UNIVERSITÉ LAVAL

Groupe de Recherche en Apprentissage Automatique de Laval

Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine

Soutenance de thèse en informatique

Pascal Germain

Département d'informatique et génie logiciel
Université Laval, Québec, Canada

11 juin 2015

Plan

Plan

1. Mise en contexte
   - Apprentissage automatique et classification
   - Les classificateurs de vote de majorité
2. L'apprentissage inductif revisité
   - Théorème PAC-bayésien ≪classique≫
   - Théorème PAC-bayésien général
3. Généralisations de la théorie PAC-bayésienne
   - Apprentissage transductif
   - Adaptation de domaine
4. Conclusion et travaux futurs

Une définition de l'apprentissage automatique

≪ *Field of study that gives computers the ability to learn without being explicitly programmed* ≫

– Arthur Samuel, 1959

Exemple

critiques de films

-1 **An insult to Douglas Adams' memory**
I agree entirely with "darkgenius'" comments. This movie is a travesty of the book and the TV series; a cutesy version totally lacking in the wit and satire of the original. Read more
Published 5 months ago by John W Beare

+1 **Don't Panic!**
If you haven't listened to the BBC radio-play, this isn't bad! Purists, no doubt, will dispute my verdict but the fact of the matter is THGTTG (see title) does have Douglas Adams'...
Read more
Published on Mar 13 2011 by Sid Matheson

+1 **On Blu-ray, even better**
I've seen this movie on TV and wanted to add it to my collection. I couldn't find it locally so when I saw it on amazon and on Blu-ray, I picked it up. Read more
Published on April 18 2009 by J. W. Little

-1 **An insult to Douglas Adams' memory**
The filmmaker's reverence for Adams' legacy? What kind of rubbish statement is that? As a loyal fan of Douglas Adams for more than a quarter of a century, I was appalled and...
Read more
Published on Aug 22 2006 by Daniel Jolley

??? **Mindbending**
I will not recommend this movie for people who haven't read at least two or three of Douglas Adams' books on hitchhiking. Read more
Published on Mar 28 2006 by alper bac

Exemple

critiques de films

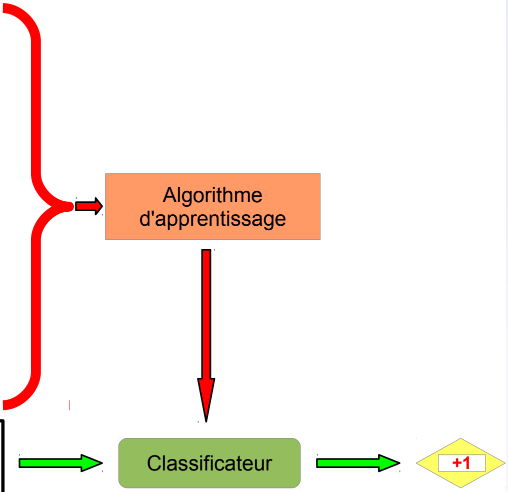
-1 **An insult to Douglas Adams' memory**
I agree entirely with "darkgenius'" comments. This movie is a travesty of the book and the TV series; a cutesy version totally lacking in the wit and satire of the original. Read more
Published 5 months ago by John W Beare

+1 **Don't Panic!**
If you haven't listened to the BBC radio-play, this isn't bad! Purists, no doubt, will dispute my verdict but the fact of the matter is THGTTG (see title) does have Douglas Adams'... Read more
Published on Mar 13 2011 by Sid Matheson

+1 **On Blu-ray, even better**
I've seen this movie on TV and wanted to add it to my collection. I couldn't find it locally so when I saw it on amazon and on Blu-ray, I picked it up. Read more
Published on April 18 2009 by J. W. Little

-1 **An insult to Douglas Adams' memory**
The filmmaker's reverence for Adams' legacy? What kind of rubbish statement is that? As a loyal fan of Douglas Adams for more than a quarter of a century, I was appalled and... Read more
Published on Aug 22 2006 by Daniel Jolley

??? **Mindbending**
I will not recommend this movie for people who haven't read at least two or three of Douglas Adams' books on hitchhiking. Read more
Published on Mar 28 2006 by alper bac



Le pourquoi et le comment

Nombreuses applications de l'apprentissage automatique

- Classification de texte
- Reconnaissance de la parole
- Recherche en bio-informatique
- ...

Un problème d'actualité

- Grandes quantités de données à traiter
- Grandes capacités de traitement de l'information

Mon approche

- Mieux comprendre le problème à l'aide d'outils mathématiques
- Formuler des garanties statistiques
- Concevoir de nouveaux algorithmes d'apprentissage

Définitions de base

Exemple d'apprentissage

Un exemple $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une **paire description-étiquette**.

Distribution génératrice des données

Chaque exemple est une **observation d'une distribution** D sur $\mathcal{X} \times \mathcal{Y}$.

Échantillon d'apprentissage

$$S \stackrel{\text{def}}{=} \{\, (x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_m, y_m) \,\} \ \sim\ D^m$$

Classificateur

$$h : \mathcal{X} \to \mathcal{Y}$$

Classificateur binaire

$$h : \mathcal{X} \to \{-1, +1\}$$

Algorithme d'apprentissage

$$A(S) \longrightarrow h$$

Risque d'un classificateur

Risque (ou erreur de généralisation)

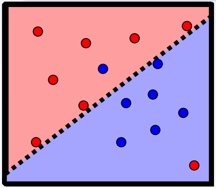
Probabilité d'erreur sur un exemple généré par la distribution D :

$$R_D(h) \stackrel{\text{def}}{=} \Pr_{(x,y)\sim D}\left(h(x) \neq y\right) = \mathbf{E}_{(x,y)\sim D} \mathrm{I}\left[y \cdot h(x) \leq 0\right],$$

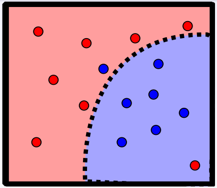
Risque empirique

Taux d'erreur sur l'échantillon d'apprentissage $S \sim D^m$:

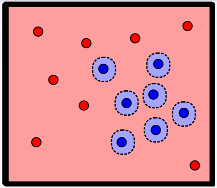
$$R_S(h) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}\left[y_i \cdot h(x_i) \leq 0\right].$$



$R_S(h) = \frac{2}{15} \simeq 13\%$



$R_S(h) = \frac{2}{15} \simeq 13\%$



$R_S(h) = \frac{0}{15} = 0\%$

Problème de l'estimation du risque

Afin d'évaluer la qualité d'un classificateur h, nous désirons connaître son risque $R_D(h)$.

