

Trois principales catégories d'approches :

-
-
- Les approches itératives d'auto-étiquetage



Contributions

Contributions de cette thèse

1. La première, approche de reprojection basée sur le boosting, travaille sur les données vectorielles.
2. La seconde, approche d'auto-étiquetage, s'affranchit de certaines contraintes permettant un traitement plus direct des données structurées.

Méthode Ensembliste pour l'Adaptation de Domaine

Publications

- Habrard A., Peyrache J-P., Sebban M.
  Boosting for Unsupervised Domain Adaptation
  ECML/PKDD 2013, Proceedings part II, 433-448, **2013**

- Habrard A., Peyrache J-P., Sebban M.
  Un Cadre Formel de Boosting pour l'Adaptation de Domaine
  CAp' 2012, 1-16, **2012**
  Prix du meilleur papier

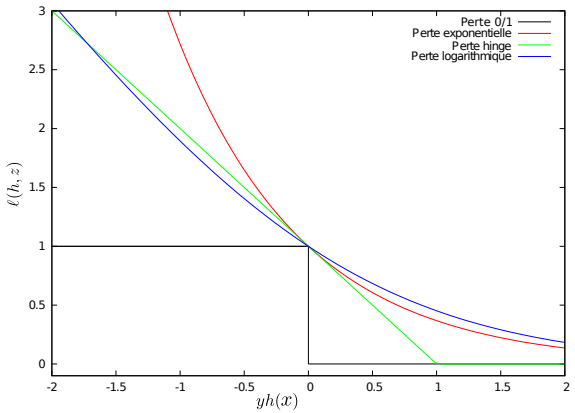
Intuition

L'idée

- Utiliser le boosting pour reprojeter les données dans un nouvel espace
- Rapprocher les deux distributions dans cet espace

Les objectifs

- Être efficace à la fois sur les données sources (**erreur de classification**) et cibles (**maximisation de marge**)
- Réduire la divergence entre les deux distributions

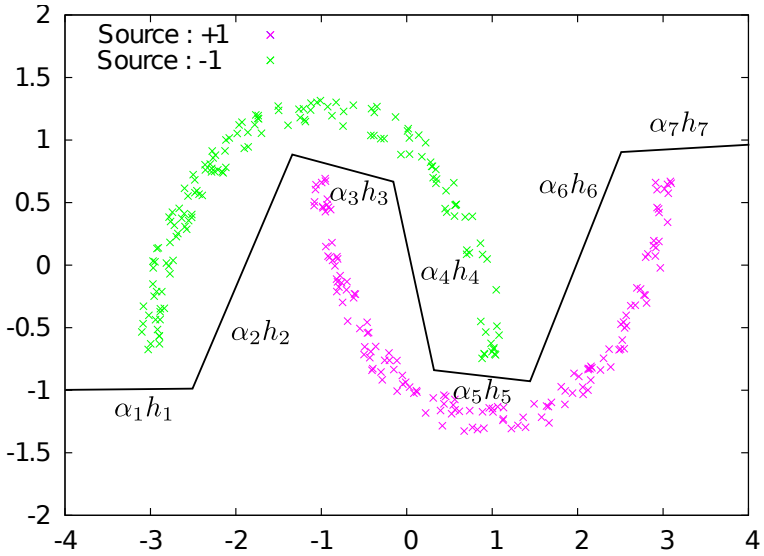


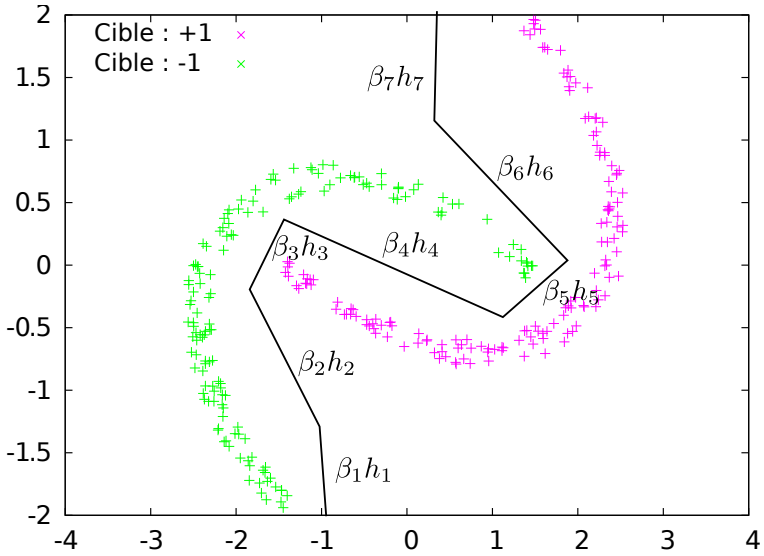
Intuition

AdaBoost [Freund and Schapire, 1997]

- Apprend itérativement des classifieurs binaires faibles h_n (ayant un taux d'erreur légèrement inférieur à l'aléatoire), typiquement des séparateurs linéaires.
- Optimise une perte exponentielle en augmentant les poids des exemples mal classés.
- Construit une combinaison convexe $\sum_{n=1}^{N} \alpha_n h_n$ des classifieurs faibles.

Intuition



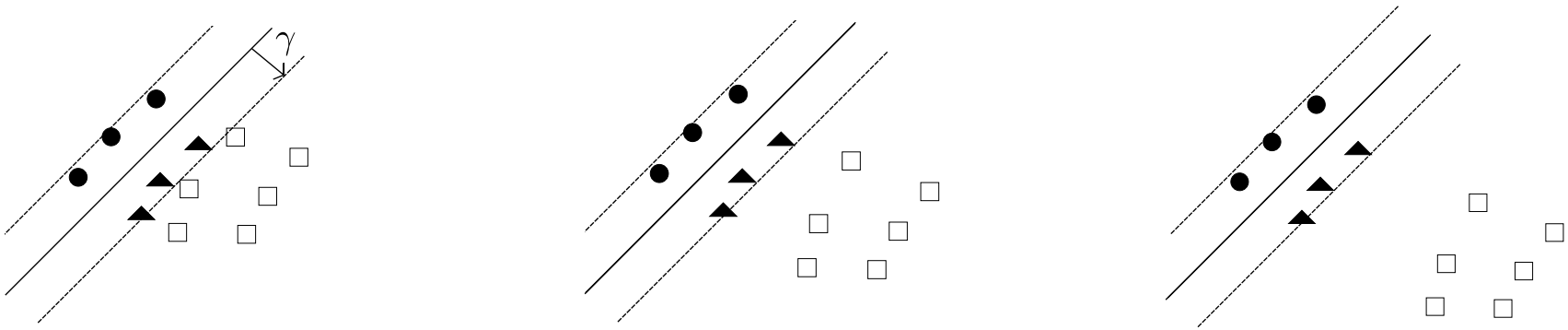


Caractéristiques

- Même espace de projection (sorties des hypothèses).
- Mêmes hypothèses faibles.
- Poids optimisés selon le domaine d'origine (source ou cible).

