PAC-Bayesian Theory
and Domain Adaptation Algorithms

Pascal Germain

November 25, 2015

Plan

Plan

1. Basic Definitions
2. PAC-Bayesian Theory
   - Majority Vote Classifiers
   - A Classical PAC-Bayesian Theorem
   - A General PAC-Bayesian Theorem
   - Transductive Learning
   - Rényi-Based Theorem
3. Domain Adaptation Algorithms
   - Ben-David et al.'s Domain Divergence
   - A First PAC-Bayesian Algorithm
   - A Second PAC-Bayesian Algorithm
   - A Neural Network / Representation Learning Algorithm
4. Conclusion and future works

Definitions

Learning example

An example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a **description-label** pair.

Data generating distribution

Each example is an **observation from distribution** D on $\mathcal{X} \times \mathcal{Y}$.

Learning sample

$$S \stackrel{\text{def}}{=} \{\, (x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)\, \} \ \sim\ D^n$$

Classifier (or hypothesis)

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

Binary classifier

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

Learning algorithm

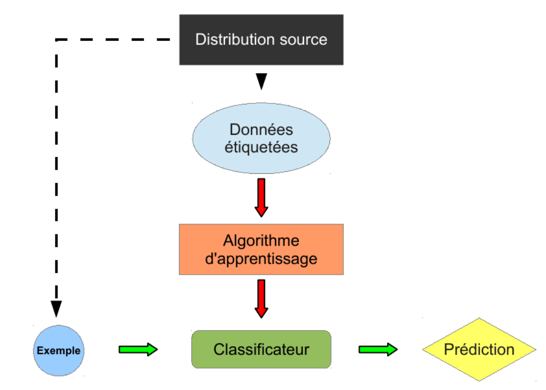
$$A(S) \longrightarrow h$$

I.I.D. Assumption

Assumption

Examples are generated *i.i.d.* by a distribution D on $\mathcal{X} \times \mathcal{Y}$.

$$S \;=\; \{\, (x_1, y_1),\; (x_2, y_2),\; \ldots,\; (x_n, y_n) \,\} \;\sim\; D^n$$



Risks

Risk (or generalization error)

Probability of misclassifying an example generated by distribution D :

$$R_D(h) \stackrel{\text{def}}{=} \Pr_{(x,y)\sim D}\big(h(x) \neq y\big)$$
$$= \mathop{\mathbf{E}}_{(x,y)\sim D} \mathrm{I}\big[y \cdot h(x) \leq 0\big], \quad \langle \textit{ binary classification } \rangle$$

where $\mathrm{I}[a] = 1$ if predicate a is *true*; $\mathrm{I}[a] = 0$ otherwise.

Empirical risk

Error rate on the learning sample $S \sim D^n$:

$$\widehat{R}_S(h) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\big[y_i \cdot h(x_i) \leq 0\big].$$

Plan

PAC-Bayesian Theory

Initiated by David McAllester (1999), the PAC-Bayesian theory gives generalization guarantees on **majority votes** of classifiers.

PAC guarantees (Probably Approximately Correct)

With probability at least «$1-\delta$», the risk of classifier h is less than «ϵ»

$$\Pr_{S\sim D^n}\left(R_D(h) \leq \epsilon(R_S(h), n, \ldots) \right) \geq 1-\delta$$

Bayesian flavor

Incorporates *a priori* knowledge about the learning problem as a probability distribution over a family of classifiers.

Training bounds

- Gives generalization guarantees **not based on testing sample**;
- Inspiration for conceiving **new learning algorithms**.

Plan

Majority Vote Classifiers

Given :

- A set of **voters** $\mathcal{H} = \{h_1, h_2, h_3, \ldots\}$ (discrete or continuous) ;
- A **weight** distribution Q on $\mathcal{H}$.

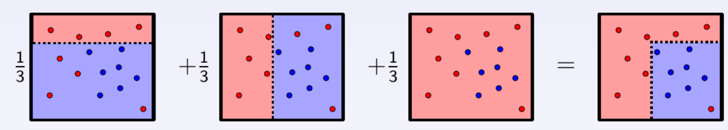
Weighted majority vote

To predict the label of $x \in \mathcal{X}$, the classifier asks for the *prevailing opinion*

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

Many learning algorithms output majority vote classifiers

AdaBoost, Random Forests, Bagging, ...



A Surrogate Loss

Given

- A data distribution D on $\mathcal{X} \times \{-1,+1\}$;
- A weight distribution Q on the set of voters $\mathcal{H}$.

Majority vote risk (or *Bayes Risk*)

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

Gibbs Risk

The stochastic Gibbs classifier $G_Q(x)$ draws $h' \in \mathcal{H}$ according to Q and output $h'(x)$.