Borne de type PAC (Probablement approximativement correctes)

Avec probabilité «$1-\delta$», le risque du classificateur h est inférieur à «ϵ»

$$\Pr\Big(R_D(h) \leq \epsilon\Big) \;\geq\; 1-\delta$$

Deux catégories de garanties de généralisation

1. Bornes sur l'échantillon de test ;
2. Bornes sur l'échantillon d'entraînement.

Théorie PAC-bayésienne

Initiée par David McAllester (1999), la théorie PAC-bayésienne permet de formuler des garanties sur le risque de **votes de majorité** de classificateurs.

Inspiration bayésienne

Permet d'incorporer des connaissances *a priori* sur le problème d'apprentissage

Bornes sur l'échantillon d'entraînement

- Permettent d'obtenir des garanties sur l'acuité des classificateurs **sans recourir à un ensemble test**
- Source d'inspiration pour la conception de **nouveaux algorithmes d'apprentissage**.

Les classificateurs de vote de majorité

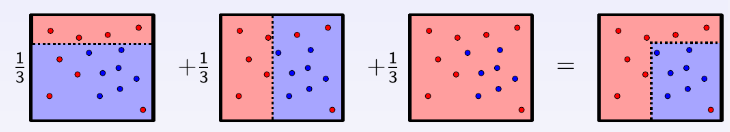
Étant donné :

- Un ensemble de **votants** $\mathcal{H} = \{h_1, h_2, h_3, \ldots\}$;
- Une distribution de **poids** Q sur $\mathcal{H}$.

Vote de majorité

Pour classifier x, le classificateur considère l'*opinion majoritaire* parmi $\mathcal{H}$

$$B_Q(x) \stackrel{\text{def}}{=} \operatorname{sgn}\left(\mathop{\mathbf{E}}_{h\sim Q} h(x)\right)$$



Plusieurs algorithmes d'apprentissage construisent des votes de majorité

AdaBoost, Random Forests, Bagging, ...

Risques

Étant donné

- Une distribution de données D sur $\mathcal{X} \times \mathcal{Y}$
- Une distribution Q sur un ensemble de votants $\mathcal{H}$

Risque du vote de majorité (risque de Bayes)

$$R_D(B_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} \mathrm{I}\left[\mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x) \leq 0 \right]$$

Risque de Gibbs

Le classificateur de Gibbs $G_Q(x)$ pige un h selon Q et retourne $h(x)$.

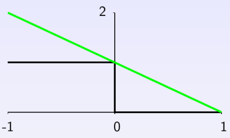
$$R_D(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} \mathop{\mathbf{E}}_{h\sim Q} \mathrm{I}\left[y\cdot h(x) \leq 0 \right]$$

$$= \mathop{\mathbf{E}}_{(x,y)\sim D} \left(\frac{1}{2} - \frac{1}{2} \mathop{\mathbf{E}}_{h\sim Q} y\cdot h(x) \right)$$

Facteur 2

Il est connu dans la littérature que

$$R_D(B_Q) \leq 2 \times R_D(G_Q)$$



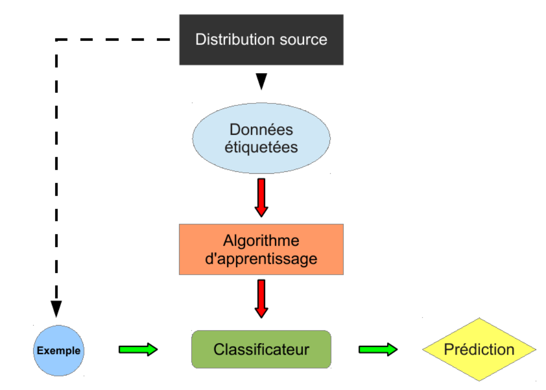
Plan

Apprentissage inductif

Hypothèse

Les exemples sont générées *i.i.d.* par une distribution D sur $\mathcal{X} \times \mathcal{Y}$.

$$S \;=\; \{\, (x_1, y_1),\; (x_2, y_2),\; \ldots,\; (x_m, y_m) \,\} \;\sim\; D^m$$



Théorème PAC-bayésien ≪classique≫

Ingrédients de la théorie PAC-bayésiennes

- Le **risque empirique du classificateur de Gibbs** G_Q :

$$R_S(G_Q) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \left(\frac{1}{2} - \frac{1}{2} \mathop{\mathbf{E}}_{h\sim Q} y_i \cdot h(x_i) \right)$$

- La **divergence Kullback-Leibler** entre le *prior* P et le *posterior* Q :

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}$$

Théorème PAC-bayésien (*McAllester, 2003*)

Pour toute distribution D sur $\mathcal{X} \times \mathcal{Y}$, pour tout ensemble $\mathcal{H}$ de votants, pour toute distribution P sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, on a, avec probabilité au moins $1-\delta$ sur le choix de $S \sim D^m$,

$$\forall\, Q \text{ sur } \mathcal{H} : \quad R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln \frac{2\sqrt{m}}{\delta}\right]}$$

Théorie PAC-bayésienne pour le risque de Gibbs

Δ-fonction : «distance» entre le $R_S(G_Q)$ et $R_D(G_Q)$

Fonction $\Delta : [0,1] \times [0,1] \to \mathbb{R}$ convexe.

Théorème général

Pour toute distribution D sur $\mathcal{X} \times \mathcal{Y}$, pour tout ensemble $\mathcal{H}$ de votants, pour toute distribution P sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, et pour toute Δ-fonction, on a, avec probabilité au moins $1-\delta$ sur le choix de $S \sim D^m$,

$$\forall\, Q \text{ sur } \mathcal{H} : \quad \Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \;\leq\; \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_\Delta(m)}{\delta}\right],$$

où

$$\mathcal{I}_\Delta(m) \;\stackrel{\text{def}}{=}\; \sup_{r \in [0,1]} \left[\sum_{k=0}^{m} \mathbf{Bin}\big(k; m, r\big)\, e^{m\Delta\left(\frac{k}{m}, r\right)}\right]$$

Théorème

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ sur } \mathcal{H} :\ \Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(m)}{\delta}\right]\right) \ \ge\ 1-\delta\,.$$

Démonstration

$$m\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} R_S(h),\ \mathop{\mathbf{E}}_{h\sim Q} R_D(h)\Big)$$

