Apprentissage de vote de majorité pour la classification supervisée et l'adaptation de domaine

approches PAC-Bayésiennes et combinaison de similarités

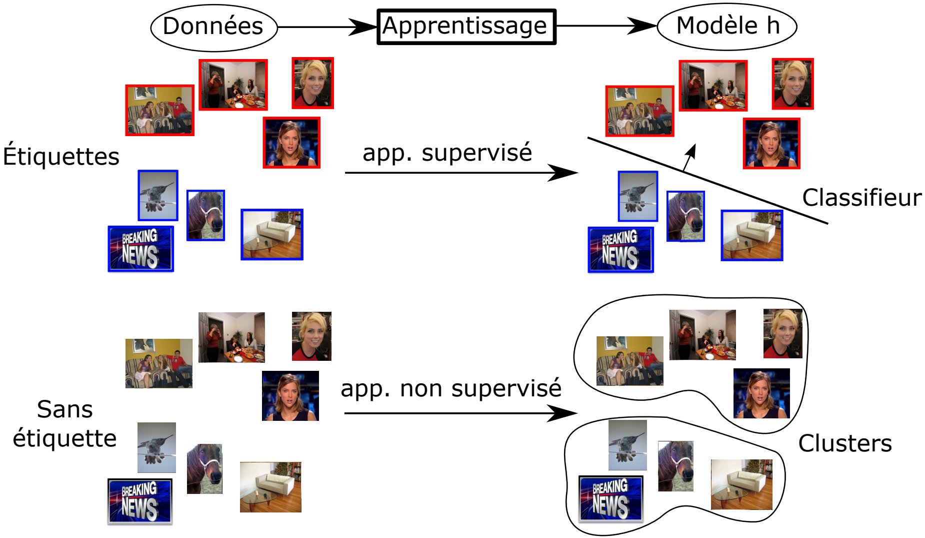
Emilie Morvant

Laboratoire d'Informatique Fondamentale, QARMA Group, Aix*Marseille Université, France

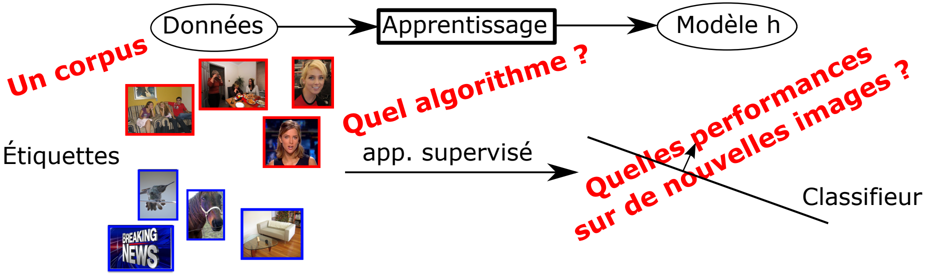
Rapporteurs : Michèle Sebag, Mario Marchand
Examinateurs : Antoine Cornuéjols, Rémi Gilleron, Liva Ralaivola
Directeur : Amaury Habrard
Co-directeur : Stéphane Ayache

Soutenance de thèse
18 septembre 2013

Tâche : Y-a-t-il une `Personne` dans l'image/la vidéo ?

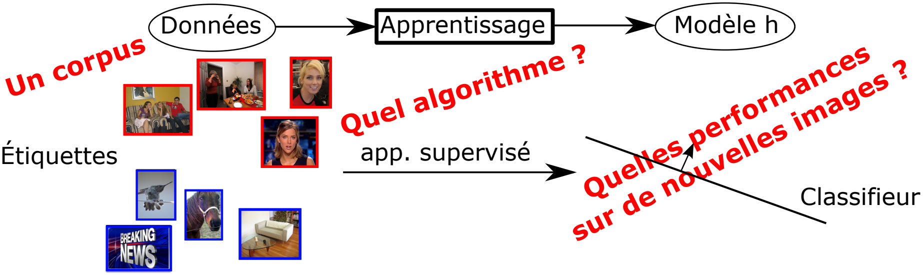


Tâche : Y-a-t-il une `Personne` dans l'image/la vidéo ?



Comment apprendre *h* pour qu'il se trompe le moins possible sur de nouvelles images ?

Tâche : Y-a-t-il une `Personne` dans l'image/la vidéo ?



Comment apprendre h pour qu'il se trompe le moins possible sur de nouvelles images ?

Solution : minimiser l'erreur empirique mesurée sur les données d'apprentissage

$\Rightarrow$ Requiert des garanties $\Rightarrow$ Borne en généralisation

erreur réelle $\leq$ erreur empirique $+$ f (complexité, nb de données)

Problématiques

1. Comment tirer bénéfice de différentes descriptions ? *ex : son, texte, image, couleur, . . .*

2. Plusieurs objets différents à reconnaître. *ex : personne, canapé, cheval, . . .*

3. Les nouvelles images proviennent d'un corpus différent. *ex : photos* $\rightarrow$ *vidéos*

Problématiques

1. Comment tirer bénéfice de différentes descriptions ? *ex : son, texte, image, couleur, . . .*

$\Rightarrow$ **multimodalité, multivue, combinaisons/fusion de modèles** $\Rightarrow$ **vote de majorité**

2. Plusieurs objets différents à reconnaître. *ex : personne, canapé, cheval, . . .*

3. Les nouvelles images proviennent d'un corpus différent. *ex : photos $\rightarrow$ vidéos*

Problématiques

1. Comment tirer bénéfice de différentes descriptions ? *ex : son, texte, image, couleur, . . .*

$\Rightarrow$ **multimodalité, multivue, combinaisons/fusion de modèles** $\Rightarrow$ **vote de majorité**

Qu'est ce qu'un vote de majorité ?

$\mathcal{H}$: un ensemble de modèles/votants retournant -1 ou $+1$

Objectif : Construire un vote de majorité pondéré sur $\mathcal{H}$: $\quad \operatorname{sign}\left[\sum_{h\in\mathcal{H}} \rho(h)h(\mathbf{x})\right]$

Question : Comment apprendre les poids $\rho(h)$?

2. Plusieurs objets différents à reconnaître. *ex : personne, canapé, cheval, . . .*

3. Les nouvelles images proviennent d'un corpus différent. *ex : photos $\rightarrow$ vidéos*

Problématiques

1. Comment tirer bénéfice de différentes descriptions ? *ex : son, texte, image, couleur, . . .*

$\Rightarrow$ **multimodalité, multivue, combinaisons/fusion de modèles** $\Rightarrow$ **vote de majorité**

Qu'est ce qu'un vote de majorité ?

$\mathcal{H}$: un ensemble de modèles/votants retournant -1 ou $+1$

Objectif : Construire un vote de majorité pondéré sur $\mathcal{H}$: $\operatorname{sign}\left[\sum_{h\in\mathcal{H}} \rho(h)h(\mathbf{x})\right]$

Question : Comment apprendre les poids $\rho(h)$?