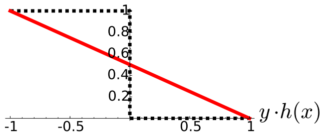
$$R_D(G_Q) \overset{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} \mathop{\mathbf{E}}_{h\sim Q} \mathrm{I}\left[y\cdot h(x) \leq 0 \right]$$

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

Factor two

It is well-known that

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

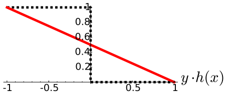


From the *Factor 2* to the $\mathcal{C}$-bound

From Markov's inequality ($\Pr(X \geq a) \leq \frac{\mathbf{E}\, X}{a}$), we obtain :

Factor 2 bound

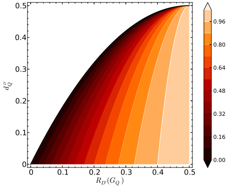
$$\begin{aligned} R_D(B_Q) &= \Pr_{(x,y)\sim D}\big(1 - y\cdot h(x) \geq 1\big) \\ &\leq \mathop{\mathbf{E}}_{(x,y)\sim D}\big(1 - y\cdot h(x)\big) \;=\; 2\, R_D(G_Q)\,. \end{aligned}$$



From Chebyshev's inequality ($\Pr(X - \mathbf{E}\, X \geq a) \leq \frac{\mathbf{Var}\, X}{a^2 + \mathbf{Var}\, X}$), we obtain :

The $\mathcal{C}$-bound (Lacasse et al., 2006)

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



where d_Q^D is the **expected disagreement** :

$$d_Q^D \;\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\; \mathop{\mathbf{E}}_{h_1\sim Q}\; \mathop{\mathbf{E}}_{h_2\sim Q} \mathrm{I}\Big[h_1(x) \neq h_2(x)\Big] \;=\; \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(x,\cdot)\sim D'}\left[\mathop{\mathbf{E}}_{h\sim Q} h(x)\right]^2\right)$$

Plan

A *Classical* PAC-Bayesian Theorem

Two principal components

- The **Gibbs empirical risk** :

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

- The **Kullback-Leibler divergence** between the *prior* P and the *posterior* Q :

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

PAC-Bayesian theorem (*McAllester, 2003*)

For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set of voters $\mathcal{H}$, for any distribution P on $\mathcal{H}$, for any $\delta\in(0,1]$, we have, with probability at least $1-\delta$ over the choice of $S\sim D^n$,

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

Plan

A General PAC-Bayesian Theorem

Δ-function : «distance» between $\widehat{R}_S(G_Q)$ et $R_D(G_Q)$

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

General theorem

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of voters, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any Δ-function, we have, with probability at least $1-\delta$ over the choice of $S \sim D^n$,

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

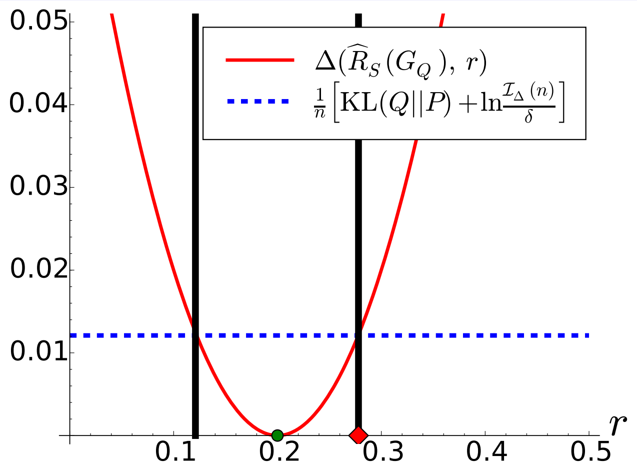
où

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

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Interpretation.



General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Proof ideas.

Change of Measure Inequality

For any P and Q on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have

$$\mathop{\mathbf{E}}_{h\sim Q}\phi(h) \leq \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} e^{\phi(h)}\right).$$

Markov's inequality

$$\Pr\left(X \geq a\right) \leq \frac{\mathbf{E}\,X}{a} \quad\Longleftrightarrow\quad \Pr\left(X \leq \frac{\mathbf{E}\,X}{\delta}\right) \geq 1-\delta\,.$$

Probability of observing k misclassifications among n examples

Given a voter h, consider a **binomial variable** of n trials with **success** $R_D(h)$:

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

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Proof.

$$
\begin{array}{lcl}
 & & n\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{R}_S(h),\ \mathop{\mathbf{E}}_{h\sim Q} R_D(h)\Big)\\
\textbf{Jensen's Inequality} & \leq & \mathop{\mathbf{E}}_{h\sim Q} n\cdot\Delta\Big(\widehat{R}_S(h), R_D(h)\Big)\\
\textbf{Change of measure} & \leq & \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{n\Delta\left(\widehat{R}_S(h), R_D(h)\right)}\\
\textbf{Markov's Inequality} & \leq_{1-\delta} & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim D^n} \mathop{\mathbf{E}}_{h\sim P} e^{n\cdot\Delta(R_{S'}(h), R_D(h))}\\
\textbf{Expectation swap} & = & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^n} e^{n\cdot\Delta(R_{S'}(h), R_D(h))}\\
\textbf{Binomial law} & = & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{n} \mathbf{Bin}\Big(k; n, R_D(h)\Big)\, e^{n\cdot\Delta(\frac{k}{n}, R_D(h))}\\
\textbf{Supremum over risk} & \leq & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \sup_{r\in[0,1]} \left[\sum_{k=0}^{n} \mathbf{Bin}\Big(k; n, r\Big)\, e^{n\Delta(\frac{k}{n}, r)}\right]\\
 & = & \mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\, \mathcal{I}_\Delta(n)\,.
\end{array}
$$