Intuition

Situation dégénérée



Situation à l'itération 1 | Situation à l'itération i | Situation à l'itération N

Divergence

Pour éviter de telles situations causées par des hypothèses *dégénérées*, nous introduisons une nouvelle notion de divergence $g_n \in [0, 1]$ induite par l'hypothèse h_n entre S et T.

SLDAB (Self-Labeling Domain Adaptation Boosting)

Apprenant faible pour l'AD

Une hypothèse h_n apprise à une itération n est un apprenant faible pour T si :

1. h_n est un apprenant faible pour S (comme dans AdaBoost).
2. $\hat{L}_n = \mathbb{E}_{x \sim D_n^T}[|f_{DA}(h_n(x))| \leq \gamma] < \frac{\gamma}{\gamma + \max(\gamma, \lambda g_n)}$.

- Si $\max(\gamma, \lambda g_n) = \gamma$, $\frac{\gamma}{\gamma + \max(\gamma, \lambda g_n)} = \frac{1}{2}$. La divergence est plutôt petite, la situation ressemble à **l'apprentissage semi-supervisé**.
- Si $\max(\gamma, \lambda g_n) = \lambda g_n$, la contrainte 2 est plus forte pour compenser une divergence plus grande. **L'AD est nécessaire**.

SLDAB

Entrée : un ensemble S d'exemples étiquetés T d'exemples non étiquetés, un nombre d'itérations N, une marge $\gamma \in [0,1]$, un paramètre de compromis $\lambda \in [0,1]$, $l = |S|$, $m = |T|$.

Sortie : deux classifieurs source H_N^S et cible H_N^T.

Initialisation : $\forall (x',y') \in S, D_1^S(x') = \frac{1}{l}$, $\forall x \in T, D_1^T(x) = \frac{1}{m}$.

pour $n = 1$ **à** N **faire**

Apprendre un apprenant faible pour l'AD h_n.

Calculer la valeur de divergence g_n.

$\alpha_n = \frac{1}{2} \ln \frac{1-\hat{\epsilon}_n}{\hat{\epsilon}_n}$ et $\beta_n = \frac{1}{\gamma + \max(\gamma, \lambda g_n)} \ln \frac{\gamma W_n^+}{\max(\gamma, \lambda g_n) W_n^-}$

$\forall (x',y') \in S,\ D_{n+1}^S(x') = D_n^S(x') . \frac{e^{-\alpha_n sgn(h_n(x')).y'}}{Z_n'}$.

$\forall x \in T,\ D_{n+1}^T(x) = D_n^T(x) . \frac{e^{-\beta_n f_{DA}(h_n(x)).y^n}}{Z_n}$,

où $y^n = sgn(f_{DA}(h_n(x)))$ si $|f_{DA}(h_n(x))| > \gamma$,

$y^n = -sgn(f_{DA}(h_n(x)))$ sinon,

et Z_n' et Z_n sont des coefficients de normalisation.

fin

$\forall (x',y') \in S, F_N^S(x') = \sum_{n=1}^N \alpha_n sgn(h_n(x'))$,

$\forall x \in T, F_N^T(x) = \sum_{n=1}^N \beta_n sgn(h_n(x))$.

SLDAB

Entrée : un ensemble S d'exemples étiquetés T d'exemples non étiquetés, un nombre d'itérations N, une marge $\gamma \in [0,1]$, un paramètre de compromis $\lambda \in [0,1]$, $l = |S|$, $m = |T|$.

Sortie : deux classifieurs source H_N^S et cible H_N^T.

Initialisation : $\forall (x',y') \in S, D_1^S(x') = \frac{1}{l}, \forall x \in T, D_1^T(x) = \frac{1}{m}$.

pour $n = 1$ **à** N **faire**

Apprendre un apprenant faible pour l'AD h_n.

Calculer la valeur de divergence g_n.

$\alpha_n = \frac{1}{2} \ln \frac{1-\hat{\epsilon}_n}{\hat{\epsilon}_n}$ et $\beta_n = \frac{1}{\gamma + \max(\gamma, \lambda g_n)} \ln \frac{\gamma W_n^+}{\max(\gamma, \lambda g_n) W_n^-}$

$\forall (x',y') \in S,\ D_{n+1}^S(x') = D_n^S(x') . \frac{e^{-\alpha_n sgn(h_n(x')).y'}}{Z_n'}$.

$\forall x \in T,\ D_{n+1}^T(x) = D_n^T(x) . \frac{e^{-\beta_n f_{DA}(h_n(x)).y^n}}{Z_n}$,

où $y^n = sgn(f_{DA}(h_n(x)))$ si $|f_{DA}(h_n(x))| > \gamma$,

$y^n = -sgn(f_{DA}(h_n(x)))$ sinon,

et Z_n' et Z_n sont des coefficients de normalisation.

fin

$\forall (x',y') \in S, F_N^S(x') = \sum_{n=1}^N \alpha_n sgn(h_n(x'))$,

$\forall x \in T, F_N^T(x) = \sum_{n=1}^N \beta_n sgn(h_n(x))$.

Analyse théorique de SLDAB

Théorème (Borne supérieure de la perte de marge $\hat{L}_{H_N^T}$)

Soit $\hat{L}_{H_N^T}$ la proportion d'exemples de T possédant une marge inférieure à γ par rapport aux divergences successives $g_n (n = 1 \dots N)$ après N itérations :

$$\hat{L}_{H_N^T} = \mathbb{E}_{x \sim T}[\mathbf{y}\mathbf{F_N^T}(x) < 0] \leq \frac{1}{|T|} \sum_{x \sim T} e^{-\mathbf{y}\mathbf{F_N^T}(x)} = \prod_{n=1}^{N} Z_n,$$

où $\mathbf{y} = (y^1, \dots, y^n, \dots, y^N)$ *et*
$\mathbf{F_N^T}(x) = (\beta_1 f_{DA}(h_1(x)), \dots, \beta_n f_{DA}(h_n(x)), \dots, \beta_N f_{DA}(h_N(x)))$.

Analyse théorique de SLDAB

Théorème (Convergence de la perte empirique)

Le borne suivante tient pour la perte empirique $\hat{L}_{H_N^T}$ *du classifieur final* H_N *:*

$$\hat{L}_{H_N^T} \leq \exp \sum_{n=1}^{N} \left(\frac{1}{1+c_n} \ln \tau_n + \ln\left(\frac{c_n+1}{\tau_n+c_n}\right) \right),$$

où $c_n = \frac{\max(\gamma, \lambda g_n)}{\gamma}$ *et* $\tau_n = W_n^+ - 0.5$.
Le terme entre parenthèses étant strictement inférieur à 0, ce théorème fait état du fait que **la perte empirique** $\hat{L}_{H_N^T}$ **décroît exponentiellement vite vers 0 avec** N.