2. Plusieurs objets différents à reconnaître. *ex : personne, canapé, cheval, . . .*

$\Rightarrow$ **classification multiclasse ou multilabel**

3. Les nouvelles images proviennent d'un corpus différent. *ex : photos $\rightarrow$ vidéos*

Problématiques

1. Comment tirer bénéfice de différentes descriptions ? *ex : son, texte, image, couleur, ...*

$\Rightarrow$ **multimodalité, multivue, combinaisons/fusion de modèles** $\Rightarrow$ **vote de majorité**

Qu'est ce qu'un vote de majorité ?

$\mathcal{H}$: un ensemble de modèles/votants retournant -1 ou $+1$

Objectif : Construire un vote de majorité pondéré sur $\mathcal{H}$: $\quad \operatorname{sign}\left[\sum_{h\in\mathcal{H}} \rho(h)h(\mathbf{x})\right]$

Question : Comment apprendre les poids $\rho(h)$?

2. Plusieurs objets différents à reconnaître. *ex : personne, canapé, cheval, ...*

$\Rightarrow$ **classification multiclasse ou multilabel**

3. Les nouvelles images proviennent d'un corpus différent. *ex : photos $\rightarrow$ vidéos*

$\Rightarrow$ **on doit adapter le modèle** $\Rightarrow$ **adaptation de domaine**

Contributions de la thèse

Classification supervisée et théorie PAC-Bayésienne

- Vote de majorité contraint et classification binaire
  - Application à des classifieurs de type *k* plus proches voisins (CAp'13)
  - Spécialisation à la fusion de classifieurs en multimédia
- Classification multiclasse
  - Borne PAC-Bayésienne sur la confusion du classifieur de Gibbs (ICML'12, CAp'12)
  - Borne sur le risque du vote de majorité pondéré

Adaptation de domaine

- Adaptation de domaine par pondération de fonctions de similarité (ϵ,γ,τ)-bonnes (ICDM'11, CAp'11, KAIS'12)
- Analyse PAC-Bayésienne de l'adaptation de domaine (ICML'13, CAp'13)

Contributions de la thèse

Classification supervisée et théorie PAC-Bayésienne

- Vote de majorité contraint et classification binaire
  - Application à des classifieurs de type k plus proches voisins (CAp'13)
  - Spécialisation à la fusion de classifieurs en multimédia
- Classification multiclasse
  - Borne PAC-Bayésienne sur la confusion du classifieur de Gibbs (ICML'12, CAp'12)
  - Borne sur le risque du vote de majorité pondéré

Adaptation de domaine

- Adaptation de domaine par pondération de fonctions de similarité (ϵ,γ,τ)-bonnes (ICDM'11, CAp'11, KAIS'12)
- Analyse PAC-Bayésienne de l'adaptation de domaine (ICML'13, CAp'13)

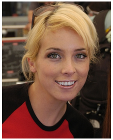
Sommaire

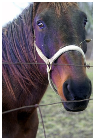
Sommaire

Quand a-t-on besoin d'adaptation de domaine (DA) ?

Lorsque la distribution **d'apprentissage** **diffère** de la distribution **de test**

Exemple





`Personne` `pas de Personne`

Corpus de **Photos** **étiquetées**

Domaine **source**

$\xrightarrow{?}$

Y-a-t'il une `Personne` ?

Corpus de **Videos** **non étiquetées**

Domaine **cible**

$\Rightarrow$ Comment apprendre, à partir d'une distribution,
un classifieur performant sur une distribution différente ?

$X \in \mathbb{R}^d$ espace d'entrée

$Y = \{-1, +1\}$ espace de sortie

$\mathcal{H}$ ensemble de classifieurs

P_S domaine **source**
P_T domaine **cible**
} distributions sur $X \times Y$

D_S, D_T distributions marginales sur X

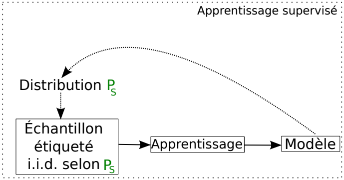
$X \in \mathbb{R}^d$ espace d'entrée

$Y = \{-1, +1\}$ espace de sortie

$\mathcal{H}$ ensemble de classifieurs

$\left.\begin{array}{r} P_S \text{ domaine } \textbf{source} \\ P_T \text{ domaine } \textbf{cible} \end{array}\right\}$ distributions sur $X \times Y$

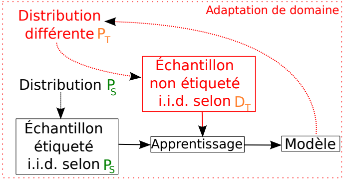
D_S, D_T distributions marginales sur X



Classification supervisée

Objectif : Trouver $h \in \mathcal{H}$ minimisant **l'erreur source** : $R_{P_S}(h) = \mathop{\mathbf{E}}_{(\mathbf{x}^s, y^s) \sim P_S} \mathbf{I}\left[h(\mathbf{x}^s) \neq y^s\right]$

$X \in \mathbb{R}^d$	espace d'entrée
$Y = \{-1, +1\}$	espace de sortie
$\mathcal{H}$	ensemble de classifieurs
P_S domaine **source** P_T domaine **cible**	distributions sur $X \times Y$
D_S, D_T	distributions marginales sur X



Classification supervisée

Objectif : Trouver $h \in \mathcal{H}$ minimisant **l'erreur source** : $R_{P_S}(h) = \underset{(\mathbf{x}^s, y^s) \sim P_S}{\mathbf{E}} \mathbf{I}\left[h(\mathbf{x}^s) \neq y^s\right]$

Adaptation de domaine

Contexte

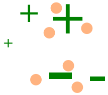
$S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m_s}$ Échantillon **source** tiré *i.i.d.* selon P_S

$T = \{\mathbf{x}_i^t\}_{i=1}^{m_t}$ Échantillon **cible** tiré *i.i.d.* selon D_T

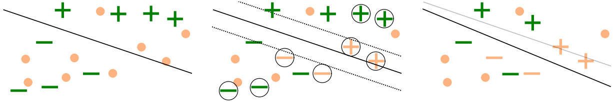
Objectif : Trouver $h \in \mathcal{H}$ minimisant **l'erreur cible** : $R_{P_T}(h) = \underset{(\mathbf{x}^t, y^t) \sim P_T}{\mathbf{E}} \mathbf{I}\left[h(\mathbf{x}^t) \neq y^t\right]$

Trois grands types d'approches

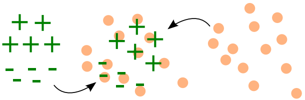
Repondérations des données
[Huang et al., 2007, Jiang and Zhai, 2007, Mansour et al., 2009b]



Auto-étiquetage
[Bruzzone and Marconcini, 2010]

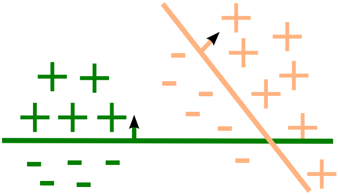


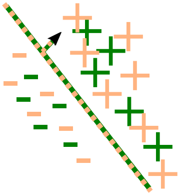
Recherche d'un espace de représentation commun
[Blitzer et al., 2006, Blitzer et al., 2011, Chen et al., 2011, Daumé III, 2007, Daumé III et al., 2010]



Problématique principale

Si h est appris depuis le domaine **source**,
quelle sera sa performance sur le domaine **cible** ?