□

General theorem

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(n)}{\delta}\right]\right) \geq 1-\delta\,.$$

Corollary

[...] with probability at least $1-\delta$ *over the choice of* $S \sim D^n$,

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

(a) $\mathrm{kl}\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]$, *(Langford and Seeger, 2001)*

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

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

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

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

Bounding the Expected Disagreement

Expected disagreement

$$d_Q^D \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,\cdot)\sim D} \mathop{\mathbf{E}}_{h_1\sim Q} \mathop{\mathbf{E}}_{h_2\sim Q} \mathrm{I}\Big[h_1(x) \neq h_2(x)\Big] = \tfrac{1}{2}\bigg(1 - \mathop{\mathbf{E}}_{(x,\cdot)\sim D'} \Big[\mathop{\mathbf{E}}_{h\sim Q} h(x)\Big]^2\bigg)$$

General theorem

[...] with probability at least $1{-}\delta$ *over the choice of* $S \sim D^n$,

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

Corollary

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

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

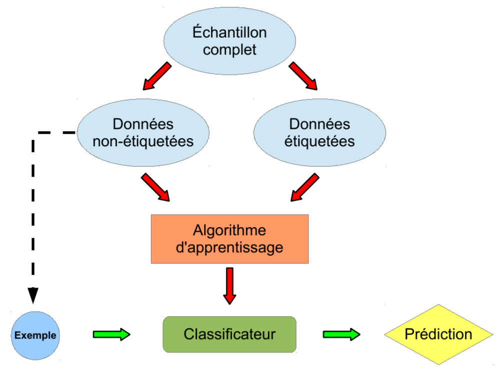
(c) $d_Q^D \leq \frac{1}{1-e^{-c}}\Big(c \cdot d_Q^S + \frac{1}{n}\big[2\,\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta}\big]\Big)$.

Plan

Assumption

Examples are drawn *without replacement* from a finite set Z of size N.

$$\begin{array}{rcllll} S & = & \{\ (x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n)\ \} & \subset Z \\ U & = & \{\ (x_{n+1}, \cdot), & (x_{n+2}, \cdot), & \ldots, & (x_N, \cdot)\ \} & = Z \setminus S \end{array}$$



General Theorem for Transductive Learning

Observation

Inductive learning : n draws with replacement according to $D \Rightarrow$ Binomial law.

Transductive learning : n draws without replacement in $Z \Rightarrow$ Hypergeometric law.

Theorem

For any set Z of N examples, for any set $\mathcal{H}$ of voters, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any Δ-function, we have, with probability at least $1-\delta$ over the choice of n examples among Z,

$$\forall\, Q \text{ on } \mathcal{H} : \quad \Delta\Big(\widehat{R}_S(G_Q), \widehat{R}_Z(G_Q)\Big) \leq \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{T}_\Delta(n,N)}{\delta}\right],$$

where

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

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

Theorem

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

Poof.

$$n\cdot\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}\widehat{R}_S(h), \mathop{\mathbf{E}}_{h\sim Q}\widehat{R}_Z(h)\Big)$$

Jensen's inequality $$\le \mathop{\mathbf{E}}_{h\sim Q} n\cdot\Delta\Big(\widehat{R}_S(h), \widehat{R}_Z(h)\Big)$$

Change of measure $$\le \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{n\Delta\left(\widehat{R}_S(h),\widehat{R}_Z(h)\right)}$$

Markov's inequality $$\le_{1-\delta} \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim[Z]^n}\mathop{\mathbf{E}}_{h\sim P} e^{n\cdot\Delta(R_{S'}(h),\widehat{R}_Z(h))}$$

Expectations swap $$= \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim[Z]^n} e^{n\cdot\Delta(R_{S'}(h),\widehat{R}_Z(h))}$$

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

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

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

□

A New Transductive Bound for the Gibbs Risk

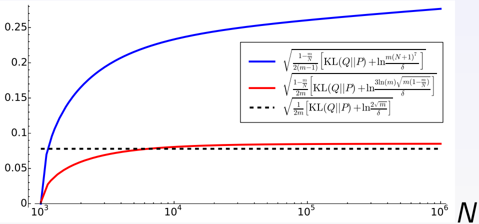
Corollary

[...] with probability at least $1-\delta$ *over the choice of* n *examples among* Z,

$$\forall\, Q \text{ on } \mathcal{H} : \widehat{R}_Z(G_Q) \;\le\; \widehat{R}_S(G_Q) + \sqrt{\frac{1-\frac{n}{N}}{2n}\left[\mathrm{KL}(Q\|P) + \ln \frac{3\ln(n)\sqrt{n(1-\frac{n}{N})}}{\delta}\right]}\,.$$

