PAC-Bayesian Learning and Domain Adaptation

Pascal Germain[1] François Laviolette[1] Amaury Habrard[2] Emilie Morvant[3]

1 **GRAAL Machine Learning Research Group**
Département d'informatique et de génie logiciel
Université Laval, Québec, Canada
{pascal.germain,francois.laviolette}@ift.ulaval.ca

2 **Laboratoire Hubert Curien**
Saint-Étienne University, France
amaury.habrard@univ-st-etienne.fr

3 **Laboratoire d'Informatique Fondamentale**
QARMA Group
Aix-Marseille University, France
emilie.morvant@lif.univ-mrs.fr

NIPS 2012 Workshop: Multi-Trade-offs in Machine Learning,
December 7, 2012

Outline

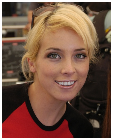
Domain Adaptation (DA) : Problem Description

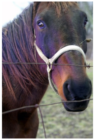
When we need DA

The **Learning** distribution is **different** from the **Testing** distribution.

An example of a DA problem

- We have **labeled** images from a **Web image** corpus
- Is there a `Person` in **unlabeled** images from a **Video** corpus ?





$\xrightarrow{?}$

`Person` `no Person`

Is there a `Person` ?

⇒ How to learn, from the **source domain**, a low-error classifier on the **target one** ?

Domain Adaptation (DA) : Problem Description

Supervised Classification

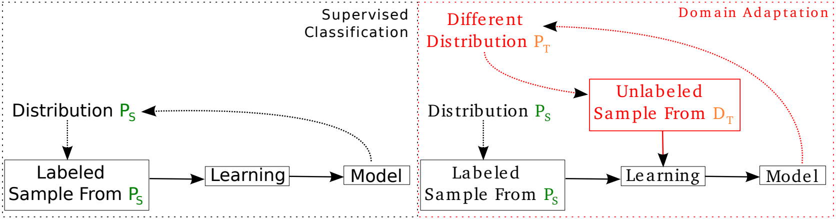
- We consider **binary classification task**: X input space, $Y = \{-1, 1\}$ label set
- P_S **source** domain: distribution over $X \times Y$; D_S marginal distribution over X
- $S \sim (P_S)^m$ a labeled source sample

$\Longrightarrow$ **Objective:** Find a classifier $h \in \mathcal{H}$ with a **low source risk** $R_{P_S}(h)$.

Domain Adaptation

- P_T **target** domain: distribution over $X \times Y$; D_T marginal distribution over X
- $T \sim (D_T)^{m'}$ a unlabeled target sample

$\Longrightarrow$ **Objective:** Find a classifier $h \in \mathcal{H}$ with a **low target risk** $R_{P_T}(h)$.



- Let $\mathcal{H}$ be an hypothesis space.

Theorem [Ben-David et al., 2010]

For every $h \in \mathcal{H}$ and for all $\delta \in]0,1]$, with probability at least $1-\delta$:

$$R_{P_T}(h) \;\leq\; R_{P_S}(h) \;+\; \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \;+\; \lambda,$$

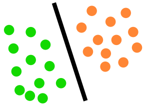
with $\lambda = \min_{h^* \in \mathcal{H}} \left(R_{P_S}(h^*) + R_{P_T}(h^*) \right)$.

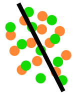
Trade-off between:

- ☐ $R_{P_S}(h)$ is the classical expected error on the source domain

- ☐ $d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T)$ is the $\mathcal{H}\Delta\mathcal{H}$-distance between source and target domains

$$d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) = 2 \sup_{h,h' \in \mathcal{H}\Delta\mathcal{H}} \left| \Pr_{\mathbf{x}\sim D_S}(h(\mathbf{x}) \neq h'(\mathbf{x})) - \Pr_{\mathbf{x}\sim D_T}(h(\mathbf{x}) \neq h'(\mathbf{x})) \right|$$





A New Domain Adaptation Bound *(PAC-Bayesian approach)*

- Let $\mathcal{H}$ be an hypothesis space.
- Given a weight distribution $\rho \sim \mathcal{H}$, we study the ρ-average errors:

$$R_{P_S}(G_\rho) \;=\; \mathop{\mathbf{E}}_{h\sim\rho} R_{P_S}(h)\,, \qquad R_{P_T}(G_\rho) \;=\; \mathop{\mathbf{E}}_{h\sim\rho} R_{P_T}(h)\,.$$

Theorem

For all $\delta \in]0,1]$, with probability at least $1-\delta$, for every posterior distribution ρ:

$$\mathop{\mathbf{E}}_{h\sim\rho} R_{P_T}(h) \;\leq\; \mathop{\mathbf{E}}_{h\sim\rho} R_{P_S}(h) \;+\; \mathrm{dis}_\rho(D_S, D_T) \;+\; \lambda_\rho,$$

with $\lambda_\rho = R_{P_S}(h^\star) + R_{P_T}(h^\star)$, and $h^\star = \operatorname*{argmin}_{h\in\mathcal{H}} \left\{ \mathop{\mathbf{E}}_{h'\sim\rho} \left(R_{D_T}(h,h') - R_{D_S}(h,h') \right) \right\}$.

☐ **Domain disagreement:** $\mathrm{dis}_\rho(D_S,D_T) = \mathop{\mathbf{E}}_{h_1,h_2\sim\rho^2} \left[\Pr_{\mathbf{x}\sim D_S}(h(\mathbf{x}) \neq h'(\mathbf{x})) - \Pr_{\mathbf{x}\sim D_T}(h(\mathbf{x}) \neq h'(\mathbf{x})) \right]$.