$\Longrightarrow$ Si les domaines sont "**proches**"
alors un classifieur d'erreur **source** faible peut être un classifieur d'erreur **cible** faible

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \ R_{P_S}(h) \ + \ \tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) \ + \ \nu$$

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \frac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu$$

$R_{P_S}(h)$: erreur classique sur le domaine source

Minimisable via une méthode de classification supervisée sans adaptation

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu}_{\text{divergences}}$$

$\frac{1}{2} d_{\mathcal{H}}(D_S, D_T)$: la $\mathcal{H}$-divergence entre D_S et D_T

$$\begin{aligned}
\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) &= \sup_{(h,h')\in\mathcal{H}^2} \Big| R_{D_T}(h, h') - R_{D_S}(h, h') \Big| \\
&= \sup_{(h,h')\in\mathcal{H}^2} \Big| \mathop{\mathbf{E}}_{\mathbf{x}^t \sim D_T} \mathbf{I}\big[h(\mathbf{x}^t) \neq h'(\mathbf{x}^t)\big] - \mathop{\mathbf{E}}_{\mathbf{x}^s \sim D_S} \mathbf{I}\big[h(\mathbf{x}^s) \neq h'(\mathbf{x}^s)\big] \Big|
\end{aligned}$$

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu}_{\text{divergences}}$$

ν : divergence entre les étiquetages

$\nu = \inf_{h' \in \mathcal{H}} \left(R_{P_S}(h') + R_{P_T}(h')\right)$,
erreur jointe optimale [Ben-David et al., 2010]

ou $\nu = R_{P_T}(h^*_T) + R_{P_T}(h^*_T, h^*_S)$,
$h^*_{\mathcal{X}}$ est la meilleure hypothèse sur le domaine $\mathcal{X}$ [Mansour et al., 2009a]

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T)}_{\text{divergence}} + \cancel{\nu}$$

Idée

Minimiser cette borne pour construire un nouvel espace de projection

- rapprochant les deux distributions marginales
- tout en gardant de bonnes performances sur le domaine source
- en supposant que les deux domaines sont liés

Théorème classique [Ben-David et al., 2010, Mansour et al., 2009a]

Soit $\mathcal{H}$ un espace d'hypothèses. Si D_S et D_T sont deux distributions sur X, alors :

$$\forall h \in \mathcal{H}, \quad \overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T)}_{\text{divergence}} + \cancel{\nu}$$

Première contribution

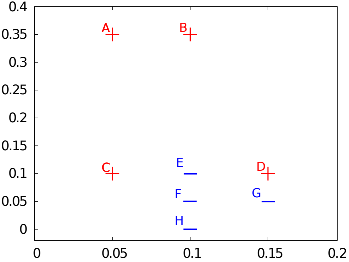
Travailler sur l'espace de projection explicite défini à partir de fonctions de similarités (ϵ,γ,τ)-bonnes

Sommaire

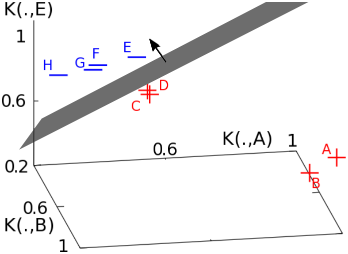
AD par pondération de fonctions de similarité (ϵ,γ,τ)-bonnes

Apprentissage supervisé avec des fonctions de similarité (ϵ,γ,τ)-bonnes [Balcan et al., 2008] (1/2)

$$
\begin{array}{l} K : X \times X \to [-1,+1] \\ R = \{\mathbf{x}'_j\}_{j=1}^{r} \end{array} \implies \phi^R : \left\{ \begin{array}{ccc} X & \to & \mathbb{R}^r \\ \mathbf{x} & \mapsto & \langle K(\mathbf{x},\mathbf{x}'_1), \dots, K(\mathbf{x},\mathbf{x}'_r) \rangle \end{array} \right.
$$



$$\xrightarrow{\phi^R}$$



un classifieur linéaire dans le ϕ^R-espace $\Leftrightarrow$ un vote de majorité pondéré sur $\left\{K(\cdot,\mathbf{x}'_j)\right\}_{\mathbf{x}'_j \in R}$

$$h(\mathbf{x}) = \operatorname{sign}\left[\sum_{j=1}^{r} \alpha_j K(\mathbf{x}, \mathbf{x}'_j)\right]$$

Définition [Balcan et al., 2008]

$K : X \times X \to [-1,+1]$ est dite (ϵ,γ,τ)-**bonne** sur un domaine P si

(i) une proportion de $1-\epsilon$ des exemples $(\mathbf{x}, y)$ satisfont : $\mathop{\mathbf{E}}_{(\mathbf{x}',y')\sim P} \left[yy'K(\mathbf{x},\mathbf{x}') | \mathbf{x}' \in R \right] \ \geq \ \gamma$

(ii) $\mathop{\mathbf{Pr}}_{\mathbf{x}'} [\mathbf{x}' \in R] \geq \tau$

Intuitivement

Pour $(\mathbf{x}_1, y_1) \sim P$, on veut qu'**en moyenne** pour $(\mathbf{x}'_2, y'_2) \in R$

si $y_1 = y'_2$	si $y_1 \neq y'_2$
$\mathbf{x}_1$ est similaire à $\mathbf{x}'_2$	$\mathbf{x}_1$ est similaire à $\mathbf{x}'_2$
$K(\mathbf{x}_1, \mathbf{x}'_2) \geq \gamma$	$K(\mathbf{x}_1, \mathbf{x}'_2) \leq -\gamma$

Remarque : Moins de contraintes qu'une fonction noyau

K ni symétrique, ni SDP $\Longrightarrow$ Facilite la modification l'espace pour adapter

AD par pondération de fonctions de similarité (ϵ,γ,τ)-bonnes

Minimiser l'erreur sur le domaine source

Rappel

Minimiser $R_{P_T}(h)$ à l'aide de la borne

$$\overbrace{R_{P_T}(h)}^{\text{erreur cible}} \quad\leq\quad \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T)}_{\text{divergence}} + \cancel{\nu}$$

Rappel

Minimiser $R_{P_T}(h)$ à l'aide de la borne

$$\overbrace{R_{P_T}(h)}^{\text{erreur cible}} \leq \underbrace{R_{P_S}(h)}_{\text{erreur source}} + \underbrace{\tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T)}_{\text{divergence}} + \cancel{\nu}$$

On minimise $R_{P_S}(h)$ via le problème d'optimisation pour les SF avec $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m_s}$ et $R = \{\mathbf{x}_j'\}_{j=1}^{r}$

$$\begin{cases} \displaystyle\min_{\boldsymbol{\alpha}} \frac{1}{m_s} \sum_{i=1}^{m_s} L\big(h, (\mathbf{x}_i^s, y_i^s)\big) + \lambda \|\boldsymbol{\alpha}\|_1, \\ \quad \text{avec } L\big(h, (\mathbf{x}_i^s, y_i^s)\big) = \left[1 - y_i^s \displaystyle\sum_{j=1}^{r} \alpha_j K(\mathbf{x}_i^s, \mathbf{x}_j')\right]_+ \end{cases}$$

$\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^{r} |\alpha_j|$

Minimiser $d_{\mathcal{H}}(D_S, D_T) \Longleftrightarrow$ Rapprocher les marginales

Problème : minimiser simultanément la divergence et l'erreur source est difficile. . .