Theorem (Derbeko et al., 2004)

$$\widehat{R}_Z(G_Q) \;\le\; \widehat{R}_S(G_Q) + \sqrt{\frac{1-\frac{n}{N}}{2(n-1)}\left[\mathrm{KL}(Q\|P) + \ln \frac{n(N+1)^7}{\delta}\right]}\,.$$



A New Transductive Bound for the Bayes Risk

Majority Vote Bound

[...] with probability at least $1-\delta$ *over the choice of* n *examples among* Z*,*

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

$$\text{(a)} \quad \widehat{R}_Z(B_Q) \;\leq\; 2 \times \overline{r} \qquad \textit{(Factor 2)}$$

$$\text{(b)} \quad \widehat{R}_Z(B_Q) \;\leq\; 1 - \frac{\left(1 - 2 \times \overline{r}\right)^2}{1 - 2 \times d_Q^Z} \qquad \textit{($\mathcal{C}$-bound)}$$

where

$$\overline{r} \;:=\; \widehat{R}_S(G_Q) + \sqrt{\frac{1-\frac{n}{N}}{2n}\left[\mathrm{KL}(Q\|P) + \ln \frac{3\ln(n)\sqrt{n(1-\frac{n}{N})}}{\delta}\right]}\,,$$

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

Empirical Comparison

Majority votes of *decision stumps* obtained with *AdaBoost*.

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

Plan

A New Change of Measure

Kullback-Leibler Change of Measure Inequality

For any P and Q on $\mathcal{H}$, and for any $\phi : \mathcal{H} \to \mathbb{R}$, we have

$$\mathop{\mathbf{E}}_{h\sim Q} \phi(h) \ \le\ \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} e^{\phi(h)}\right).$$

Rényi Change of Measure Inequality

For any P and Q on $\mathcal{H}$, any $\phi : \mathcal{H} \to \mathbb{R}$, and for any $\alpha > 1$, we have

$$\frac{\alpha}{\alpha-1} \ln \mathop{\mathbf{E}}_{h\sim Q} \phi(h) \ \le\ D_\alpha(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} \phi(h)^{\frac{\alpha}{\alpha-1}}\right),$$

with $\quad D_\alpha(Q\|P) \ \stackrel{\text{def}}{=}\ \dfrac{1}{\alpha-1} \ln \left[\mathop{\mathbf{E}}_{h\sim P}\left(\dfrac{Q(h)}{P(h)}\right)^{\alpha}\right].$

Rényi-Based General Theorem

General theorem

[...] for any $\alpha > 1$*, with probability at least* $1-\delta$ *over the choice of* $S \sim D^n$*,*

$$\forall Q \text{ on } \mathcal{H}\colon \quad \ln \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{\mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha')}{\delta}\right],$$

with

$$\mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha') \stackrel{\text{def}}{=} \sup_{r\in[0,1]} \left[\sum_{k=0}^{n} \mathbf{Bin}\big(k;n,r\big)\Delta\big(\tfrac{k}{n},\, r\big)^{\alpha'}\right],$$

and $\alpha' := \frac{\alpha}{\alpha-1} > 1\,.$

General theorem (*Rényi-Based*)

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ on } \mathcal{H}:\ \ln \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \le \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{\mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha')}{\delta}\right]\right) \ \ge\ 1-\delta\,.$$

Proof. $\alpha' := \frac{\alpha}{\alpha-1}$

$$\alpha' \cdot \ln \Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} \widehat{R}_S(h), \mathop{\mathbf{E}}_{h\sim Q} R_D(h)\Big)$$

Jensen's Inequality $$\le \quad \alpha' \cdot \ln \mathop{\mathbf{E}}_{h\sim Q} \Delta\Big(\widehat{R}_S(h), R_D(h)\Big)$$

Change of measure $$\le \quad D_\alpha(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} \Delta\big(\widehat{R}_S(h), R_D(h)\big)^{\alpha'}$$

Markov's Inequality $$\le_{1-\delta} \quad D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim D^n} \mathop{\mathbf{E}}_{h\sim P} \Delta(R_{S'}(h), R_D(h))^{\alpha'}$$

Expectation swap $$= \quad D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^n} \Delta(R_{S'}(h), R_D(h))^{\alpha'}$$

