PAC-Bayesian Learning and Domain Adaptation

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

[1] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada — {pascal.germain,francois.laviolette}@ift.ulaval.ca

[2] Laboratoire Hubert Curien UMR CNRS 5516, Univ. Jean Monnet, 42000 St-Etienne, France — amaury.habrard@univ-st-etienne.fr

[3] Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France — emilie.morvant@lif.univ-mrs.fr

Introduction, Notations and Motivation

Example:

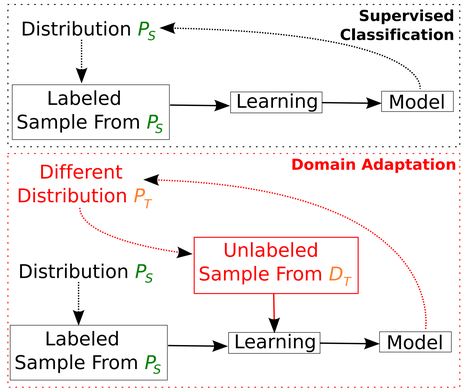
- We have **labeled** images from a **Web image** corpus, *i.e.* $\sim P_S$
- Is there a `Person` in **unlabeled** images from a **Video** corpus, *i.e.* $\sim D_T$?

`Person` `no Person` Is there a `Person` ?

$P_S \neq P_T$

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

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



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

P_T **target** domain: different distribution over $X \times Y$; D_T marginal distribution over X

Risks of a hypothesis $h : X \to Y$: $R_{P_S}(h)$, $R_S(h)$ **source** risks ; $R_{P_T}(h)$, $R_T(h)$ **target** risks

In a classical context

Supervised Classification objective:

$h \in \mathcal{H}$ with a **low** $R_{P_S}(h)$

Domain Adaptation objective:

$h \in \mathcal{H}$ with a **low** $R_{P_T}(h)$

In a PAC-Bayesian context

Supervised Classification objective:

Posterior distribution ρ with a low $R_{P_S}(G_\rho) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim\rho} R_{P_S}(h)$

Domain Adaptation objective:

Posterior distribution ρ with a low $R_{P_T}(G_\rho) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim\rho} R_{P_T}(h)$

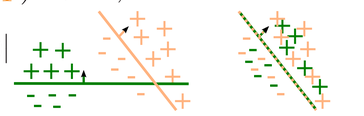
A Classical Domain Adaptation Bound

Theorem 1 ([1]). *Let $\mathcal{H}$ an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$, for every $h \in \mathcal{H}$:*

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

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

and $\nu \stackrel{\text{def}}{=} R_{P_S}(h^*) + R_{P_T}(h^*)$, *with* $h^* \stackrel{\text{def}}{=} \operatorname{argmin}_{h\in\mathcal{H}} (R_{P_S}(h) + R_{P_T}(h))$.



Far domains

Close domains

PAC-Bayesian Learning of Linear Classifier

Theorem 2 ([3]). *Let $\mathcal{H}$ be a set of linear classifiers $h_{\mathbf{v}}(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{sgn}(\mathbf{v}\cdot\mathbf{x})$ such that $\mathbf{v} \in \mathbb{R}^d$ is a weight vector. Consider a prior π_0 and a posterior $\rho_{\mathbf{w}}$ defined as isotropic Gaussians respectively centered on vectors $\mathbf{0}$ and $\mathbf{w}$. For any domain $P_S \subseteq \mathbb{R}^d \times Y$ and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim(P_S)^m} \left(\forall\, \mathbf{w} \in \mathbb{R}^d \;:\; \mathrm{kl}\Big(R_S(G_{\rho_\mathbf{w}}) \,\Big\|\, R_{P_S}(G_{\rho_\mathbf{w}})\Big) \leq \frac{1}{m}\left[\mathrm{KL}(\rho_\mathbf{w} \,\|\, \pi_0) + \ln\frac{\xi(m)}{\delta}\right] \right) \;\geq\; 1-\delta\,,$$

where $R_{P_S}(G_{\rho_\mathbf{w}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \Phi\left(y\, \frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right)$, $\Phi(a) \stackrel{\text{def}}{=} \frac{1}{2}[1 - \mathrm{Erf}(\frac{a}{\sqrt{2}})]$, *and* $\mathrm{KL}(\rho_\mathbf{w} \,\|\, \pi_0) = \frac{1}{2}\|\mathbf{w}\|^2$.