Minimiser $d_{\mathcal{H}}(D_S, D_T) \iff$ Rapprocher les marginales

Problème : minimiser simultanément la divergence et l'erreur source est difficile...

Notre solution

$\mathcal{C}_{ST}$ un ensemble de couples $(\mathbf{x}^s, \mathbf{x}^t) \in S \times T$

Construction d'un nouvel espace de projection ϕ^R_{new}

t.q. la différence entre les pertes de $\mathbf{x}^s$ et $\mathbf{x}^t$ est faible

$$\left| \left[1 - y\sum_{j=1}^{r} \alpha_j K(\mathbf{x}^s, \mathbf{x}'_j)\right]_+ - \left[1 - y\sum_{j=1}^{r} \alpha_j K(\mathbf{x}^t, \mathbf{x}'_j)\right]_+ \right| \leq \underbrace{\left\| \left({}^t\phi^R(\mathbf{x}^s) - {}^t\phi^R(\mathbf{x}^t)\right)\mathrm{diag}(\boldsymbol{\alpha}) \right\|_1}_{\left\| {}^t\phi^R_{new}(\mathbf{x}^s) - {}^t\phi^R_{new}(\mathbf{x}^t) \right\|_1}$$

Minimiser $d_{\mathcal{H}}(D_S, D_T) \iff$ Rapprocher les marginales

Problème : minimiser simultanément la divergence et l'erreur source est difficile...

Notre solution

$\mathcal{C}_{ST}$ un ensemble de couples $(\mathbf{x}^s, \mathbf{x}^t) \in S \times T$

Construction d'un nouvel espace de projection ϕ^R_{new}

t.q. la différence entre les pertes de $\mathbf{x}^s$ et $\mathbf{x}^t$ est faible

$$\left| \left[1 - y \sum_{j=1}^{r} \alpha_j K(\mathbf{x}^s, \mathbf{x}'_j) \right]_+ - \left[1 - y \sum_{j=1}^{r} \alpha_j K(\mathbf{x}^t, \mathbf{x}'_j) \right]_+ \right| \leq \underbrace{\left\| ({}^t\phi^R(\mathbf{x}^s) - {}^t\phi^R(\mathbf{x}^t)) \mathrm{diag}(\boldsymbol{\alpha}) \right\|_1}_{\left\| {}^t\phi^R_{new}(\mathbf{x}^s) - {}^t\phi^R_{new}(\mathbf{x}^t) \right\|_1}$$

$$\Rightarrow \ \phi^R_{new}(\cdot) \ = \ \langle \ \underbrace{\alpha_1 K(\cdot, \mathbf{x}'_1)}_{K_{new}(\cdot, \mathbf{x}'_1)} \ , \ \ldots \ , \ \underbrace{\alpha_r K(\cdot, \mathbf{x}'_r)}_{K_{new}(\cdot, \mathbf{x}'_r)} \ \rangle$$

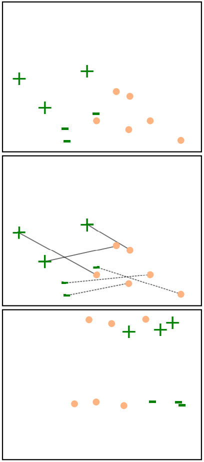
Algorithme DASF

Avec $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m_s} \sim (P_S)^{m_s}$ et $R = \{\mathbf{x}'_j\}_{j=1}^{r}$
Construction du ϕ^R-espace à l'aide $\boldsymbol{\alpha}$ appris par

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{m_s}\sum_{i=1}^{m_s} L\big(h, (\mathbf{x}_i^s, y_i^s)\big) + \lambda\|\boldsymbol{\alpha}\|_1$$

$$+ \beta \sum_{(\mathbf{x}^s,\mathbf{x}^t)\in\mathcal{C}_{ST}} \left\| \left({}^t\phi^R(\mathbf{x}^s) - {}^t\phi^R(\mathbf{x}^t)\right) \mathrm{diag}(\boldsymbol{\alpha}) \right\|_1$$

avec $L\big(h, (\mathbf{x}_i^s, y_i^s)\big) = \left[1 - y_i \sum_{j=1}^{r} \alpha_j K(\mathbf{x}_i^s, \mathbf{x}'_j)\right]_+$



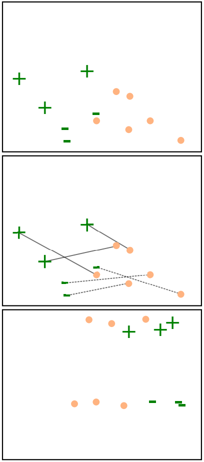
Algorithme DASF

Avec $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m_s} \sim (P_S)^{m_s}$ et $R = \{\mathbf{x}'_j\}_{j=1}^{r}$
Construction du ϕ^R-espace à l'aide $\boldsymbol{\alpha}$ appris par

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{m_s}\sum_{i=1}^{m_s} L\big(h, (\mathbf{x}_i^s, y_i^s)\big) + \lambda\|\boldsymbol{\alpha}\|_1$$

$$+\ \beta \sum_{(\mathbf{x}^s,\mathbf{x}^t)\in\mathcal{C}_{ST}} \Big\| \big({}^t\phi^R(\mathbf{x}^s) - {}^t\phi^R(\mathbf{x}^t)\big)\, \mathrm{diag}(\boldsymbol{\alpha}) \Big\|_1$$

$$\text{avec } L\big(h, (\mathbf{x}_i^s, y_i^s)\big) = \Big[1 - y_i \sum_{j=1}^{r} \alpha_j K(\mathbf{x}_i^s, \mathbf{x}'_j)\Big]_+$$



Tester tous les couples $(\mathbf{x}^s, \mathbf{x}^t) \in \mathcal{C}_{ST}$ est intraitable

$\Rightarrow$ Procédure itérative + Validation "circulaire"

L'intuition portée par la théorie

- Borne en généralisation

  $\hookrightarrow$ DASF est robuste sur le domaine **source**

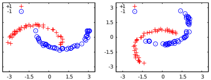
- Analyse de la parcimonie

  $\hookrightarrow$ la parcimonie dépend de l'écart entre les coordonnées

$\Rightarrow$ Les domaines sont éloignés

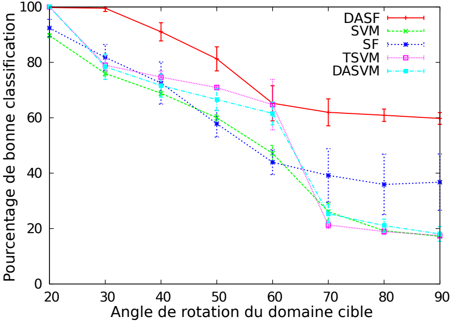
$\Rightarrow$ La différence entre les coordonnées est élevée