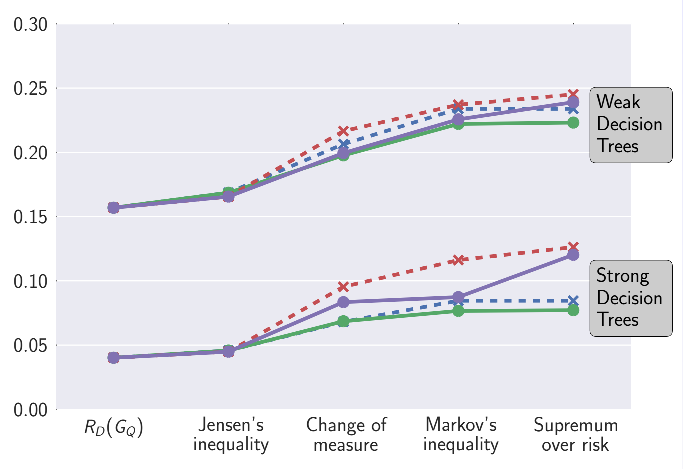
Binomial law $$= \quad D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{n} \mathbf{Bin}\big(k; n, R_D(h)\big)\Delta\big(\tfrac{k}{n}, R_D(h)\big)^{\alpha'}$$

Supremum over risk $$\le \quad D_\alpha(Q\|P) + \ln \frac{1}{\delta} \sup_{r\in[0,1]} \left[\sum_{k=0}^{n} \mathbf{Bin}\big(k; n, r\big)\Delta\big(\tfrac{k}{n}, r\big)^{\alpha'}\right]$$

$$, \quad = \quad D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathcal{I}^{\mathrm{R}}_{\Delta}(n,\alpha')\,. \qquad \square$$

Empirical Study

Majority votes of 500 decision trees on *Mushroom* dataset



× $\mathrm{KL}(Q\|P)$ and $\Delta := 2(q-p)^2$ • $D_\alpha(Q\|P)$ and $\Delta := 2(q-p)^2$

× $\mathrm{KL}(Q\|P)$ and $\Delta := \mathrm{kl}(q,p)$ • $D_\alpha(Q\|P)$ and $\Delta := \mathrm{kl}(q,p)$

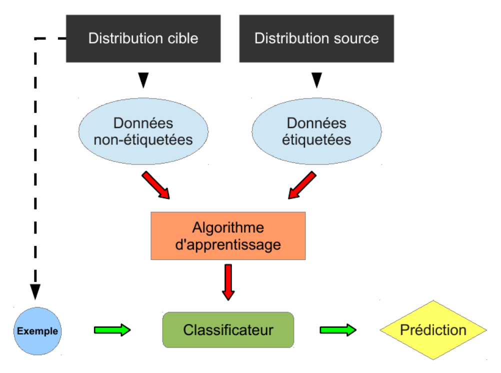
Plan

Domain Adaptation

Assumption

Source and target examples are generated by different distributions

$$
\begin{array}{rclllll}
S & = & \{\, (x_1, y_1), & (x_2, y_2), & \ldots, & (x_n, y_n) \,\} & \sim (D_{\mathrm{S}})^n \\
T & = & \{\, (x_1, \cdot), & (x_2, \cdot), & \ldots, & (x_n, \cdot) \,\} & \sim (D_{\mathrm{T}})^n
\end{array}
$$



Our Domain Adaptation Setting

Binary classification tasks

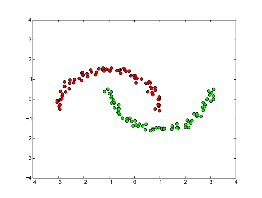
- Input space : $\mathbb{R}^d$
- Labels : $\{-1, 1\}$

Two different data distributions

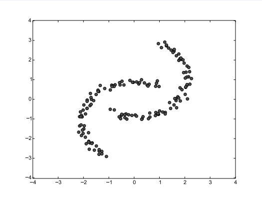
- Source domain : D_S
- Target domain : D_T

A **domain adaptation** learning algorithm is provided with

a **labeled source sample**
$S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^n \sim (D_S)^n$,



an **unlabeled target sample**
$T = \{\mathbf{x}_i^t\}_{i=1}^n \sim (D_T)^n$.



The goal is to build a classifier $h : \mathbb{R}^d \to \{-1, 1\}$ with a low **target risk**

$$R_{D_T}(h) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x}^t, y^t) \sim D_T} \left[h(\mathbf{x}^t) \neq y^t \right].$$

Plan

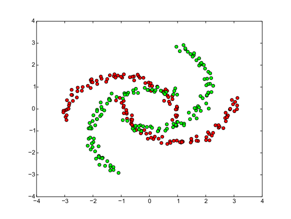
Divergence between source and target domains

Definition (Ben David et al., 2006)

Given two domain distributions D_S and D_T, and a **hypothesis class** $\mathcal{H}$, the $\mathcal{H}$-**divergence** between D_S and D_T is

$$d_{\mathcal{H}}(D_S, D_T) \stackrel{\text{def}}{=} 2 \sup_{h \in \mathcal{H}} \left| \Pr_{\mathbf{x}^s \sim D_S} \left[h(\mathbf{x}^s) = 1 \right] + \Pr_{\mathbf{x}^t \sim D_T} \left[h(\mathbf{x}^t) = -1 \right] - 1 \right| .$$

The $\mathcal{H}$-**divergence** measures the ability of an hypothesis class $\mathcal{H}$ to **discriminate** between source D_S and target D_T distributions.