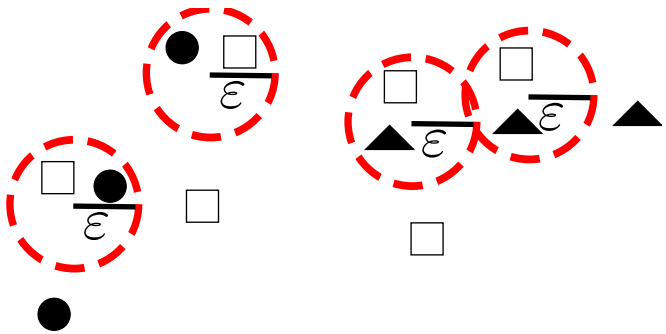
Mesure de divergence

Les mesures de divergence existantes ne conviennent pas à notre cadre. Celui-ci nécessite une divergence qui :

- est induite par **une hypothèse particulière** h_n,
- est en mesure d'évaluer la discordance entre S et T,
- évite les hypothèses dégénérées.

Mesure de divergence

Variation Perturbée (PV) [Harel and Mannor, 2012]



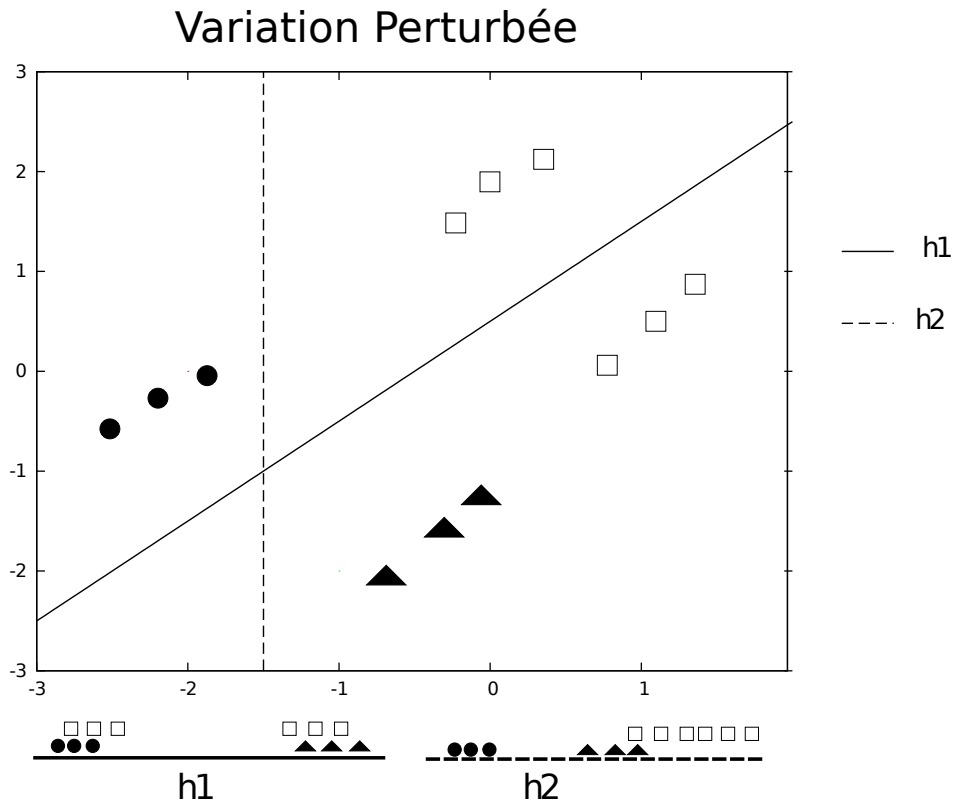
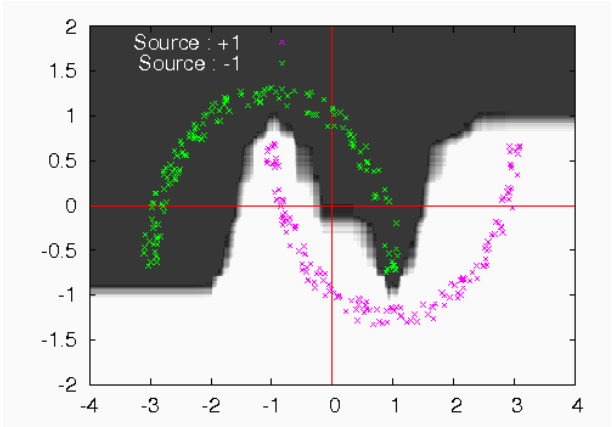
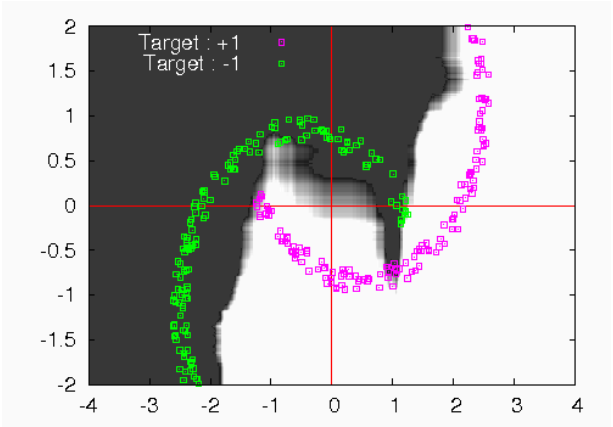


Illustration de SLDAB sur un exemple artificiel

Nous considérons deux lunes **sources** dans un espace à 2 dimensions. Le **domaine cible** est obtenu par une rotation anti-horaire du domaine source en fonction d'angles de rotation de 20 à 90 degrés.



(a) Sur la source : $F_N^S(x)$



(b) Sur la cible : $F_N^T(x)$

$F_N^S(x) = \sum_{n=1}^{N} \alpha_n sign(h_n(x))$ et $F_N^T(x) = \sum_{n=1}^{N} \beta_n sign(h_n(x))$.

Comparaison expérimentale

Comparaison entre SLDAB et deux approches d'AD largement utilisées (DASVM [Bruzzone et Marconcini, 2010] et une technique de repondération SVM-W [Huang et al., 2006]).