$\Rightarrow$ augmentation de la parcimonie



Problème jouet : "lunes jumelles"

- 1 domaine source
- 8 domaines cibles selon 8 angles de rotations
- 10 tirages aléatoire pour chaque angle
- Performances sur 1500 exemples cibles
- Noyau gaussien ou renormalisation (non SDP, non symétrique) du noyau



Taille	10	10	9	8	4	4	4	3

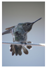
Annotation d'images

- Domaine source : PascalVOC 2007 avec un ratio $+/-$ de $1/3$
- Deux domaines cibles :
  - Ratio $+/-$ différent : PascalVOC 2007 Test
  - De même ratio $+/-$: TrecVid 2007
- F-mesure sur le domaine cible
- Noyau gaussien ou renormalisation (non SDP, non symétrique) du noyau

		SVM	SF	TSVM	DASVM	DASF
VOC vs VOC						
Moy. sur 20 conc.	F-mes.	0.22	0.19	0.17	0.20	**0.25**
	Taille	642	210	705	622	**200**
VOC vs Trec						
BOAT	F-mes.	0.56	0.49	0.56	0.52	**0.57**
	Taille	351	214	498	202	**120**
CAR	F-mes.	0.43	0.50	0.52	**0.55**	**0.55**
	Taille	1096	**176**	631	627	254
MONITOR	F-mes.	0.19	0.34	0.37	0.30	**0.42**
	Taille	698	246	741	523	**151**
PERSON	F-mes.	0.52	0.45	0.46	0.54	**0.57**
	Taille	951	226	1024	274	**19**
PLANE	F-mes.	0.32	0.54	0.61	0.52	**0.66**
	Taille	428	178	259	450	**7**
Moy. sur les 5 conc.	F-mes.	0.40	0.47	0.50	0.49	**0.55**
	Taille	705	208	631	415	**110**

VOC vs VOC : Points raisonnables pour le concept Personne



- Idée : Contrôle de la borne d'adaptation de domaine théorique
- Algorithme d'adaptation itératif
  - construit un nouvel espace de projection
  - rapproche les domaines tout en gardant de bonnes garanties sur le domaine source
  - en repondérant une fonction de similarité (ϵ,γ,τ)-bonne
- Pas d'étiquette cible (mais a été généralisé à l'utilisation d'étiquettes cibles)
- Avec des garanties en généralisation
  - l'algorithme est robuste sur le domaine source
- La parcimonie des modèles dépend de la difficulté du problème

Les inconvénients de cette approche

- le couplage $\Longrightarrow$ les points proches sont de même étiquette
- procédure itérative $\Longrightarrow$ compliquée et lourde à mettre en œuvre
- pas de justification théorique de la minimisation de la divergence
- la $\mathcal{H}$-divergence est difficile à optimiser en même temps que l'erreur source
- l'analyse classique peut être vue comme une analyse dans le pire cas

Les inconvénients de cette approche

- le couplage $\Longrightarrow$ les points proches sont de même étiquette
- procédure itérative $\Longrightarrow$ compliquée et lourde à mettre en œuvre
- pas de justification théorique de la minimisation de la divergence
- la $\mathcal{H}$-divergence est difficile à optimiser en même temps que l'erreur source
- l'analyse classique peut être vue comme une analyse dans le pire cas

Seconde contribution

Une première analyse PAC-Bayésienne de l'adaptation de domaine

Analyse en "moyenne" sur $\mathcal{H}$ $\Rightarrow$ Garanties pour les votes de majorité

Sommaire

$X \in \mathbb{R}^d$; $Y = \{-1, +1\}$ espace de sortie ; $\mathcal{H}$ ensemble de classifieurs ; P_S domaine sur $X \times Y$

Classification supervisée

Objectif : Trouver $h \in \mathcal{H}$ minimisant **l'erreur** : $R_{P_S}(h) = \mathop{\mathbf{E}}_{(\mathbf{x}^s, y^s) \sim P_S} \mathbf{I}\big[h(\mathbf{x}^s) \neq y^s\big]$

Approche PAC-Bayésienne [McAllester, 1999]

Objectif : Trouver le **vote de majorité** ρ**-pondéré** B_ρ sur $\mathcal{H}$ minimisant l'erreur $R_{P_S}(B_\rho)$

$$B_\rho(\mathbf{x}) = \mathrm{sign}\left[\sum_{h \in \mathcal{H}} \rho(h) h(\mathbf{x})\right]$$

où ρ est la distribution **posterior** sur $\mathcal{H}$ apprise à partir d'une distribution **prior** π sur $\mathcal{H}$

- La théorie PAC-Bayésienne **ne** borne **pas** directement $R_{P_S}(B_\rho)$

- mais l'erreur du **classifieur de Gibbs** G_ρ

  qui prédit l'étiquette de $\mathbf{x} \in X$ en

  - choisissant aléatoirement selon ρ un h
  - puis en retournant $h(\mathbf{x})$

- La théorie PAC-Bayésienne **ne** borne **pas** directement $R_{P_S}(B_\rho)$
- mais l'erreur du **classifieur de Gibbs** G_ρ
  qui prédit l'étiquette de $\mathbf{x} \in X$ en
  - choisissant aléatoirement selon ρ un h
  - puis en retournant $h(\mathbf{x})$

IMPORTANT — l'erreur de G_ρ correspond à l'espérance selon ρ des erreurs de $\mathcal{H}$:

$$R_{P_S}(G_\rho) = \mathop{\mathbf{E}}_{h\sim\rho} R_{P_S}(h)$$

On a : $R_{P_S}(B_\rho) \leq 2R_{P_S}(G_\rho)$

$\Rightarrow$ une majoration de $R_{P_S}(G_\rho)$ $\Rightarrow$ une majoration de $R_{P_S}(B_\rho)$

Théorème ([Catoni, 2007], comme énoncé dans [Germain et al., 2009])

Pour tout P_S sur $X \times Y$, pour tout $\mathcal{H}$, pour tout π sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, et pour tout $\mathbf{c} > 0$, avec une proba. d'au moins $1-\delta$ sur $S \sim (P_S)^{m_s}$, pour tout ρ sur $\mathcal{H}$, on a

$$\overbrace{R_{P_S}(G_\rho)}^{\text{erreur réelle}} \leq \frac{\mathbf{c}}{1-e^{-\mathbf{c}}}\left[\underbrace{R_S(G_\rho)}_{\text{erreur empirique}} + \underbrace{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{m_s \mathbf{c}}}_{f(\text{“complexité”, nb données})} \right]$$

avec $\mathrm{KL}(\rho\|\pi) = \mathbf{E}_{h\sim\rho} \ln \dfrac{\rho(h)}{\pi(h)}$ la divergence de Kullback-Leibler