Inégalité de Jensen
$$\le \mathop{\mathbf{E}}_{h\sim Q} m\cdot\Delta\Big(R_S(h), R_D(h)\Big)$$

Changement de mesure
$$\le \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{m\Delta\left(R_S(h), R_D(h)\right)}$$

Inégalité de Markov
$$\le_{1-\delta} \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta(R_{S'}(h), R_D(h))}$$

Inversion des espérances
$$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^m} e^{m\cdot\Delta(R_{S'}(h), R_D(h))}$$

Loi binomiale
$$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \mathbf{Bin}\Big(k; m, R_D(h)\Big)\, e^{m\cdot\Delta(\frac{k}{m}, R_D(h))}$$

Supremum sur le risque
$$\le \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \sup_{r\in[0,1]}\left[\sum_{k=0}^{m} \mathbf{Bin}\Big(k; m, r\Big)\, e^{m\Delta(\frac{k}{m}, r)}\right]$$

$$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\,\mathcal{I}_\Delta(m)\,.$$

□

Corollaire

[...] avec probabilité au moins $1-\delta$ *sur le choix de* $S \sim D^m$,

$\forall\, Q$ *sur* $\mathcal{H}$:

(a) $\displaystyle \mathrm{kl}\Big(R_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right],$ *(Langford et Seeger, 2001)*

(b) $\displaystyle R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]},$ *(McAllester, 1999)*

(c) $\displaystyle R_D(G_Q) \leq \frac{1}{1-e^{-c}}\left(c \cdot R_S(G_Q) + \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right).$ *(Catoni, 2007)*

$$\mathrm{kl}(q,p) \overset{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \geq 2(q-p)^2\,,$$

$$\Delta_c(q,p) \overset{\text{def}}{=} -\ln[1-(1-e^{-c})\cdot p] - c\cdot q\,.$$

Théorie PAC-bayésienne pour l'espérance de désaccord

Espérance de désaccord

$$d_Q^D \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\ \mathop{\mathbf{E}}_{h_1\sim Q}\ \mathop{\mathbf{E}}_{h_2\sim Q} \mathrm{I}\Big[h_1(x) \neq h_2(x_i)\Big]$$

Théorème général

[...] avec probabilité au moins $1-\delta$ sur le choix de $S \sim D^m$,

$$\forall\, Q \text{ sur } \mathcal{H} : \quad \Delta\Big(d_Q^S,\, d_Q^D\Big) \ \leq\ \frac{1}{m}\left[2\,\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(m)}{\delta}\right],$$

Corollaire

(a) $\mathrm{kl}\Big(d_Q^S,\, d_Q^D\Big) \leq \frac{1}{m}\left[2\,\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]$,

(b) $d_Q^D \leq d_Q^S + \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]}$,

(c) $d_Q^D \leq \frac{1}{1-e^{-c}}\left(c\cdot d_Q^S + \frac{1}{m}\left[2\,\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right)$.

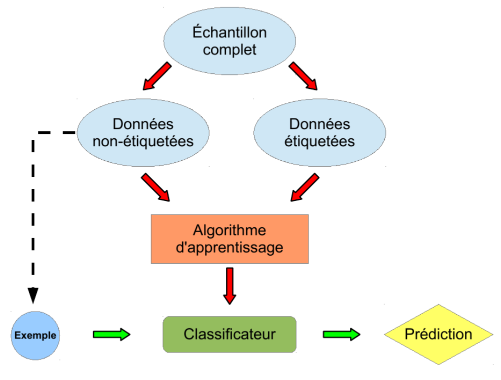
Plan

Apprentissage transductif

Hypothèse

Les données sont pigés sans remise d'un ensemble complet Z de taille N.

$$\begin{array}{rclllll} S & = & \{\,(x_1,y_1), & (x_2,y_2), & \ldots, & (x_m,y_m)\,\} & \subset Z \\ U & = & \{\,(x_{m+1},\cdot), & (x_{m+2},\cdot), & \ldots, & (x_N,\cdot)\,\} & = Z \setminus S \end{array}$$



Théorème général pour l'apprentissage transductif

Cadre inductif $\Rightarrow$ m piges avec remises selon D $\Rightarrow$ Loi binomiale.

Cadre transductif $\Rightarrow$ m piges sans remises dans Z $\Rightarrow$ Loi hypergéométrique.

Théorème

Pour tout échantillon de données Z contenant N exemples, pour tout ensemble $\mathcal{H}$ de votants, pour toute distribution P sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, et pour toute Δ-fonction, on a, avec probabilité au moins $1-\delta$ sur le choix S de m exemples parmi Z,

$$\forall\, Q \text{ sur } \mathcal{H} : \quad \Delta\big(R_S(G_Q), R_Z(G_Q)\big) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{T}_\Delta(m,N)}{\delta}\right],$$

où

$$\mathcal{T}_\Delta(m,N) \stackrel{\text{def}}{=} \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{m,N,K}} \frac{\binom{K}{k}\binom{N-K}{m-k}}{\binom{N}{m}} e^{m\Delta\left(\frac{k}{m},\frac{K}{N}\right)}\right],$$

et $\quad \mathcal{K}_{m,N,K} \stackrel{\text{def}}{=} \big\{\max[0, K+m-N], \ldots, \min[m,K]\big\}.$

Théorème

$$\Pr_{S\sim[Z]^m}\left(\forall\, Q \text{ sur } \mathcal{H} :\ \Delta\Big(R_S(G_Q), R_Z(G_Q)\Big) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{T}_\Delta(m,N)}{\delta}\right]\right) \ge 1-\delta\,.$$

Démonstration

$$m\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R_Z(h)\Big)$$

Inégalité de Jensen
$$\le \mathop{\mathbf{E}}_{h\sim Q} m\cdot\Delta\Big(R_S(h), R_Z(h)\Big)$$

Changement de mesure
$$\le \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{m\Delta\left(R_S(h),R_Z(h)\right)}$$

Inégalité de Markov
$$\le_{1-\delta} \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim[Z]^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta(R_{S'}(h),R_Z(h))}$$

Inversion des espérances
$$= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim[Z]^m} e^{m\cdot\Delta(R_{S'}(h),R_Z(h))}$$

Loi hypergéométrique
$$= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k\in\mathcal{K}_{m,N,N\cdot R_Z(h)}} \frac{\binom{N\cdot R_Z(h)}{k}\binom{N-N\cdot R_Z(h)}{m-k}}{\binom{N}{m}} e^{m\cdot\Delta(\frac{k}{m},R_Z(h))}$$