Angle	20°	30°	40°	50°	60°	70°	80°	90°	Average
SVM	10.3	24	32.2	40	43.3	55.2	67.7	80.7	44.2 ± 0.9
AdaBoost	20.9	32.1	44.3	53.7	61.2	69.7	77.9	83.4	55.4 ± 0.4
DASVM	**0.0**	21.6	28.4	33.4	38.4	74.7	78.9	81.9	44.6 ± 3.2
SVM-W	6.8	12.9	9.5	26.9	48.2	59.7	66.6	67.8	37.3 ± 5.3
SLDAB	1.2	**3.6**	**7.9**	**10.8**	**17.2**	**39.7**	**47.1**	**45.5**	$\mathbf{21.6 \pm 1.2}$

Base de données LUNES

Algorithme	Taux d'erreur (in%)
SVM	38
AdaBoost	59.4
DASVM	37.5
SVM-W	37.9
SLDAB	**35.8**

Base de données SPAMS

Auto-Étiquetage pour l'AD sur Données Structurées

Publications

- Habrard A., Peyrache J-P., Sebban M.
  Iterative Self-Labeling Domain Adaptation for Linear Structured Image Classification
  IJAIT, Volume N° 22, Issue N° 5, **2013**

- Habrard A., Peyrache J-P., Sebban M.
  DA with Good Edit Similarities : a Sparse Way to deal with Rotation and Scaling Problems in Image Classification
  ICTAI 2011, 181-188, **2011**
  Prix du meilleur papier

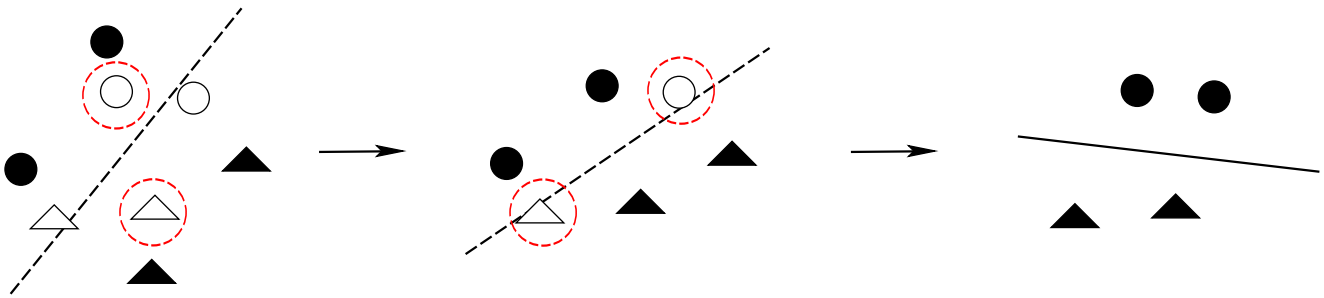
Introduction

Motivations

- Pas de cadre existant pour l'auto-étiquetage
- Approches fonctionnant en pratique mais pas théoriquement fondées
- Peu de travaux concernant les données structurées

Originalité

- Proposer un cadre théorique pour l'auto-étiquetage
- Utiliser la théorie des $(\varepsilon, \gamma, \tau)$-bonnes fonctions de similarité [Balcan et al., 2006]



Modèle théorique

Définition (Exemple semi-étiqueté)

Un point cible semi-étiqueté, inséré dans $S^{(i)}$ à une étape i, à la place d'un point source est un exemple de T (sans remplacement), étiqueté par l'hypothèse $h^{(i-1)}$ apprise depuis $S^{(i-1)}$.

Définition (Apprenant faible pour l'auto-étiquetage)

Un classifieur $h^{(i)}$ appris à une itération i depuis $S^{(i)}$ est un apprenant faible pour l'auto-étiquetage par rapport à $SL^j = \{x_1^T \dots x_{2k}^T\}$ de $2k$ exemples cibles semi-étiquetés, insérés à l'étape j si :

$$\epsilon^{(j)}_{SL^j}(h^{(i)}) = \mathbb{E}_{x_l^T \in SL^j}[h^{(i)}(x_l^T) \neq y_l^T] < \frac{1}{2}.$$

Modèle théorique

Théorème

Soit $h^{(i)}$ *un apprenant faible appris à l'étape* i *depuis* $S^{(i)}$ *et* $\tilde{\epsilon}_S^{(i)}(h^{(i)}) = \frac{1}{2} - \gamma_S^{(i)}$ *son erreur empirique correspondante. Soit* $\epsilon_T^{(i)}(h^{(i)}) = \frac{1}{2} - \gamma_T^{(i)}$ *l'erreur empirique (inconnue) de* $h^{(i)}$ *sur l'ensemble cible* T*.*

$h^{(i)}$ *est un* **apprenant faible pour l'auto-étiquetage** *par rapport à un ensemble* $SL^j = \{x_1^T \dots x_{2k}^T\}$ *de* $2k$ *exemples cibles semi-étiquetés insérés à l'itération* j $(j \leq i)$*, si* $\gamma_T^{(j-1)} > 0$*.*

Modèle théorique

Théorème

Soit $h^{(\frac{N}{2k})}$ l'apprenant faible pour l'auto-étiquetage appris par $\mathcal{A}$ après $\frac{N}{2k}$ itérations, nécessaires pour changer $S^{(0)}$ en un nouvel ensemble d'apprentissage composé uniquement d'exemples cibles. L'algorithme $\mathcal{A}$ effectue une adaptation de domaine efficace avec $h^{(\frac{N}{2k})}$ si

$$\gamma_S^{(i)} \geq \gamma_T^{(i)}, \forall i = 1, \ldots, \frac{N}{2k}, \text{(Apprendre quelque chose)} \quad (1)$$

$$\gamma_S^{max} > \sqrt{\frac{\gamma_T^{(0)}}{2}}, \text{(Être meilleur que sans adaptation)} \quad (2)$$

où $\gamma_S^{max} = max(\gamma_S^{(0)}, \ldots, \gamma_S^{(n)})$.

Modèle théorique

Conditions nécessaires :

- $h^{(i)}$ doit fonctionner correctement sur T $(\gamma_T^{(i)} > 0)$.
- $h^{(i)}$ doit fonctionner correctement sur S $(\gamma_S^{(i)} > \gamma_T^{(i)})$.
- $\mathcal{A}$ doit obtenir une meilleure performance qu'un processus non adaptatif.

Conclusion

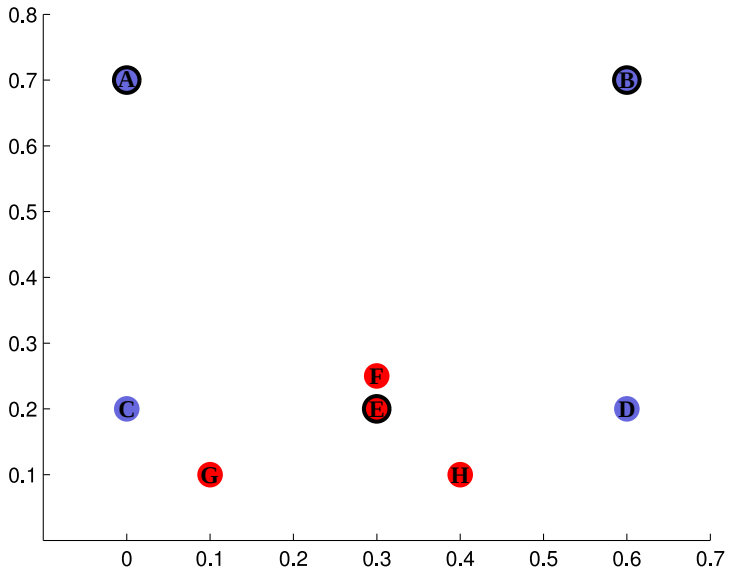
Il faut choisir des exemples qui :

- aident à apprendre quelque chose de nouveau sur le domaine cible.
- ont de grandes chances d'être correctement semi-étiquetés.