- $\mathbf{c}$ contrôle le compromis entre $R_S(G_\rho)$ et $\frac{\mathrm{KL}(\rho\|\pi)}{m_s}$

Théorème ([Catoni, 2007], comme énoncé dans [Germain et al., 2009])

Pour tout P_S sur $X \times Y$, pour tout $\mathcal{H}$, pour tout π sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, et pour tout $\mathbf{c} > 0$, avec une proba. d'au moins $1-\delta$ sur $S \sim (P_S)^{m_s}$, pour tout ρ sur $\mathcal{H}$, on a

$$\overbrace{R_{P_S}(G_\rho)}^{\text{erreur réelle}} \leq \frac{\mathbf{c}}{1-e^{-\mathbf{c}}}\left[\underbrace{R_S(G_\rho)}_{\text{erreur empirique}} + \underbrace{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{m_s\mathbf{c}}}_{f(\text{``complexité'', nb données})}\right]$$

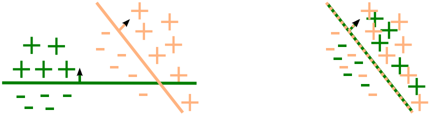
avec $\mathrm{KL}(\rho\|\pi) = \mathbf{E}_{h\sim\rho} \ln \dfrac{\rho(h)}{\pi(h)}$ la divergence de Kullback-Leibler

- $\mathbf{c}$ contrôle le compromis entre $R_S(G_\rho)$ et $\frac{\mathrm{KL}(\rho\|\pi)}{m_s}$
- L'algorithme PBGD : ensemble des classifieurs linéaires
  $\rho_\mathbf{w}$ gaussienne isotropique centrée en $\mathbf{w}$
  Étant donné $S \sim {P_S}^{m_s}$

$$\mathrm{argmin}_{\rho_\mathbf{w}} \ \mathbf{C}\, m_s\, R_S(G_{\rho_\mathbf{w}}) + \underbrace{\tfrac{1}{2}\|\mathbf{w}\|^2}_{\mathrm{KL}(\rho_\mathbf{w}\|\pi_\mathbf{0})}$$

Adaptation de domaine

Nécessité d'une divergence entre les domaines



Théorie PAC-Bayésienne

Vote de majorité ρ-pondéré

Analyse en "moyenne" / en espérance sur $\mathcal{H}$

Définition : ρ-désaccord entre D_S et D_T

$$\mathrm{dis}_\rho(D_S, D_T) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[R_{D_T}(h,h') - R_{D_S}(h,h') \right] \right|$$

où : $\displaystyle \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_D(h,h') = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x}) \right]$

Définition : ρ-désaccord entre D_S et D_T

$$\mathrm{dis}_\rho(D_S, D_T) = \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[R_{D_T}(h,h') - R_{D_S}(h,h') \right] \right|$$

où : $\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_D(h,h') = \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right]$

En suivant l'idée portée par la C-borne [Lacasse et al., 2007] :

$$R_{P_S}(B_\rho) \leq 1 - \frac{\left(1 - 2R_{P_S}(G_\rho)\right)^2}{1 - 2 \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h')}$$

L'intuition est

Si $R_{P_S}(G_\rho) \simeq R_{P_T}(G_\rho)$

$\Rightarrow R_{P_S}(B_\rho) \simeq R_{P_T}(B_\rho)$ **quand** $\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h') \simeq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_T}(h,h')$

Théorème

Soit $\mathcal{H}$ un espace d'hypothèses. Pour toute distribution ρ sur $\mathcal{H}$, on a

$$R_{P_T}(G_\rho) \ \leq \ R_{P_S}(G_\rho) \ + \ \mathrm{dis}_\rho(D_S, D_T) \ + \ \nu_\rho$$

avec $\rho_T^* = \mathrm{argmin}_\rho \ R_{P_T}(G_\rho)$ est le meilleur posterior sur le domaine cible
et $\nu_\rho = R_{P_T}(G_{\rho_T^*}) + \underset{h\sim\rho}{\mathbf{E}} \ \underset{h'\sim\rho_T^*}{\mathbf{E}} \ [R_{D_T}(h, h') + R_{D_S}(h, h')]$

$$R_{P_T}(G_\rho) \leq R_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T) + \nu_\rho \quad \textbf{Vs} \quad R_{P_T}(h) \leq R_{P_S}(h) + \tfrac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu$$

- notre borne porte sur l'erreur du classifieur de Gibbs : $R_{P_T}(G_\rho) = \mathbf{E}_{h\sim\rho} R_{P_T}(h)$
- la notion de divergence diffère

Rappel :
$$\begin{array}{rcl} \mathrm{dis}_\rho(D_S, D_T) & = & \left| \mathbf{E}_{(h,h')\sim\rho^2} \left[R_{D_T}(h,h') - R_{D_S}(h,h') \right] \right| \\ \frac{1}{2} d_{\mathcal{H}}(D_S, D_T) & = & \sup_{(h,h')\in\mathcal{H}^2} \left| R_{D_T}(h,h') - R_{D_S}(h,h') \right| \end{array}$$

  - $\frac{1}{2} d_{\mathcal{H}}(D_S, D_T)$ est une divergence "dans le pire cas"
  - $\mathrm{dis}_\rho(D_S, D_T)$ est spécifique au classifieur G_ρ considéré
  - on a : $\frac{1}{2} d_{\mathcal{H}}(D_S, D_T) \geq \mathrm{dis}_\rho(D_S, D_T)$

Théorème

Soit $\mathcal{H}$ un espace d'hypothèses. Pour toute distribution ρ sur $\mathcal{H}$, on a

$$R_{P_T}(G_\rho) \ \leq \ R_{P_S}(G_\rho) \ + \ \mathrm{dis}_\rho(D_S, D_T) \ + \ \cancel{\nu_\rho}$$

avec $\rho_T^* = \mathrm{argmin}_\rho \ R_{P_T}(G_\rho)$ est le meilleur posterior sur le domaine cible
et $\nu_\rho = R_{P_T}(G_{\rho_T^*}) + \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho_T^*} \left[R_{D_T}(h,h') + R_{D_S}(h,h')\right]$

Notre solution pour l'adaptation PAC-Bayésienne

Minimisation **simultanée** de $R_{P_S}(G_\rho)$ et $\mathrm{dis}_\rho(D_S, D_T)$ avec des justifications théoriques

Théorème : Borne en généralisation PAC-Bayésienne (à la Catoni)