Supremum sur le risque
$$\le \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{m,N,K}} \frac{\binom{K}{k}\binom{N-K}{m-k}}{\binom{N}{m}} e^{m\Delta(\frac{k}{m},\frac{K}{N})}\right]$$

$$= \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\mathcal{T}_\Delta(m,N)\,. \qquad \square$$

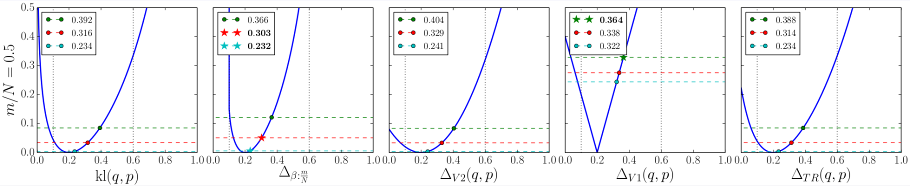
Choix de la Δ-fonction

Théorème

$$\Pr_{S\sim[Z]^m}\left(\forall\, Q \text{ sur } \mathcal{H} :\ \Delta\Big(R_S(G_Q), R_Z(G_Q)\Big) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{T}_\Delta(m,N)}{\delta}\right]\right) \ge 1-\delta\,.$$

On peut évaluer numériquement $\mathcal{T}_\Delta(m,N)$ pour toute Δ-fonction.

$$\mathcal{T}_\Delta(m,N) \overset{\text{def}}{=} \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{m,N,K}} \frac{\binom{K}{k}\binom{N-K}{m-k}}{\binom{N}{m}} e^{m\Delta\left(\frac{k}{m},\frac{K}{N}\right)}\right].$$



Une Δ-fonction pour le cas transductif

Cadre inductif (inspiré par Maurer, 2004)

$$\mathrm{kl}(q,p) \stackrel{\mathrm{def}}{=} q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \geq 2(q-p)^2 \quad \Rightarrow \quad \mathcal{I}_{\mathrm{kl}}(m) \leq 2\sqrt{m}$$

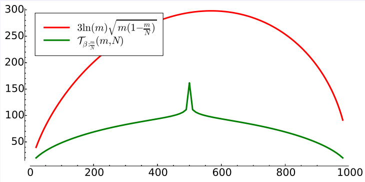
Cadre transductif

$$\Delta_\beta(q,p) = \mathrm{kl}(q,p) + \frac{1-\beta}{\beta} \mathrm{kl}\left(\frac{p-\beta q}{1-\beta}, p\right).$$

Théorème

Soit m et N des entiers tels que $20 \leq m \leq N-20$, alors

$$\mathcal{T}_{\beta:\frac{m}{N}}(m,N) \leq 3\ln(m)\sqrt{m\left(1-\frac{m}{N}\right)}.$$



Borne sur le risque de Gibbs

Corollaire

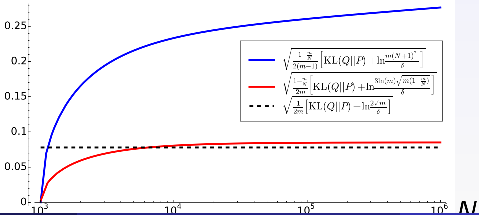
[...] avec probabilité au moins $1-\delta$ *sur le choix* S *de* m *exemples parmi* Z*,*

$\forall\, Q$ *sur* $\mathcal{H}$:

$$R_Z(G_Q) \;\leq\; R_S(G_Q) + \sqrt{\frac{1-\frac{m}{N}}{2m}\left[\mathrm{KL}(Q\|P) + \ln \frac{3\ln(m)\sqrt{m(1-\frac{m}{N})}}{\delta}\right]}\,.$$

Théorème (Derbeko et al., 2004)

$$R_Z(G_Q) \;\leq\; R_S(G_Q) + \sqrt{\frac{1-\frac{m}{N}}{2(m-1)}\left[\mathrm{KL}(Q\|P) + \ln \frac{m(N+1)^7}{\delta}\right]}\,.$$



Comparaison empirique avec Derbeko et al., 2004

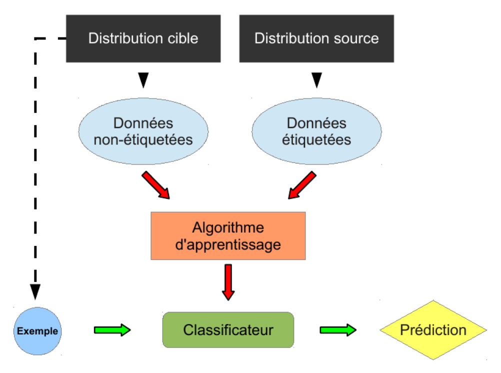
Ensemble de données	N	m/N	$R_S(G_Q)$	Nous	Derbeko
car	1728	0.1	0.193	**0.555**	0.793
		0.5	0.179	**0.418**	0.496
letter_AB	1555	0.1	0.146	**0.469**	0.718
		0.5	0.171	**0.402**	0.485
mushroom	8124	0.1	0.202	**0.486**	0.609
		0.5	0.205	**0.439**	0.479
nursery	12959	0.1	0.169	**0.404**	0.504
		0.5	0.167	**0.357**	0.391
optdigits	3823	0.1	0.208	**0.533**	0.703
		0.5	0.210	**0.460**	0.516
pageblock	5473	0.1	0.199	**0.495**	0.642
		0.5	0.208	**0.448**	0.497
pendigits	7494	0.1	0.209	**0.499**	0.629
		0.5	0.215	**0.457**	0.500
segment	2310	0.1	0.206	**0.558**	0.769
		0.5	0.206	**0.462**	0.532
spambase	4601	0.1	0.222	**0.553**	0.708
		0.5	0.225	**0.488**	0.539

Adaptation de domaine

Hypothèse

Les exemples sources et cibles sont générés par des distributions différentes.

$$\begin{array}{rcllll} S & = & \{\,(x_1, y_1), & (x_2, y_2), & \ldots, & (x_m, y_m)\,\} & \sim & (D_{\mathrm{S}})^m \\ T & = & \{\,(x_1, \cdot), & (x_2, \cdot), & \ldots, & (x_m, \cdot)\,\} & \sim & (D_{\mathrm{T}})^m \end{array}$$