The algorithm **PBGD** [2] minimizes the bound on $R_{P_S}(G_{\rho_\mathbf{w}})$ by gradient descent

A PAC-Bayesian Domain Adaptation Bound

Theorem 3. *Let $\mathcal{H}$ an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$, for every posterior distribution ρ:*

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

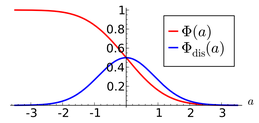
where $\mathrm{dis}_\rho(D_S,D_T) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h_1,h_2\sim\rho^2} [R_{D_T}(h_1,h_2) - R_{D_S}(h_1,h_2)]$ *is the* **domain disagreement** *between D_S and D_T,*

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

- 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 $P_S \times D_T$.
- Similarly to Theorem 2, we use **PAC-Bayesian theory specialized to linear classifiers**.

Theorem 4. *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,*

$$\Pr_{\langle S,T\rangle\sim(P_{\langle S,T\rangle})^m} \left(\forall\, \mathbf{w} \in \mathbb{R}^d : \mathrm{kl}\Big(B^*_{\langle S,T\rangle} \,\Big\|\, B^*_{P_{\langle S,T\rangle}}\Big) \leq \frac{1}{m}\left[2\mathrm{KL}(\rho_\mathbf{w} \,\|\, \pi_0) + \ln\frac{\xi(m)}{\delta}\right] \right) \geq 1-\delta\,,$$

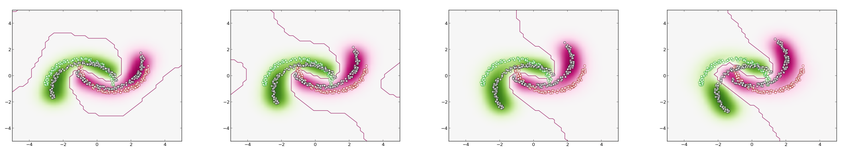


where $B^*_{P_{\langle S,T\rangle}} G_{\rho_\mathbf{w}} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x}^s,y^s,\mathbf{x}^t)\sim P_{\langle S,T\rangle}} \left[\Phi\left(y^s\, \frac{\mathbf{w}\cdot\mathbf{x}^s}{\|\mathbf{x}^s\|}\right) + \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}^s}{\|\mathbf{x}^s\|}\right) - \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}^t}{\|\mathbf{x}^t\|}\right)\right] \cdot \frac{1}{2} + \frac{1}{4}$, *and* $\Phi_{\mathrm{dis}}(a) \stackrel{\text{def}}{=} 2\,\Phi(a)\,\Phi(-a)$.

The algorithm **DA-PBGD** minimizes the bound on $B_{P_{\langle S,T\rangle}}(G_{\rho_\mathbf{w}})$ by gradient descent

Preliminary Experimental Results

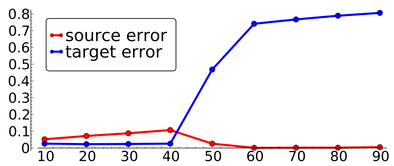
Illustration of the decision of DA-PBGD on 4 rotations angles: From left to right 20°, 30°, 40°, 50°. In green and pink is the source sample, in grey is the target sample.



Average accuracy results for 4 rotation angles. DA-PBGD is more stable than the others and outperforms all the methods for 2 angles.

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

The trade-off between target and source errors according to the difficulty of the task (*i.e.* the rotation angle).



References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Machine Learning Journal*, 2010.

[2] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian Learning of Linear Classifiers. In *Proceedings of ICML*, 2009.

[3] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *Proceedings of NIPS*, 2002.