DASVM

- Suit une stratégie similaire
- Limitations dues à la théorie des SVMs

Les $(\varepsilon, \gamma, \tau)$-bonnes fonctions de similarité

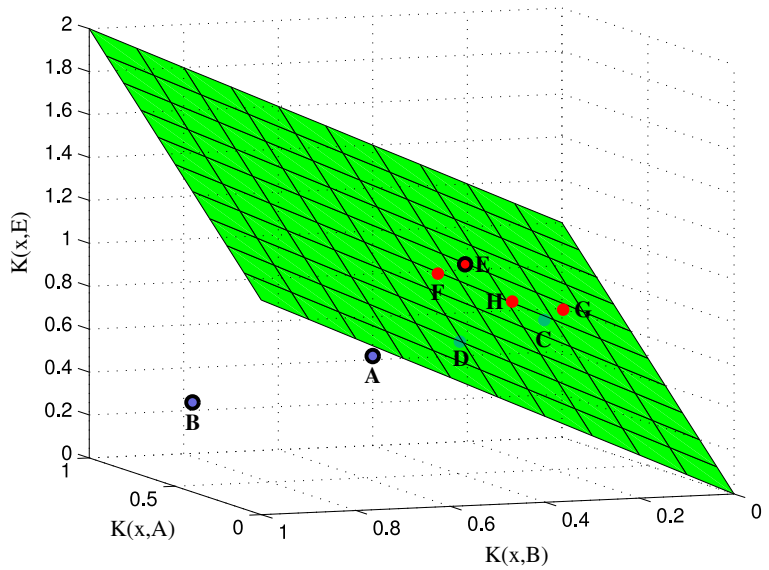


Définition

Une fonction de similarité K est $(\varepsilon,\gamma,\tau)$-bonne si :

1. *Une proportion 1-ε des exemples (x, ℓ) satisfait :* $E_{(x',\ell')} P[\ell\ell' K(x, x')|R(x')] \geq \gamma$,
2. $Pr_{x'}[R(x')] \geq \tau$.

Les $(\varepsilon, \gamma, \tau)$-bonnes fonctions de similarité



Problème d'optimisation linéaire à résoudre :

$$\min_{\alpha} \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha^j \ell_i K(x_i, x'_j) \right]_+ + \lambda ||\alpha||_1,$$

où $[1 - z]_+ = \max(0, 1 - z)$ est la perte hinge.

Les $(\varepsilon, \gamma, \tau)$-bonnes fonctions de similarité

Intérêt

- Modèles parcimonieux
- Pas de contraintes de symétrie ou de SDP (contrairement aux SVMs)
- Application directe aux données structurées

Définition (Distance d'édition)

La distance d'édition $e_d(x, x')$ entre deux chaînes de caractères x et x' est le nombre minimal d'opérations (insertion, substitution ou suppression d'un symbole) à effectuer pour transformer x en x'.

On utilise $K(x, x') = -e_d(x, x')$.

GESIDA

Itérations

- $S^{(0)} = S$ et $T^{(0)} = T$.
- Classifieur $h^{(i)}$ appris depuis $S^{(i)}$ par résolution de :

$$\min_{\alpha} \sum_{i=1}^{|S^{(i)}|} \left[1 - \sum_{j=1}^{|S^{(i)}|} \alpha^j y_i K(x_i, x'_j) \right]_+ + \lambda ||\alpha||_1,$$

- $2k$ exemples semi-étiquetés (k de chaque classe) insérés dans $S^{(i+1)}$

Sélection des exemples semi-étiquetés

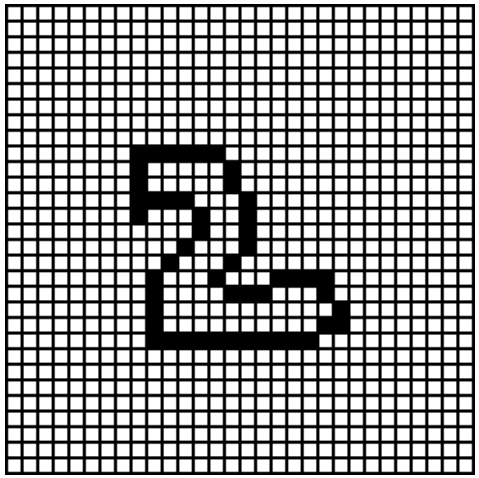
Deux phases :

- de $i = 0$ à $i = I$: sélection des exemples avec la **plus grande marge**
- pour $i > I$: sélection des exemples avec **une marge de plus en plus inférieure à** γ.

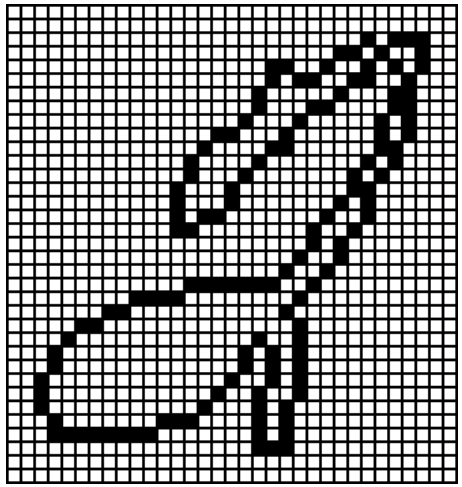
Résultats expérimentaux

Chiffres manuscrits

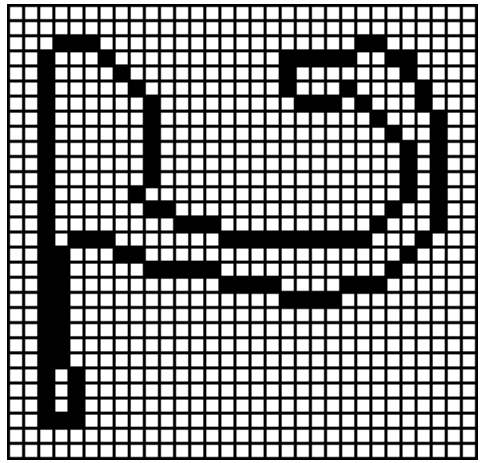
Problèmes de mise à l'échelle et de rotation



Source



Cible (mise à l'échelle)