Bound on the target risk

Theorem (Ben David et al., 2006)

Let $\mathcal{H}$ be a hypothesis class of VC-dimension d. With probability $1-\delta$ over the choice of samples $S \sim (D_S)^n$ and $T \sim (D_T)^n$, for every $h \in \mathcal{H}$:

$$R_{D_T}(h) \;\leq\; \widehat{R}_S(h) + \tfrac{4}{n}\sqrt{d \log \tfrac{2e\,n}{d} + \log \tfrac{4}{\delta}} + \hat{d}_{\mathcal{H}}(S,T) + \tfrac{4}{n^2}\sqrt{d \log \tfrac{2\,n}{d} + \log \tfrac{4}{\delta}} + \beta$$

with $\beta \geq \inf\limits_{h^* \in \mathcal{H}} \left[R_{D_S}(h^*) + R_{D_T}(h^*)\right].$

Empirical risk on the source sample :

$$\widehat{R}_S(h) \;\stackrel{\text{def}}{=}\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}[h(\mathbf{x}_i^s) \neq y_i^s]\,.$$

Empirical $\mathcal{H}$-divergence :

$$\hat{d}_{\mathcal{H}}(S,T) \;\stackrel{\text{def}}{=}\; 2\max_{h\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}[h(\mathbf{x}_i^s)=1] + \frac{1}{n}\sum_{i=1}^{n}\mathrm{I}[h(\mathbf{x}_i^t)=-1] - 1\right].$$

Plan

Nouvelle borne pour l'adapation de domaine

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

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

Distributions disagreement

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

Theorem

[...] with probability $1-\delta$ *over the choice of* $S \times T \sim (D_{\rm S} \times D_{\rm T})^n$*, we have*

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

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

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

A New Domain Adaptation Algorithm

For linear classifiers

PAC-Bayes specialization to linear classifier (Langford and Shawe-Taylor, 2002)

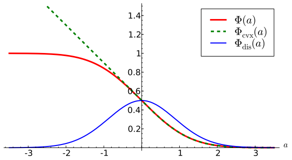
- Linear classifier : $h_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}[\mathbf{w}\cdot\mathbf{x}]$
- Voters : $\mathcal{H} = \{h_{\mathbf{v}} \mid \mathbf{v} \in \mathbb{R}^d\}$
- Prior $P_{\mathbf{0}}$: isotropic Gaussian centered on $\mathbf{0}$
- Posterior $Q_{\mathbf{w}}$: isotropic Gaussian centered on $\mathbf{w}$

- $h_{\mathbf{w}}(\mathbf{x}) = B_{Q_{\mathbf{w}}}(\mathbf{x})$
- $R_D(Q_{\mathbf{w}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \Phi\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right)$
- $\mathrm{KL}(Q_{\mathbf{w}}\|P_{\mathbf{0}}) = \frac{1}{2}\|\mathbf{w}\|^2$

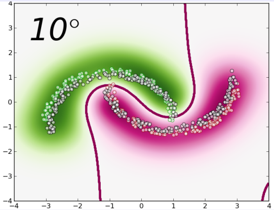
PBDA

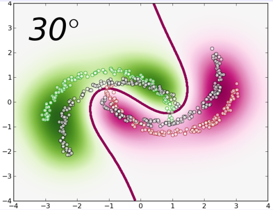
Minimize :

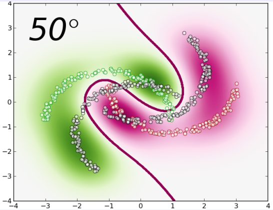
$$C\sum_{i=1}^{n} \Phi_{\mathrm{c}}\left(y_i^{\mathrm{S}}\,\frac{\mathbf{w}\cdot\mathbf{x}_i^{\mathrm{S}}}{\|\mathbf{x}_i^{\mathrm{S}}\|}\right) + A\left|\sum_{i=1}^{n} \Phi_{\mathrm{d}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^{\mathrm{S}}}{\|\mathbf{x}_i^{\mathrm{S}}\|}\right) - \Phi_{\mathrm{d}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^{\mathrm{T}}}{\|\mathbf{x}_i^{\mathrm{T}}\|}\right)\right| + \frac{\|\mathbf{w}\|^2}{2}$$



Intertwining Moons Toy Dataset

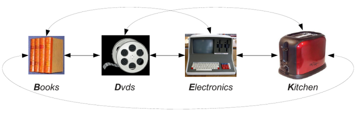






Amazon Reviews Dataset

Input : product review (bag of words) — **Output :** positive or negative rating.



source → target	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**
Mean	0.226	0.231	**0.204**	0.210	0.208