Nouvelle borne pour l'adapation de domaine

$\mathcal{H}\Delta\mathcal{H}$-distance (Ben-David et al., 2006, 2010)

$$d_{\mathcal{H}\Delta\mathcal{H}}(D_{\mathrm{S}}, D_{\mathrm{T}}) \stackrel{\text{def}}{=} 2 \sup_{h,h' \in \mathcal{H}} \left| \mathop{\mathbf{E}}_{(x^{\mathrm{S}},\cdot)\sim D_{\mathrm{S}}} \mathrm{I}[h(x^{\mathrm{S}}) \neq h'(x^{\mathrm{S}})] - \mathop{\mathbf{E}}_{(x^{\mathrm{T}},\cdot)\sim D_{\mathrm{T}}} \mathrm{I}[h(x^{\mathrm{T}}) \neq h'(x^{\mathrm{T}})] \right|$$

Désaccord entre distributions

$$\mathrm{dis}_Q(D_{\mathrm{S}}, D_{\mathrm{T}}) \stackrel{\text{def}}{=} \left| d_Q^{D_{\mathrm{T}}} - d_Q^{D_{\mathrm{S}}} \right|$$

Théorème

[...] avec probabilité au moins $1-\delta$ sur le choix de $S \times T \sim (D_{\mathrm{S}} \times D_{\mathrm{T}})^m$, on a

$\forall\, Q$ sur $\mathcal{H}$:

$$R_{D_{\mathrm{T}}}(G_Q) \;\leq\; c' \cdot R_S(G_Q) + a' \cdot \mathrm{dis}_Q(S,T) + \left(\frac{c'}{c} + \frac{2\,a'}{a}\right) \frac{\mathrm{KL}(Q\|P) + \ln\frac{3}{\delta}}{m} + \lambda_Q^\star + a' - 1$$

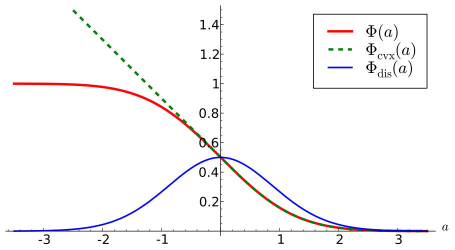
où $a' \stackrel{\text{def}}{=} \frac{2a}{1-e^{-2a}}$ et $c' \stackrel{\text{def}}{=} \frac{c}{1-e^{-c}}$.

Nouvel algorithme pour l'adaptation de domaine

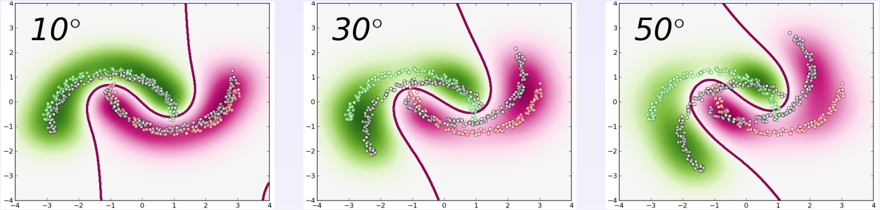
PBDA

Algorithme de minimisation de la borne pour classificateurs linéaires.

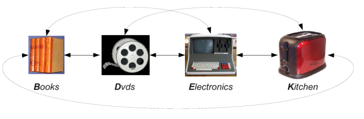
$$C\sum_{i=1}^{m}\Phi_{\rm cvx}\left(y_i^{\rm S}\frac{\mathbf{w}\cdot\mathbf{x}_i^{\rm S}}{\|\mathbf{x}_i^{\rm S}\|}\right)+A\left|\sum_{i=1}^{m}\Phi_{\rm dis}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^{\rm S}}{\|\mathbf{x}_i^{\rm S}\|}\right)-\Phi_{\rm dis}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^{\rm T}}{\|\mathbf{x}_i^{\rm T}\|}\right)\right|+\frac{\|\mathbf{w}\|^2}{2}$$



Résultats sur un problème jouet

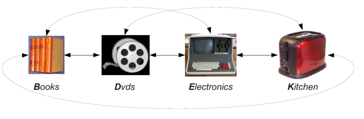


Résultats empiriques sur des données réelles



source → cible	PBGD	SVM	DASVM	CODA	PBDA
books→dvds	**0.174**	0.179	0.193	0.181	0.183
books→electronics	0.275	0.290	**0.226**	0.232	0.263
books→kitchen	0.236	0.251	**0.179**	0.215	0.229
dvds→books	**0.192**	0.203	0.202	0.217	0.197
dvds→electronics	0.256	0.269	**0.186**	0.214	0.241
dvds→kitchen	0.211	0.232	0.183	**0.181**	0.186
electronics→books	0.268	0.287	0.305	0.275	**0.232**
electronics→dvds	0.245	0.267	**0.214**	0.239	0.221
electronics→kitchen	**0.127**	0.129	0.149	0.134	0.141
kitchen→books	0.255	0.267	0.259	**0.247**	**0.247**
kitchen→dvds	0.244	0.253	**0.198**	0.238	0.233
kitchen→electronics	0.235	0.149	0.157	0.153	**0.129**
Moyenne	0.226	0.231	**0.204**	0.210	0.208

Résultats empiriques sur des données réelles



source → cible	PBGD	SVM	DASVM	CODA	PBDA	DALC
books→dvds	**0.174**	0.179	0.193	0.181	0.183	0.178
books→electronics	0.275	0.290	0.226	0.232	0.263	**0.212**
books→kitchen	0.236	0.251	**0.179**	0.215	0.229	0.194
dvds→books	0.192	0.203	0.202	0.217	0.197	**0.186**
dvds→electronics	0.256	0.269	**0.186**	0.214	0.241	0.245
dvds→kitchen	0.211	0.232	0.183	0.181	0.186	**0.175**
electronics→books	0.268	0.287	0.305	0.275	**0.232**	0.240
electronics→dvds	0.245	0.267	**0.214**	0.239	0.221	0.256
electronics→kitchen	0.127	0.129	0.149	0.134	0.141	**0.123**
kitchen→books	0.255	0.267	0.259	0.247	0.247	**0.236**
kitchen→dvds	0.244	0.253	**0.198**	0.238	0.233	0.225
kitchen→electronics	0.235	0.149	0.157	0.153	**0.129**	0.131
Moyenne	0.226	0.231	0.204	0.210	0.208	**0.200**

Plan

Résumé des contributions

Analyse PAC-bayésienne de trois cadres d'apprentissages :

1. Apprentissage inductif
   - Approche générale permettant de déduire plusieurs résultats existants.
   - Approche modulaire permettant d'adapter la théorie à d'autres cadres.
2. Apprentissage transductif
   - Amélioration substantielle de la borne existante.
3. Adaptation de domaine
   - Première borne PAC-bayésienne pour l'adaptation de domaine.
   - Algorithme d'apprentissage avec des assises théoriques.