Pour tout P_S et P_T sur $X \times Y$, pour tout $\mathcal{H}$, pour tout π sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, pour tout $\mathbf{a} > 0$, $\mathbf{c} > 0$, avec une proba. d'au moins $1-\delta$ sur $S \times T \sim (P_S \times D_T)^m$, on a

$$\forall \rho \sim \mathcal{H}, \quad \overbrace{R_{P_T}(G_\rho)}^{\text{erreur réelle cible}} \leq \mathbf{c}' \underbrace{R_S(G_\rho)}_{\text{erreur empirique source}} + \mathbf{a}' \underbrace{\mathrm{dis}_\rho(S,T)}_{\rho\text{-desaccord empirique}} + \left(\frac{\mathbf{c}'}{\mathbf{c}} + \frac{2\,\mathbf{a}'}{\mathbf{a}}\right) \underbrace{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{\delta}}{m}}_{f(\text{"complexité"},\ \text{nb données})} + \nu_\rho + \alpha' - 1$$

où $\mathbf{c}' = \frac{\mathbf{c}}{1-e^{-\mathbf{c}}}$ et $\mathbf{a}' = \frac{2\mathbf{a}}{1-e^{-2\mathbf{a}}}$

$\Longrightarrow$ Comme pour PBGD : spécialisation aux classifieurs linéaires

L'algorithme PBDA

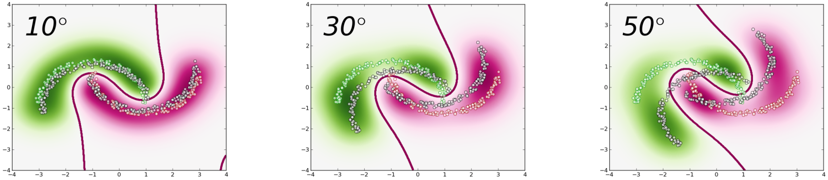
Étant donné $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (P_S)^m$, $T = \{\mathbf{x}_i^t\}_{i=1}^m \sim (P_T)^m$

$$\operatorname{argmin}_{\rho_{\mathbf{w}}} \; C\, m\, R_S(G_{\rho_{\mathbf{w}}}) + A\, m\, \mathrm{dis}_{\rho_{\mathbf{w}}}(S, T) + \underbrace{\frac{1}{2}\left\|\mathbf{w}\right\|^2}_{\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})}$$

où $\mathrm{dis}_{\rho_{\mathbf{w}}}(S, T) = \left| \underset{(h,h')\sim\rho_{\mathbf{w}}^2}{\mathbf{E}} R_S(h, h') - \underset{(h,h')\sim\rho_{\mathbf{w}}^2}{\mathbf{E}} R_T(h, h') \right|$

Problème jouet : "les lunes jumelles"

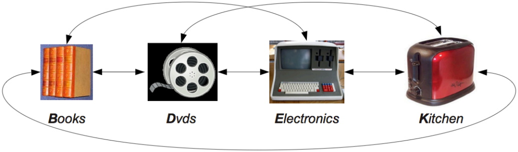
- 1 domaine source
- 8 domaines cibles selon 8 angles de rotations
- 10 tirages aléatoire pour chaque angle
- Performances sur un ensemble de test de 1 500 exemples cibles
- Noyau gaussien



(a) (b) (c)

Avis sur des produits *Amazon.com*

- 12 tâches d'adaptation d'un type vers un autre (*ex.* books → kitchen)



  - Domaine source : 2 000 exemples étiquetés
  - Domaine cible : 2 000 exemples non étiquetés
  - Performances sur l'échantillon cible de test : entre 3 000 et 6 000 exemples
- Dimension des données : $\sim 40,000$
- Noyau linéaire

	Erreur moyenne
PBGD3 [Germain et al., 2009]	0.226
SVM	0.231
DASF	0.275
DASVM [Bruzzone and Marconcini, 2010]	*0.204*
CODA [Chen et al., 2011]	0.210
PBDA (5× *plus rapide*)	0.208

Conclusion

- La première analyse PAC-Bayésienne de l'adaptation de domaine
  - exprimée comme une espérance selon ρ sur une classe d'hypothèses
  - une divergence qui dépend de ρ
  - directement optimisable (avec des justifications théoriques)
- Un premier algorithme spécialisé aux classifieurs linéaires
  - résultats prometteurs

Points forts

- Minimisation directe divergence/erreur source
- Nouvelles pistes pour l'adaptation de domaine

Conclusion

- La première analyse PAC-Bayésienne de l'adaptation de domaine
  - exprimée comme une espérance selon ρ sur une classe d'hypothèses
  - une divergence qui dépend de ρ
  - directement optimisable (avec des justifications théoriques)
- Un premier algorithme spécialisé aux classifieurs linéaires
  - résultats prometteurs

Points forts

- Minimisation directe divergence/erreur source
- Nouvelles pistes pour l'adaptation de domaine

Perspectives

- Utilité de distribution(s) prior
- Traitement du terme ν_ρ *(une première amélioration de la borne)*

Sommaire

Conclusion

- DASF $\rightarrow$ un algorithme “classique” d’adaptation
- PBDA $\rightarrow$ un premier algorithme pour l’adaptation PAC-Bayésienne

Mais aussi :

- P-MinCq $\rightarrow$ extension d’un algorithme PAC-Bayésien pour considérer un prior
  $\rightarrow$ minimise la C-borne
- Étude PAC-Bayésienne du multiclasse $\rightarrow$ met en jeu la matrice de confusion
  $\rightarrow$ généralisation de la C-borne

Perspectives générales

Comment dériver un algorithme pour **plusieurs classes, des sorties structurées** ?

$\hookrightarrow$ la matrice de confusion

$\hookrightarrow$ vote de majorité multiclasse

$\hookrightarrow$ code correcteur

Perspectives générales

Comment dériver un algorithme pour **plusieurs classes, des sorties structurées** ?

↪ la matrice de confusion

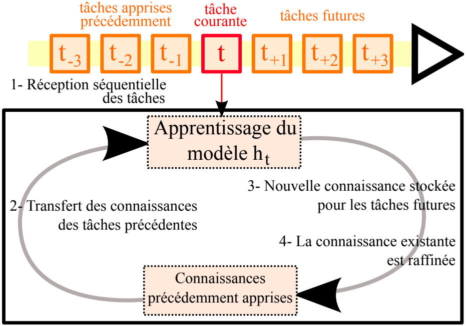
↪ vote de majorité multiclasse

↪ code correcteur

Comment utiliser des modèles de **tâches** précédentes/simultanées ?

Comment tirer parti de **connaissances** ?

Comment **valider** les hyperparamètres ?



Le long-life learning

Perspectives générales

Comment dériver un algorithme pour **plusieurs classes, des sorties structurées** ?

↪ la matrice de confusion

↪ vote de majorité multiclasse

↪ code correcteur

Comment utiliser des modèles de **tâches** précédentes/simultanées ?

↪ vote de majorité

↪ apprentissage en ligne

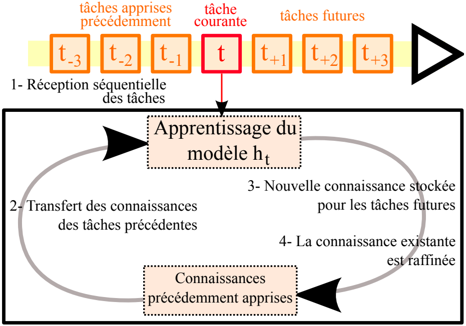
Comment tirer parti de **connaissances** ?