Given empirical observations $S \sim (P_S)^m$ and $T \sim (D_S)^{m'}$,

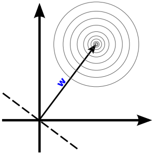
$\Rightarrow$ **We want to minimize :** $B_{P_{\langle S,T\rangle}}(G_\rho) \stackrel{\text{def}}{=} R_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T)\,,$

where $P_{\langle S,T\rangle}$ denotes the joint distribution over $P_S \times D_T$.

PAC-Bayesian Learning of Linear Classifier

[Germain, Lacasse, Laviolette and Marchand, 2009]

- Let $\mathcal{H}$ be a set of linear classifiers $h_{\mathbf{v}}(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{sgn}(\mathbf{v} \cdot \mathbf{x})$
- Consider a prior $\pi_{\mathbf{0}}$ and a posterior $\rho_{\mathbf{w}}$ defined as isotropic Gaussians respectively centered on vectors $\mathbf{0}$ and $\mathbf{w}$.



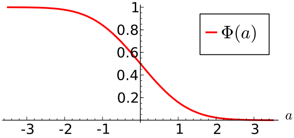
Theorem [Langford and Shawe-Taylor, 2002]

For any domain $P_S \subseteq \mathbb{R}^d \times Y$ *and any* $\delta \in (0,1]$, *we have,*

$$\Pr_{S\sim(P_S)^m}\left(\forall\, \mathbf{w} \in \mathbb{R}^d \;:\; \mathrm{kl}\Big(R_S(G_{\rho_\mathbf{w}}) \,\Big\|\, R_{P_S}(G_{\rho_\mathbf{w}})\Big) \leq \frac{1}{m}\left[\mathrm{KL}(\rho_\mathbf{w} \,\|\, \pi_\mathbf{0}) + \ln\frac{\xi(m)}{\delta}\right]\right) \geq 1-\delta.$$

Trade-off between:

- ☐ $R_S(G_{\rho_\mathbf{w}}) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \Phi\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right)$ is the **sigmoidal loss**
- ☐ $\mathrm{KL}(\rho_\mathbf{w} \,\|\, \pi_\mathbf{0}) \;=\; \frac{1}{2}\|\mathbf{w}\|^2$ is a **regularizer**



Given empirical observations $S \sim (P_S)^m$ and $T \sim (D_S)^{m'}$,

$\Rightarrow$ **We want to minimize :** $B_{P_{\langle S,T\rangle}}(G_\rho) \stackrel{\text{def}}{=} R_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T)\,,$

where $P_{\langle S,T\rangle}$ denotes the joint distribution over $P_S \times D_T$.

Theorem

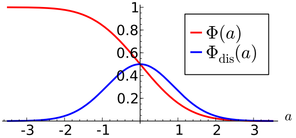
For any domain $P_{\langle S,T\rangle} \subseteq \mathbb{R}^d \times Y \times \mathbb{R}^d$ *and any* $\delta \in (0,1]$, *we have,*

$$\Pr_{\langle S,T\rangle\sim(P_{\langle S,T\rangle})^m}\left(\forall\, \mathbf{w} \in \mathbb{R}^d : \mathrm{kl}\left(B^*_{\langle S,T\rangle} \,\middle\|\, B^*_{P_{\langle S,T\rangle}}\right) \le \frac{1}{m}\left[2\mathrm{KL}(\rho_{\mathbf{w}} \| \pi_{\mathbf{0}}) + \ln\frac{\xi(m)}{\delta}\right]\right) \ge 1-\delta\,,$$

where $B_{P_{\langle S,T\rangle}}(G_{\rho_{\mathbf{w}}}) \;=\; R_{P_S}(G_{\rho_{\mathbf{w}}}) + \mathrm{dis}_{\rho_{\mathbf{w}}}(D_S, D_T)\,.$

Trade-off between:

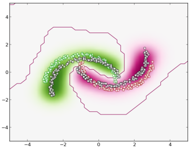
- ☐ $R_{P_S}(G_{\rho_{\mathbf{w}}}) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x}^s,y^s)\sim P_S} \Phi\left(y^s\,\frac{\mathbf{w}\cdot\mathbf{x}^s}{\|\mathbf{x}^s\|}\right)$
- ☐ $\mathrm{dis}_{\rho_{\mathbf{w}}}(D_S, D_T) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x}^s,y^s)\sim P_S} \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}^s}{\|\mathbf{x}^s\|}\right) - \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}^t}{\|\mathbf{x}^t\|}\right)$
- ☐ $\mathrm{KL}(\rho_{\mathbf{w}} \| \pi_{\mathbf{0}}) \;=\; \frac{1}{2}\|\mathbf{w}\|^2$



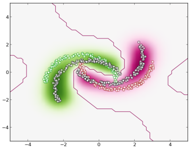
Preliminary Experimental Results

Bound minimization by gradient descent

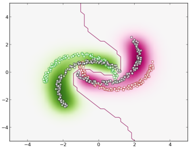
Illustration of the decision boundary on 4 rotations angles:



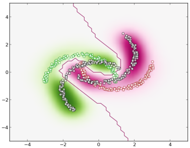
20°



30°

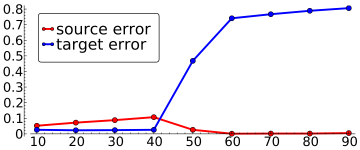


40°



50°

Rotation angle	20°	30°	40°	50°
PBGD	99.5	89.8	78.6	60
SVM	89.6	76	68.8	60
TSVM	**100**	78.9	74.6	**70.9**
DASVM	**100**	78.4	71.6	66.6
DASF	98	92	83	70
DA-PBGD	97.7	**97.6**	**97.4**	53.2



Thank you!

See you at our poster.