Autres contributions :

- Théorie PAC-bayésienne pour votants dépendants des données
- Théorèmes PAC-bayésiens sans terme $\mathrm{KL}(Q\|P)$
- Réseaux de neurones adaptatif (apprentissage de représentation)

Perspectives

1. Apprentissage transductif
   - Conception de nouveaux algorithmes d'apprentissage
2. Adaptation de domaine
   - Améliorer le temps de calcul des algorithmes PBDA / DALC
   - Combiner avec d'autre approches d'adaptation de domaines (repondération des exemples source, apprentissage des représentations)
3. Étudier d'autres cadres d'apprentissage !

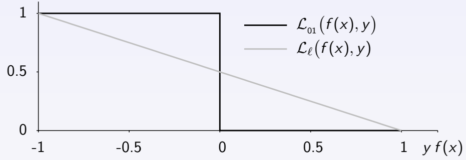
Plan

5 Annexe

Fonctions de pertes $\mathcal{L} : \overline{\mathcal{Y}} \times \mathcal{Y} \to [0, 1]$

Perte zéro-un : $\mathcal{L}_{01}\big(f(x), y\big) \stackrel{\text{def}}{=} \mathrm{I}\big[\, y\, f(x) \leq 0 \,\big]$,

Perte linéaire : $\mathcal{L}_{\ell}\big(f(x), y\big) \stackrel{\text{def}}{=} \frac{1}{2}\Big(1 - y\, f(x)\Big)$.

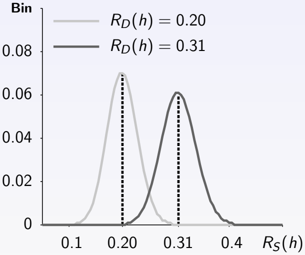


Risque d'un votant et loi binomiale

Probabilité d'observer k erreurs parmi m exemples

Pour un votant $h(\cdot)$ de risque $R_D(h)$, on considère une **variable binomiale** de m essais avec probabilité de succès $R_D(h)$:

$$\begin{aligned}\mathbf{Bin}\Big(k; m, R_D(h)\Big) &\stackrel{\text{def}}{=} \Pr_{S\sim D^m}\Big(R_S(h) = \tfrac{k}{m}\Big)\\ &= \binom{m}{k}\Big(R_D(h)\Big)^k\Big(1 - R_D(h)\Big)^{m-k}.\end{aligned}$$



Risque et loi hypergéométrique

Cadre inductif $\Rightarrow$ m piges avec remises selon D $\Rightarrow$ Loi binomiale.

Cadre transductif $\Rightarrow$ m piges sans remises dans Z $\Rightarrow$ Loi hypergéométrique.

Probabilité d'observer k erreurs parmi m exemples

Pour un votant h de risque $R_Z(h)$, on considère une **variable hypergéométrique** de m piges parmi une population de taille N contenant $N \cdot R_Z(h)$ succès.

$$\Pr_{S \sim [Z]^m}\left(R_S(h) = \tfrac{k}{m}\right) \quad = \quad \frac{\binom{N\cdot R_Z(h)}{k}\binom{N - N\cdot R_Z(h)}{m-k}}{\binom{N}{m}}\,,$$

pour tout $k \in \left\{ \max[0, N\cdot R_Z(h) + m - N], \ldots, \min[m, N\cdot R_Z(h)] \right\}$.

Marge du vote de majorité

Marge sur un exemple (x, y)

$$\color{red}{M_Q(x,y)} \overset{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \color{red}{y \cdot h(x)}\,.$$

Marge sur une distribution D

La **variable aléatoire** $\color{red}{M_Q^D}$ donne la marge sur exemple généré par D.

Risque de Bayes

$$\color{green}{R_D(B_Q)} = \Pr_{(x,y)\sim D}\Big(\color{red}{M_Q(x,y)} \leq 0\Big)$$

Risque de Gibbs

$$\color{green}{R_D(G_Q)} = \frac{1}{2}\Big(1 - \color{red}{\mu_1(M_Q^D)}\Big)$$

Espérance de désaccord

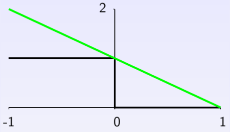
$$\color{green}{d_Q^D} \overset{\text{def}}{=} \mathop{\mathbf{E}}_{(x,\cdot)\sim D} \mathop{\mathbf{E}}_{h_1\sim Q} \mathop{\mathbf{E}}_{h_2\sim Q} \mathrm{I}\Big[h_1(x) \neq h_2(x)\Big] = \frac{1}{2}\Big(1 - \color{red}{\mu_2(M_Q^D)}\Big)$$

De la borne du facteur 2 à la $\mathcal{C}$-borne

En appliquant l'inégalité de Markov ($\Pr(X \geq a) \leq \frac{\mathbf{E}\, X}{a}$), on obtient :

Borne du facteur 2

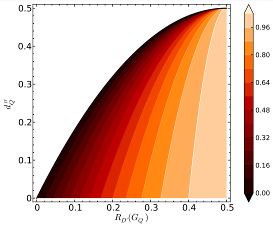
$$
\begin{aligned}
R_D(B_Q) &= \Pr_{(x,y)\sim D}\big(1 - M_Q(x,y) \geq 1\big) \\
&\leq \mathop{\mathbf{E}}_{(x,y)\sim D}\big(1 - M_Q(x,y)\big) \\
&= 1 - \mu_1(M_Q^D) \;=\; 2\, R_D(G_Q)\,.
\end{aligned}
$$



Par l'inégalité de Tchebychev ($\Pr(X - \mathbf{E}\, X \geq a) \leq \frac{\mathbf{Var}\, X}{a^2 + \mathbf{Var}\, X}$), on obtient :

La $\mathcal{C}$-borne (Lacasse et al., 2006)

$$
R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \;\stackrel{\mathrm{def}}{=}\; 1 - \frac{\big(1 - 2\cdot R_D(G_Q)\big)^2}{1 - 2\cdot d_Q^D}
$$