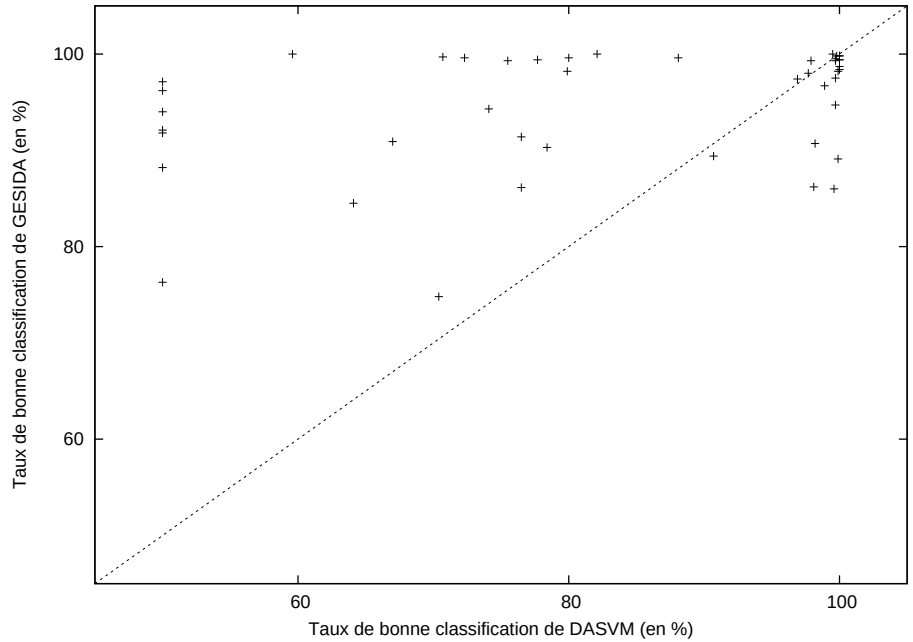


Cible (rotation)

Résultats expérimentaux

Problèmes de changement d'échelle

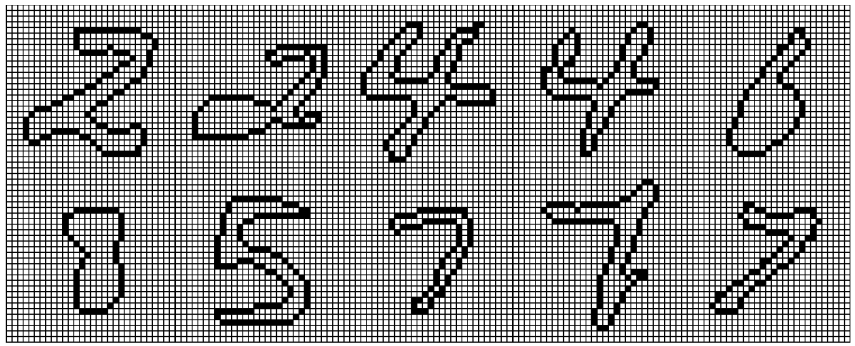
% de bonne classification de DASVM	83.3 ± 3.2
Nombre final de vecteurs de support	120 ± 7.8
% de bonne classification de GESIDA	$\mathbf{94.7 \pm 2.1}$
Nombre final de points raisonnables	$\mathbf{11 \pm 2.4}$
% de bonne classification par sélection aléatoire pour l'AD	52.14 ± 0.6
% de bonne classification sans adaptation	50.21 ± 1.7



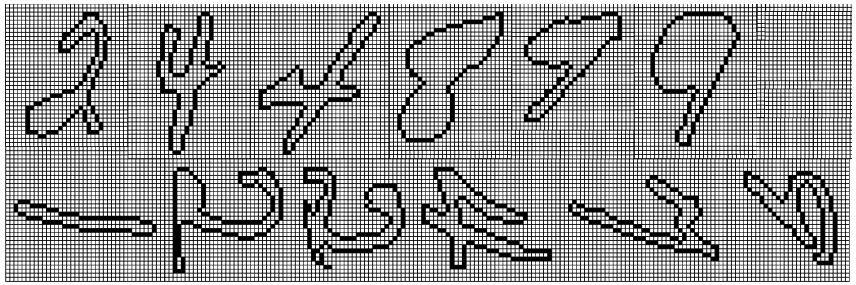
Résultats expérimentaux

Étude des points raisonnables

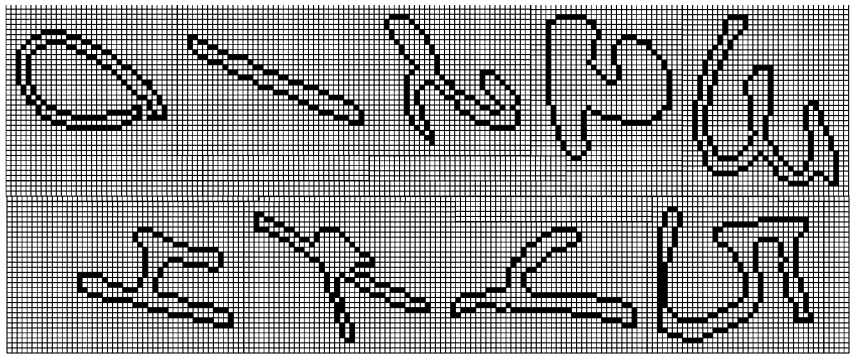
Au début du processus



Au milieu du processus



À la fin du processus



Conclusion et Perspectives Générales

Conclusion

SLDAB

- Approche basée sur le boosting
- Mêmes apprenants faibles sur S et T, tenant compte d'une mesure de divergence
- Convergence de la perte empirique
- Efficace dans les cadres d'AD et d'apprentissage semi-supervisé

GESIDA

- Approche d'auto-étiquetage traitant directement les données structurées
- Utilisation du cadre des $(\varepsilon, \gamma, \tau)$-bonnes fonctions de similarité
- Étude théorique des conditions nécessaires à la réussite d'un algorithme d'auto-étiquetage
- Résultats expérimentaux probants

Perspectives

...des contributions

- Étendre l'approche aux données structurées (SLDAB)
- Dériver des garanties en généralisation (SLDAB)
- Étendre l'analyse théorique à notre stratégie de sélection des exemples (GESIDA)

Plus généralement...

- Définir une notion générale d'apprenant faible
- Étudier la problématique de la représentation des données
- Évaluer λ pour éviter un processus de transfert négatif
- Étendre à d'autres domaines : régression, *Life-long Learning*, AD supervisée...