Plan

A New Perspective on Domain Adaptation

Theorem

For all pairs D_{S} and D_{T} on $\mathcal{X}\times\mathcal{Y}$, for all set $\mathcal{H}$ of voters, and for all $q>0$,

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

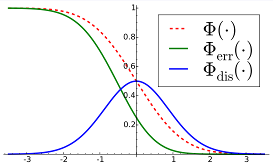
where $\beta_q(D_{\mathrm{T}}\|D_{\mathrm{S}}) \;=\; \left[\mathop{\mathbf{E}}_{(x,y)\sim D_{\mathrm{S}}}\left(\frac{D_{\mathrm{T}}(x,y)}{D_{\mathrm{S}}(x,y)}\right)^q\right]^{\frac{1}{q}}$,

and $e_Q^D \stackrel{\mathrm{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D}\ \mathop{\mathbf{E}}_{h_1\sim Q}\ \mathop{\mathbf{E}}_{h_2\sim Q} \mathrm{I}\big[h_1(x)\neq y\big]\times \mathrm{I}\big[h_2(x)\neq y\big]$.

DALC

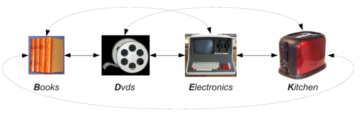
Minimize :

$$A\sum_{i=1}^{n_t}\Phi_{\mathrm{d}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^{\mathrm{T}}}{\|\mathbf{x}_i^{\mathrm{T}}\|}\right) + B\sum_{i=1}^{n_s}\Phi_{\mathrm{err}}\left(y_i^{\mathrm{S}}\frac{\mathbf{w}\cdot\mathbf{x}_i^{\mathrm{S}}}{\|\mathbf{x}_i^{\mathrm{S}}\|}\right) + \frac{\|\mathbf{w}\|^2}{2}$$



Amazon Reviews Dataset

Input : product review (bag of words) — **Output :** positive or negative rating.



source → target	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
Mean	0.226	0.231	0.204	0.210	0.208	**0.200**

Plan

Bound on the target risk

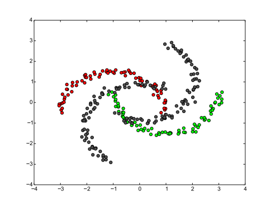
Theorem (Ben David et al., 2006)

Let $\mathcal{H}$ be a hypothesis class of VC-dimension d. With probability $1-\delta$ over the choice of samples $S \sim (D_S)^n$ and $T \sim (D_T)^n$, for every $h \in \mathcal{H}$:

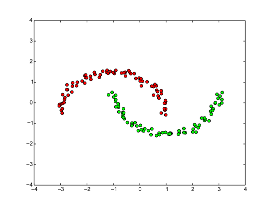
$$R_{D_T}(h) \;\le\; \widehat{R}_S(h) + \frac{4}{n}\sqrt{d\log\frac{2e\,n}{d} + \log\frac{4}{\delta}} + \hat{d}_{\mathcal{H}}(S,T) + \frac{4}{n^2}\sqrt{d\log\frac{2\,n}{d} + \log\frac{4}{\delta}} + \beta$$

with $\beta \ge \inf\limits_{h^*\in\mathcal{H}} \left[R_{D_S}(h^*) + R_{D_T}(h^*)\right].$

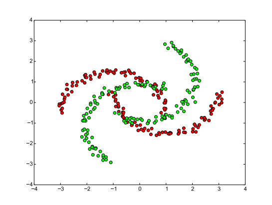
Target risk $R_{D_T}(h)$ is low
if, given S and T,



$\widehat{R}_S(h)$ is small,
i.e., $h \in \mathcal{H}$ is good on



and $\hat{d}_{\mathcal{H}}(S,T)$ is small,
i.e., all $h' \in \mathcal{H}$ are bad on



Domain-Adversarial Neural Network (DANN)

Empirical $\mathcal{H}$-divergence

$$\hat{d}_{\mathcal{H}}(S, T) \stackrel{\text{def}}{=} 2 \max_{h \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathrm{I}[h(\mathbf{x}_i^s) = 1] + \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}[h(\mathbf{x}_i^t) = -1] - 1 \right].$$

We estimate the $\mathcal{H}$-divergence by a logistic regressor that model the probability that a given input (either $\mathbf{x}^s$ or $\mathbf{x}^t$) is from the source domain :

$$o(\mathbf{h}(\mathbf{x})) \stackrel{\text{def}}{=} \mathrm{sigm}(d + \mathbf{w}^\top \mathbf{h}(\mathbf{x})).$$

Given a representation output by the hidden layer $\mathbf{h}(\cdot)$:

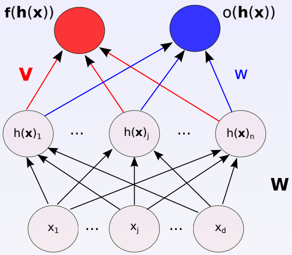
$$\hat{d}_{\mathcal{H}}\Big(\mathbf{h}(S), \mathbf{h}(T)\Big) \approx 2 \max_{\mathbf{w}, d} \left[\frac{1}{n} \sum_{i=1}^{n} \log\big(o(\mathbf{h}(\mathbf{x}_i^s))\big) + \frac{1}{n} \sum_{i=1}^{n} \log\big(1 - o(\mathbf{h}(\mathbf{x}_i^t))\big) - 1 \right].$$

Domain-Adversarial Neural Network (DANN)

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

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

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

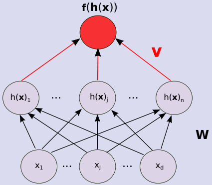


1. Pick a $\mathbf{x}^s \in S$ and $\mathbf{x}^t \in T$
2. Update $\mathbf{V}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
3. Update $\mathbf{W}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
4. Update $\mathbf{w}$ towards $o(\mathbf{h}(\mathbf{x}^s)) = 1$ and $o(\mathbf{h}(\mathbf{x}^t))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

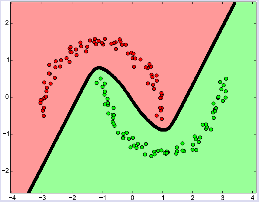
DANN finds a representation $\mathbf{h}(\cdot)$ that are good on S;
but unable to discriminate between S and T.

Toy Dataset

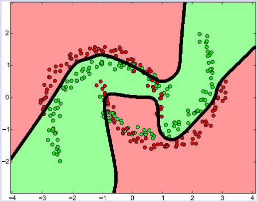
Standard Neural Network (NN)



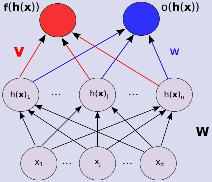
Trained to classify source



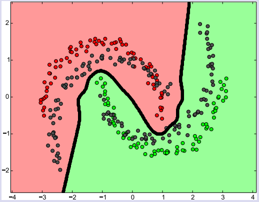
Trained to classify domains



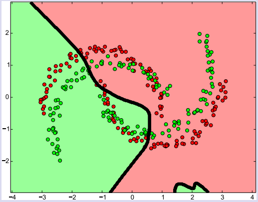
Domain-Adversarial Neural Networks (DANN)



Classification output : $\mathbf{f}(\mathbf{h}(\mathbf{x}))$



Domain output : $o(\mathbf{h}(\mathbf{x}))$



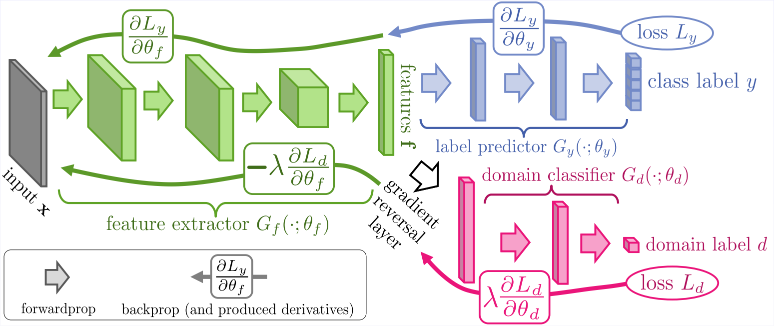
Amazon Reviews

Input : product review (bag of words) — **Output :** positive or negative rating.

Dataset	**DANN**	**NN**
books $\rightarrow$ dvd	0.201	**0.199**
books $\rightarrow$ electronics	**0.246**	0.251
books $\rightarrow$ kitchen	**0.230**	0.235
dvd $\rightarrow$ books	**0.247**	0.261
dvd $\rightarrow$ electronics	**0.247**	0.256
dvd $\rightarrow$ kitchen	0.227	0.227
electronics $\rightarrow$ books	**0.280**	0.281
electronics $\rightarrow$ dvd	**0.273**	0.277
electronics $\rightarrow$ kitchen	**0.148**	0.149
kitchen $\rightarrow$ books	**0.283**	0.288
kitchen $\rightarrow$ dvd	0.261	0.261
kitchen $\rightarrow$ electronics	0.161	0.161

To appear in JMLR : **Domain-Adversarial Neural Networks.**

by Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand and Lempitsky



Plan

Short Summary

An original PAC-Bayesian approach

- General theorem from which we recover existing results ;
- Modular proof, easy to adapt to various frameworks.

Domain adaptation algorithms

- Two algorithms for linear classifiers derived from PAC-Bayesian bounds ;
- One *representation learning* Network inspired by the seminal work of Ben-David et al.

Perspectives

- Explore the relationships between PAC-Bayesian and *truly* Bayesian approaches ;
- Speed-up domain adaptation algorithms with stochastic gradient ;
- Go beyond simple binary classification setting ;
- Apply PAC-Bayes to your problems !

If you only have a hammer,
you tend to see every problem as a nail.

— Abraham Maslow, 1966