Borne sur le risque de Bayes

Borne du vote de majorité

Pour toute distribution D sur $\mathcal{X} \times \mathcal{Y}$, pour tout ensemble $\mathcal{H}$ de votants, pour toute distribution P sur $\mathcal{H}$, et pour tout $\delta \in (0,1]$, on a, avec probabilité au moins $1-\delta$ sur le choix de $S \sim D^m$,

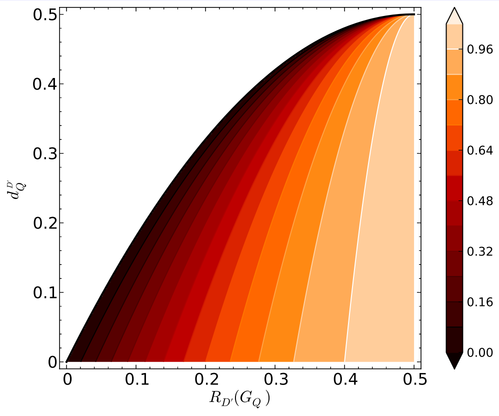
$$\forall\, Q \text{ sur } \mathcal{H}: \quad R_D(B_Q) \;\leq\; \mathcal{C}_Q^D \leq\; 1 - \frac{\left(1 - 2 \cdot \sup \mathcal{R}_{Q,S}^{\delta/2}\right)^2}{1 - 2 \cdot \inf \mathcal{D}_{Q,S}^{\delta/2}}\,,$$

où

$$\mathcal{R}_{Q,S}^{\delta/2} \;\stackrel{\text{def}}{=}\; \left\{ r \in [0, \tfrac{1}{2}] \;\middle|\; \mathrm{kl}\big(R_S(G_Q),\, r\big) \;\leq\; \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \tfrac{2\sqrt{m}}{\delta/2}\right] \right\},$$

$$\mathcal{D}_{Q,S}^{\delta/2} \;\stackrel{\text{def}}{=}\; \left\{ d \in [0, \tfrac{1}{2}] \;\middle|\; \mathrm{kl}(d_Q^S,\, d) \;\leq\; \frac{1}{m}\left[2\,\mathrm{KL}(Q\|P) + \ln \tfrac{2\sqrt{m}}{\delta/2}\right] \right\}.$$

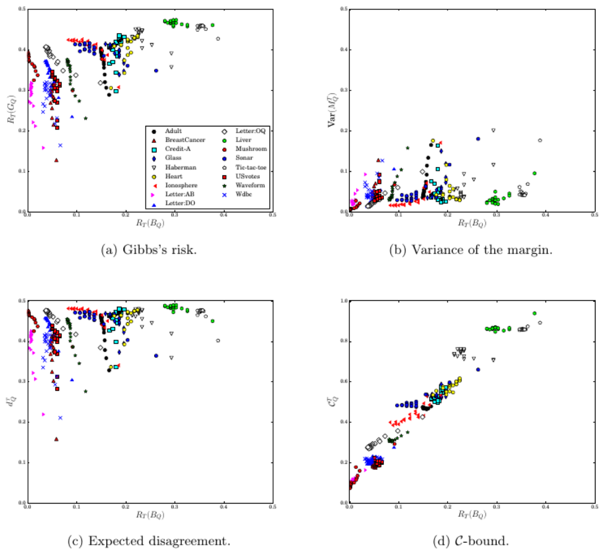
Comportement de la $\mathcal{C}$-borne



Proposition

$$R_{D'}(G_Q) \leq d_Q^{D'} \iff \mathcal{C}_Q^{D'} \leq 2R_{D'}(G_Q)$$

Étude empirique de la $\mathcal{C}$-borne



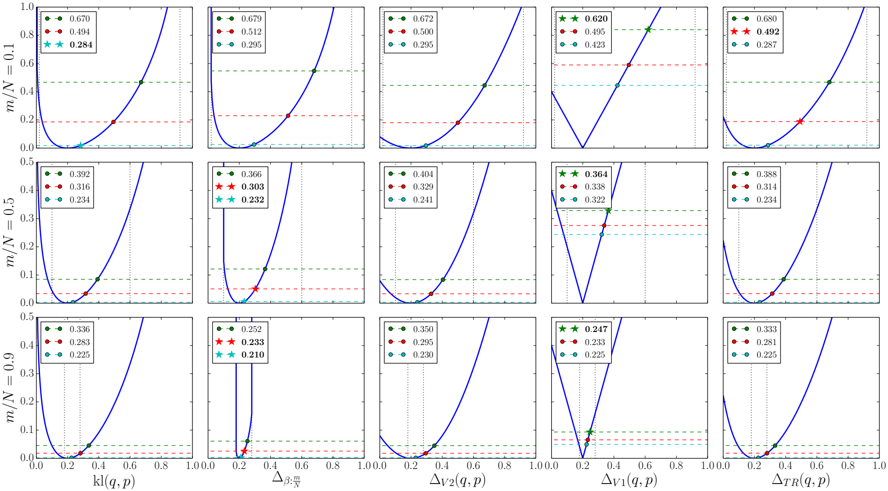
(a) Gibbs's risk.

(b) Variance of the margin.

(c) Expected disagreement.

(d) $\mathcal{C}$-bound.

Comparaisons de plusieurs Δ-fonctions



Conception d'une Δ-fonction pour le cas transductif

$$\Delta_\beta(q,p) \stackrel{\text{def}}{=} \frac{H(\beta) - pH(\beta\frac{q}{p}) - (1-p)H(\beta\frac{1-q}{1-p})}{\beta}\,.$$

avec $H(q) \stackrel{\text{def}}{=} -q\ln q - (1-q)\ln(1-q)$

Fixons $\beta := \frac{m}{N}$

$$\mathcal{T}_{\beta:\frac{m}{N}}(m,N) = \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{m,N,K}} \frac{\binom{K}{k}\binom{N-K}{m-k}}{\binom{N}{m}} e^{NH(\frac{m}{N}) - KH(\frac{k}{K}) - (N-K)H(\frac{m-k}{N-K})}\right]$$

$$= \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{m,N,K}} \frac{\alpha(k,K)\,\alpha(m-k,N-K)}{\alpha(m,N)}\right]$$

où $\alpha(a,b) \stackrel{\text{def}}{=} \binom{b}{a}\left(\frac{a}{b}\right)^a\left(1-\frac{a}{b}\right)^{b-a}$

Bornes sur le risque de Bayes (transductif)