↪ notion de prior et PAC-Bayes

↪ divergence entre tâches

Comment **valider** les hyperparamètres ?

↪ une validation "PAC-Bayésienne"



Le long-life learning

Merci pour votre attention.

References

Balcan, M. F., Blum, A. and Srebro, N. (2008).
Improved Guarantees for Learning via Similarity Functions.
In Proceedings of Annual Conference on Computational Learning Theory pp. 287–298,.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Vaughan, J. (2010).
A theory of learning from different domains.
Machine Learning Journal *79*, 151–175.

Blitzer, J., Foster, D. and Kakade, S. (2011).
Domain adaptation with coupled subspaces.
Journal of Machine Learning Research-Proceedings Track *15*, 173–181.

Blitzer, J., McDonald, R. and Pereira, F. (2006).
Domain Adaptation with Structural Correspondence Learning.
In Proceedings of Conference on Empirical Methods on Natural Language Processing pp. 120–128,.

Bruzzone, L. and Marconcini, M. (2010).
Domain Adaptation Problems : A DASVM Classification Technique and a Circular Validation Strategy.
IEEE Transactions on Pattern Analysis and Machine Intelligence *32*, 770–787.

Catoni, O. (2007).
PAC-Bayesian supervised classification : the thermodynamics of statistical learning, vol. 56,.
Institute of Mathematical Statistic.

Chen, M., Weinberger, K. Q. and Blitzer, J. (2011).
Co-Training for Domain Adaptation.
In Proceedings of Annual Conference on Neural Information Processing Systems pp. 2456–2464,.

Daumé III, H. (2007).
Frustratingly Easy Domain Adaptation.
In ACL.

Daumé III, H., Kumar, A. and Saha, A. (2010).
Co-regularization Based Semi-supervised Domain Adaptation.
In NIPS pp. 478–486,.

Germain, P., Lacasse, A., Laviolette, F. and Marchand, M. (2009).
PAC-Bayesian Learning of Linear Classifiers.
In Proceedings of International Conference on Machine Learning.

Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M. and Scholkopf, B. (2007).
Correcting sample selection bias by unlabeled data.
Advances in neural information processing systems *19*, 601.

Jiang, J. and Zhai, C. (2007).
Instance Weighting for Domain Adaptation in NLP.
In Proceedings of Association for Computational Linguistics.

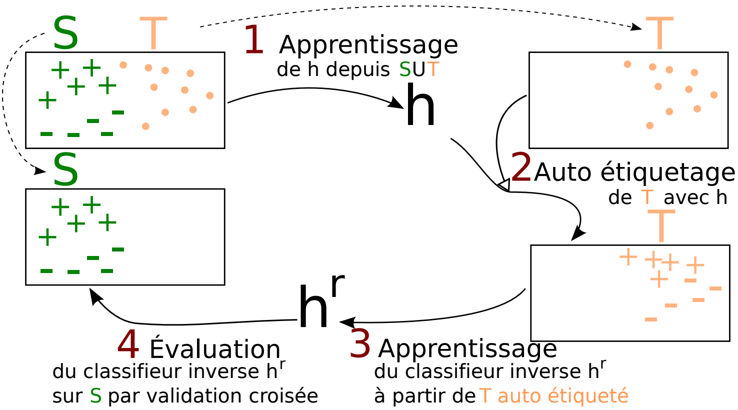
Lacasse, A., Laviolette, F., Marchand, M., Germain, P. and Usunier, N. (2007).
PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier.
In Proceedings of annual conference on Neural Information Processing Systems.

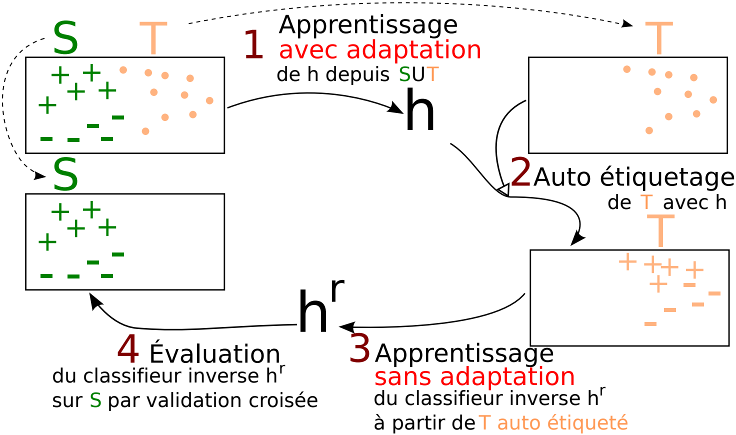
Mansour, Y., Mohri, M. and Rostamizadeh, A. (2009a).
Domain Adaptation : Learning Bounds and Algorithms.
In Proceedings of Annual Conference on Learning Theory pp. 19–30,.

Mansour, Y., Mohri, M. and Rostamizadeh, A. (2009b).
Multiple Source Adaptation and the Rényi Divergence.
In Proceedings of annual Conference on Uncertainty in Artificial Intelligence pp. 367–374,.

McAllester, D. A. (1999).
Some PAC-Bayesian Theorems.
Machine Learning Journal *37*, 355–363.

Xu, H. and Mannor, S. (2010).
Robustness and Generalization.
In Proceedings of Annual Conference on Computational Theory pp. 503–515,.


S
T
1 Apprentissage
de h depuis SUT
h
T
2 Auto étiquetage
de T avec h
S
T
h^r
4 Évaluation
du classifieur inverse h^r
sur S par validation croisée
3 Apprentissage
du classifieur inverse h^r
à partir de T auto étiqueté


S
T
1 Apprentissage
avec adaptation
de h depuis SUT
h
T
2 Auto étiquetage
de T avec h
S
T
hr
4 Évaluation
du classifieur inverse hr
sur S par validation croisée
3 Apprentissage
sans adaptation
du classifieur inverse hr
à partir de T auto étiqueté

- Algorithme robuste [Xu and Mannor, 2010]
  - Idée : *"if a testing sample is similar to a training sample then the testing error is close to the training error"* (sans adaptation)

⇒ DASF est robuste sur le domaine source

$$R^L_{P_T}(h) \le R^L_S(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2\ln\frac{1}{\delta}}{m_s}} + \frac{1}{2} d_{\mathcal{H}}(D_S, D_T) + \nu,$$

Lemme

Soit $B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}^s, \mathbf{x}^t) \in \mathcal{C}_{ST}} |K(\mathbf{x}^s, \mathbf{x}'_j) - K(\mathbf{x}^t, \mathbf{x}'_j)| \right\} > 0.$

Si $\boldsymbol{\alpha}^*$ est la solution optimale du problème, alors

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\boldsymbol{\beta} B_R + \boldsymbol{\lambda}}.$$

La parcimonie dépend des hyperparameters et de B_R

$\Rightarrow$ Les domaines sont éloignés

$\Rightarrow$ La différence entre les coordonnées est élevée

$\Rightarrow$ B_R tend à croître

$\Rightarrow$ augmentation de la parcimonie