Merci pour votre attention

SLDAB - Apprenant faible pour l'AD

k stumps aléatoires

- $\frac{k}{2}$ remplissant la condition sur S
- $\frac{k}{2}$ remplissant la condition sur T

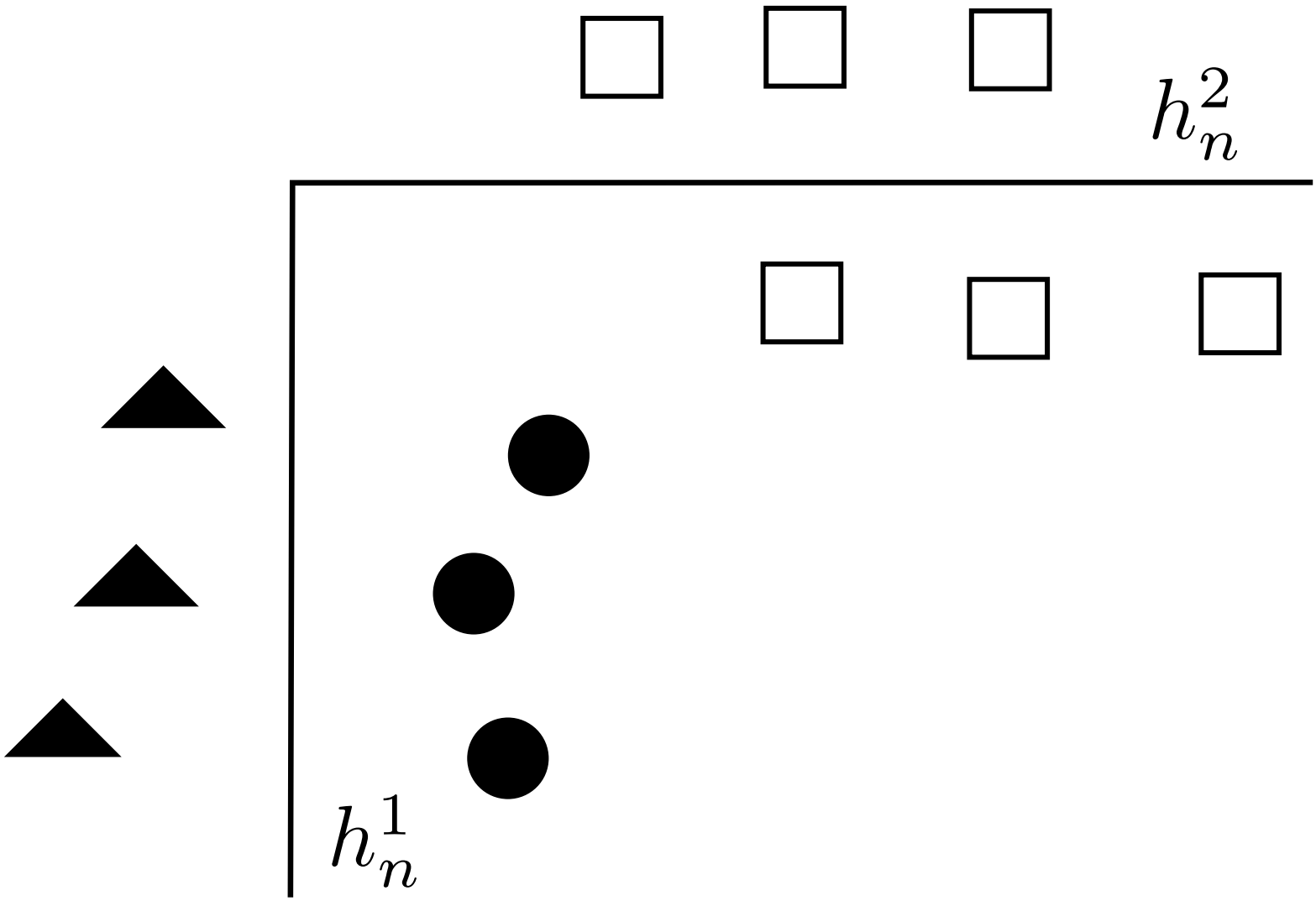
Nous suggérons ensuite de résoudre le problème d'optimisation suivant, afin de trouver la meilleure combinaison convexe $h_n = \sum_k \kappa_k h_n^k$ (avec $\sum_k \kappa_k = 1$) :

Problème d'optimisation

$$\underset{\kappa}{\text{argmin}} \sum_{(x',y')\in S} D_n^S(x') \left[-y' \sum_k \kappa_k sgn(h_n^k(x')) \right]_+ + \sum_{x\in T} D_n^T(x) \left[1 - \left(\sum_k \kappa_k marg(f_{DA}(h_n^k(x))) \right) \right]_+$$

où $[1-x]_+ = max(0, 1-x)$ est la perte hinge, et $marg(f_{DA}(h_n^k(x)))$ renvoie -1 si $f_{DA}(h_n^k(x))$ est plus petit que γ et $+1$ sinon.

SLDAB - Apprenant faible pour l'AD



SLDAB - Généralisation

Théorème (Borne sur l'erreur en généralisation d'AdaBoost)

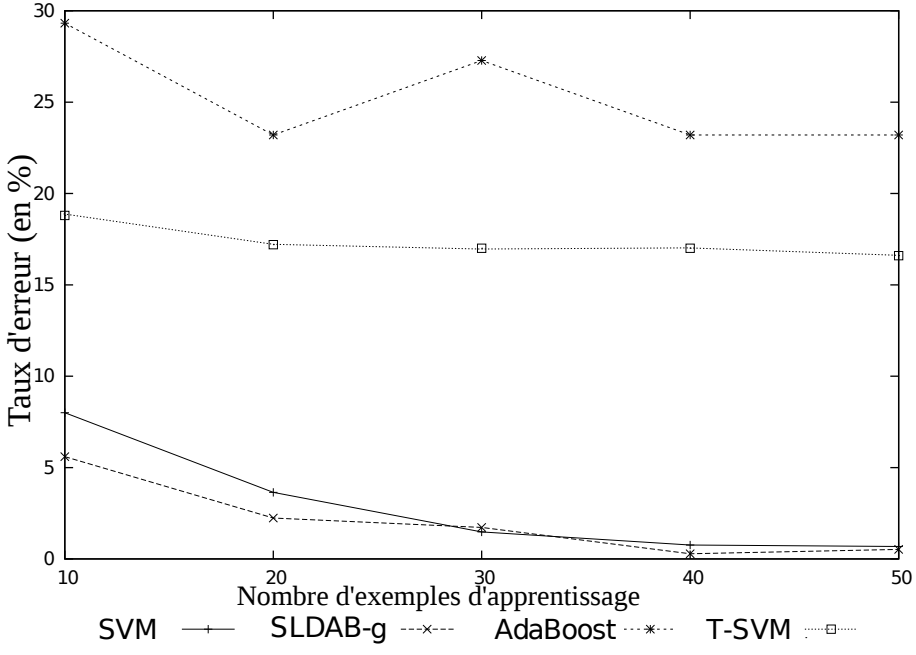
Soit $\mathcal{H}$ une classe de classifieurs de VC-dimension d. $\forall \delta > 0$ et $\gamma > 0$, avec une probabilité $1 - \delta$, n'importe quel ensemble de N classifieurs construit depuis un échantillon d'apprentissage S de taille $|S|$ issu d'une distribution $\mathcal{D}_S$ satisfait l'inégalité suivante sur l'erreur en généralisation $\epsilon_{\mathcal{D}_S}(H_S^N)$:

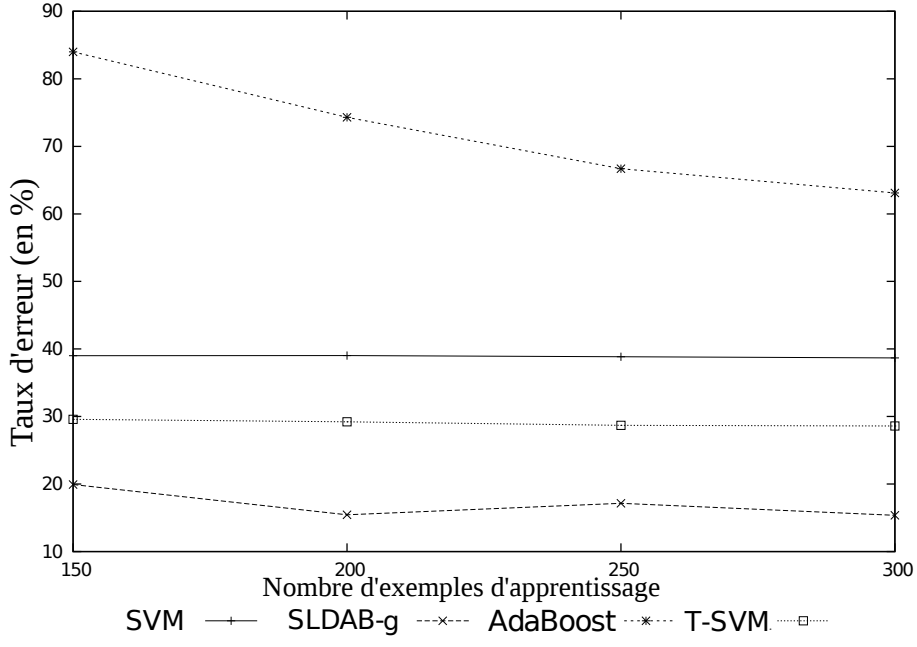
$$\epsilon_{\mathcal{D}_S}(H_S^N) \leq \widehat{Pr}_{\mathbf{x}\sim S}[marge(\mathbf{x}) \leq \gamma] + \mathcal{O}\left(\sqrt{\frac{d}{|S|}\frac{log^2(|S|/d)}{\gamma^2} + log(1/\delta)}\right).$$

Borne sur l'erreur en généralisation de SLDAB

$$\epsilon_{\mathcal{D}_T}(H_T^N) \quad \leq \quad \hat{L}_{H_S^N} + div(S,T) + \lambda^\star + \mathcal{O}\left(\sqrt{\frac{d}{|T|}\frac{log^2(|T|/d)}{\gamma^2} + log(1/\delta)}\right)$$

SLDAB - Cadre semi-supervisé





GESIDA - Algorithme

Entrée :

- un ensemble S de N exemples sources étiquetés,
- un ensemble T de $M > N$ exemples cibles non étiquetés,
- deux paramètres k et I.

Sortie : un classifieur h.

Initialisation : $S^{(0)} = S$; $T^{(0)} = T$ **pour** $i = 1$ **à** $\frac{N}{2k}$ **faire**

Apprendre $h^{(i)}$ sur $S^{(i)}$

Calculer la marge des exemples de $S^{(i)}$ et $T^{(i)}$

Mettre à jour $S^{(i)}$ et $T^{(i)}$ afin de traiter des *outliers* semi-étiquetés dans $S^{(i)}$

Construire $SL^{(i)}$ et $Sup^{(i)}$ en fonction des paramètres k et I

$S^{(i+1)} \leftarrow SL^{(i)} \cup (S^{(i)} \setminus Sup^{(i)})$

$T^{(i+1)} \leftarrow T^{(i)} \setminus SL^{(i)}$

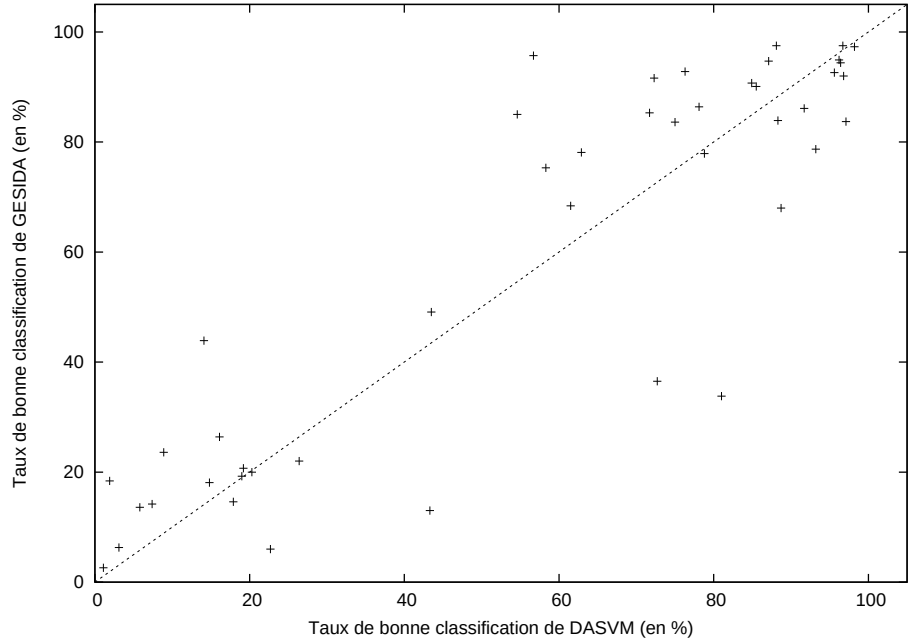
fin

Renvoyer le classifieur final appris sur $S^{(\frac{N}{2k})}$

GESIDA - Résultats expérimentaux

Problèmes de rotation

% de bonne classification de DASVM Nombre final de vecteurs de support	57.1 ± 11.7 113 ± 9.3
% de bonne classification de GESIDA Nombre final de points raisonnables	$\mathbf{59.2 \pm 8.1}$ $\mathbf{17 \pm 2.7}$
% de bonne classification par sélection aléatoire pour l'AD	56.63 ± 7.8
% de bonne classification sans adaptation	55.48 ± 7.7



GESIDA - Résultats expérimentaux

Étude sur la sélection aléatoire

Itération	P_1			P_2		
	$\gamma_S^{(i)}$	$\gamma_T^{(i-1)}$	1 - $\epsilon_T^{(i)}$	$\gamma_S^{(i)}$	$\gamma_T^{(i-1)}$	1 - $\epsilon_T^{(i)}$
1	0.5	0	0.585	0.50	-0.1	0.32
2	0.475	0.085	0.75	0.50	-0.18	0.285
3	0.48	0.25	0.73	0.50	-0.215	0.285
4	0.49	0.23	0.795	0.50	-0.215	0.24
5	0.49	0.295	0.875	0.50	-0.26	0.18
6	0.49	0.375	0.94	0.50	-0.32	0.205
7	0.49	0.44	0.94	0.50	-0.295	0.19
8	0.49	0.44	0.94	0.50	-0.31	0.12
9	0.49	0.44	0.94	0.50	-0.38	0.145
10	0.495	0.44	0.985	0.50	-0.355	0.115
11	0.5	0.485	0.99	0.495	-0.385	0.115