Borne du vote de majorité

Pour tout échantillon de données Z contenant $N \geq 42$ exemples, pour tout ensemble $\mathcal{H}$ de votants, pour toute distribution P sur $\mathcal{H}$, et pour tout $\delta \in (0,1]$, on a, avec probabilité au moins $1-\delta$ sur le choix S de m exemples parmi Z,

$\forall\, Q$ *sur* $\mathcal{H}$:

(a) $R_Z(B_Q) \;\leq\; 2 \cdot \sup \mathcal{R}_{Q,S}^{\delta,\beta}$ *(Facteur 2)*

(b) $R_Z(B_Q) \;\leq\; 1 - \dfrac{\left(1 - 2 \cdot \sup \mathcal{R}_{Q,S}^{\delta,\beta}\right)^2}{1 - 2 \cdot d_Q^Z}$ *($\mathcal{C}$-borne)*

où

$$\mathcal{R}_{Q,S}^{\delta,\beta} \;\stackrel{\text{def}}{=}\; \left\{ r \in [0, \tfrac{1}{2}] \;\middle|\; \Delta_{\beta:\frac{m}{N}}\big(R_S(G_Q),\, r\big) \;\leq\; \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{3\ln(m)\sqrt{m(1-\frac{m}{N})}}{\delta}\right] \right\},$$

$$d_Q^Z \;=\; \frac{1}{2}\left(1 - \sum_{i=1}^{N}\left[\mathop{\mathbf{E}}_{h\sim Q} h(x_i)\right]^2\right).$$

Bornes sur le risque de Bayes (transductif)

Nom	N	m/N	$R_S(B_Q)$	Facteur 2	$\mathcal{C}$-borne
car	1728	0.1	0.105	1.092	-
car	1728	0.5	0.115	0.830	**0.819**
letter_AB	1555	0.1	0.000	**0.914**	0.961
letter_AB	1555	0.5	0.000	0.797	**0.626**
mushroom	8124	0.1	0.000	**0.964**	0.966
mushroom	8124	0.5	0.000	0.875	**0.546**
nursery	12959	0.1	0.009	0.798	**0.692**
nursery	12959	0.5	0.010	0.711	**0.379**
optdigits	3823	0.1	0.000	1.055	-
optdigits	3823	0.5	0.026	0.917	**0.793**
pageblock	5473	0.1	0.048	**0.979**	0.992
pageblock	5473	0.5	0.057	0.894	**0.697**
pendigits	7494	0.1	0.023	**0.989**	0.997
pendigits	7494	0.5	0.041	0.912	**0.706**
segment	2310	0.1	0.000	1.101	-
segment	2310	0.5	0.014	0.920	**0.834**
spambase	4601	0.1	0.115	1.096	-
spambase	4601	0.5	0.137	0.973	**0.961**

Nouvelle approche pour l'adaptation de domaine

Théorème

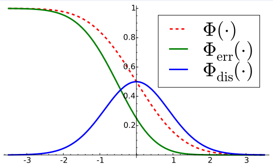
Pour toute paire de distributions D_{S} et D_{T} sur $\mathcal{X} \times \mathcal{Y}$, pour tout ensemble $\mathcal{H}$ de votants $\mathcal{X} \to [-1,1]$, pour tout nombre réel $q > 0$,

$$\forall Q \text{ sur } \mathcal{H}, \quad R_{D_{\mathrm{T}}}(G_Q) \;\leq\; \frac{1}{2}\, d_Q^{D_{\mathrm{T}}} + \beta_q(D_{\mathrm{T}} \| D_{\mathrm{S}}) \times \left[e_Q^{D_{\mathrm{S}}} \right]^{1-\frac{1}{q}} .$$

où $\beta_q(D_{\mathrm{T}} \| D_{\mathrm{S}}) \;=\; \left[\mathop{\mathbf{E}}_{(x,y)\sim D_{\mathrm{S}}} \left(\frac{D_{\mathrm{T}}(x,y)}{D_{\mathrm{S}}(x,y)} \right)^q \right]^{\frac{1}{q}} .$

DALC

$$C \sum_{i=1}^{m_s} \Phi_{\mathrm{err}} \left(y_i^{\mathrm{S}} \frac{\mathbf{w} \cdot \mathbf{x}_i^{\mathrm{S}}}{\|\mathbf{x}_i^{\mathrm{S}}\|} \right) + A \sum_{i=1}^{m_t} \Phi_{\mathrm{dis}} \left(\frac{\mathbf{w} \cdot \mathbf{x}_i^{\mathrm{T}}}{\|\mathbf{x}_i^{\mathrm{T}}\|} \right) + \frac{\|\mathbf{w}\|^2}{2}$$



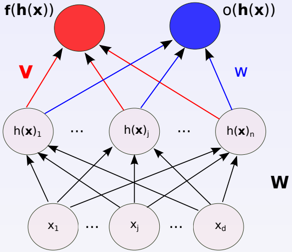
Domain-Adversarial Neural Network (DANN)

$$\min_{\mathbf{W},\mathbf{V},\mathbf{b},\mathbf{c}} \left[\underbrace{\frac{1}{m}\sum_{i=1}^{m} -\log\left(f_{y_i^s}(\mathbf{x}_i^s)\right)}_{\text{source loss}} + \underbrace{\lambda \max_{\mathbf{w},d} \left(\frac{1}{m}\sum_{i=1}^{m}\log\left(o(\mathbf{h}(\mathbf{x}_i^s))\right) + \frac{1}{m}\sum_{i=1}^{m}\log\left(1-o(\mathbf{h}(\mathbf{x}_i^t))\right)\right)}_{\text{adaptation regularizer}} \right],$$

where $\lambda > 0$ weights the domain adaptation regularization term.

Given a source sample $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,
and a target sample $T = \{(\mathbf{x}_i^t)\}_{i=1}^m \sim (\mathcal{D}_T)^m$,

1. Pick a $\mathbf{x}^s \in S$ and $\mathbf{x}^t \in T$
2. Update $\mathbf{V}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
3. Update $\mathbf{W}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
4. Update $\mathbf{w}$ towards $o(\mathbf{h}(\mathbf{x}^s)) = 1$ and $o(\mathbf{h}(\mathbf{x}^t)) = -1$
5. Update $\mathbf{W}$ towards $o(\mathbf{h}(\mathbf{x}^s)) = -1$ and $o(\mathbf{h}(\mathbf{x}^t)) = 1$



**DANN finds a representation $\mathbf{h}(\cdot)$ that are good on S ;
but unable to discriminate between S and T.**