A New PAC-Bayesian Perspective on Domain Adaptation

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

[1] **INRIA Paris – SIERRA Project-Team**
École Normale Supérieure, Paris, France

[2] **Laboratoire Hubert Curien**
University of Saint-Étienne, France

[3] **Département d'informatique et génie logiciel – GRAAL**
Université Laval, Québec, Canada

ICML New York
June 20, 2016

Unsupervised Domain Adaptation Problem

Binary Classification

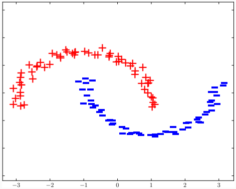
- Input space: $\mathbf{X}$
- Labels: $Y = \{-1, +1\}$

Two different data distributions

- Source domain: $\mathcal{S}$
- Target domain: $\mathcal{T}$

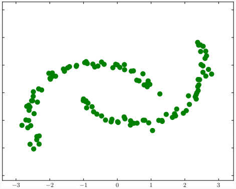
A **domain adaptation learning algorithm** is provided with

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



ADAPTATION $\Longrightarrow$

an **unlabeled target sample**
$T = \{\mathbf{x}_i\}_{i=1}^{m'} \sim (\mathcal{T}_{\mathbf{X}})^{m'}.$



The goal is to build a classifier $h : \mathbf{X} \to Y$ with a low **target risk**:

$$\widehat{\mathbf{R}}_{\mathcal{T}}(h) := \Pr_{\mathcal{T}}\Big(h(\mathbf{x}) \neq y\Big) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathrm{I}\big[h(\mathbf{x}) \neq y\big].$$

$\big(\mathrm{I}[\,\cdot\,]$ *is the indicator function*$\big)$

Previous Approaches

Let $\mathcal{H}$ be a hypothesis class.

Classical domain adaptation theorem (Ben David et al., 2006)

For all hypothesis h in $\mathcal{H}$:

$$\mathbf{R}_{\mathcal{T}}(h) \leq \overbrace{\mathbf{R}_{\mathcal{S}}(h)}^{\text{source risk}} + \overbrace{\sup_{(h,h')\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \mathrm{I}\big[h(\mathbf{x})\neq h'(\mathbf{x})\big] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \mathrm{I}\big[h(\mathbf{x})\neq h'(\mathbf{x})\big] \right|}^{\text{domain divergence}} + \overbrace{\mu_{h^*}}^{\text{non-estimable term}}.$$

Our First PAC-Bayesian domain adaptation theorem (ICML 2013)

For all distribution ρ over $\mathcal{H}$:

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_{\mathcal{T}}(h) \leq \overbrace{\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_{\mathcal{S}}(h)}^{\text{source risk}} + \overbrace{\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \mathrm{I}\big[h(\mathbf{x})\neq h'(\mathbf{x})\big] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \mathrm{I}\big[h(\mathbf{x})\neq h'(\mathbf{x})\big] \right) \right|}^{\text{domain divergence}} + \overbrace{\lambda_{\rho}}^{\text{non-estimable term}}.$$

- **Pro:** The divergence supremum is replaced by a ρ-average. We learned ρ.
- **Con:** The non-estimable term λ_{ρ} relies on ρ. We have to ignore it.

New Approach : Expected Risk Decomposition

Observation (Lacasse, **Laviolette**, Marchand, **Germain**, Usunier, 2006)

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_{\mathcal{T}}(h) \;=\; \frac{1}{2}\,\overbrace{\mathbf{d}_{\mathcal{T}_{\mathbf{X}}}(\rho)}^{\substack{\text{expected}\\ \text{disagreement}}} \;+\; \overbrace{\mathbf{e}_{\mathcal{T}}(\rho)}^{\substack{\text{expected}\\ \text{joint error}}}\,,$$

where, considering $h \sim \rho$ and $h' \sim \rho$,

$$\mathbf{d}_{\mathcal{T}_{\mathbf{X}}}(\rho) := \Pr_{\mathcal{T}}\Big(h(\mathbf{x}) \neq h'(\mathbf{x})\Big) \qquad = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathrm{I}\big[h(\mathbf{x}) \neq h'(\mathbf{x})\big],$$

$$\mathbf{e}_{\mathcal{T}}(\rho) := \Pr_{\mathcal{T}}\Big(h(\mathbf{x}) \neq y \wedge h'(\mathbf{x}) \neq y\Big) \quad = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathrm{I}\big[h(\mathbf{x}) \neq y\big]\,\mathrm{I}\big[h'(\mathbf{x}) \neq y\big].$$

We can estimate $\mathbf{d}_{\mathcal{T}_{\mathbf{X}}}(\rho)$ from a target sample,
but we cannot estimate $\mathbf{e}_{\mathcal{T}}(\rho)$ (since it relies on target labels).

New Approach : Joint Error and Domain Divergence

Estimating target joint error $\mathbf{e}_{\mathcal{T}}$

Let $q > 0$ **:**

$$\mathbf{e}_{\mathcal{T}}(\rho) \;\leq\; \overbrace{\beta_q(\mathcal{T}\|\mathcal{S})}^{\substack{\text{domain}\\ \text{divergence}}} \times \Big[\overbrace{\mathbf{e}_{\mathcal{S}}(\rho)}^{\substack{\text{source}\\ \text{joint error}}} \Big]^{1-\frac{1}{q}} + \overbrace{\eta_{\mathcal{T}\setminus\mathcal{S}}}^{\substack{\text{difference}\\ \text{of supports}}},$$

where

$$\beta_q(\mathcal{T}\|\mathcal{S}) \;:=\; \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \Bigg(\underbrace{\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}}_{\text{weight ratio}} \Bigg)^q \right]^{\frac{1}{q}} \;\in\; [1,\infty)\,,$$

and

$$\eta_{\mathcal{T}\setminus\mathcal{S}} \;:=\; \underbrace{\Pr_{\mathcal{T}}\Big((\mathbf{x},y) \notin \textsc{support}(\mathcal{S}) \Big)}_{\substack{\text{target area}\\ \text{outside source support}}} \times \underbrace{\sup_{h\in\mathcal{H}} \mathbf{R}_{\mathcal{T}\setminus\mathcal{S}}(h)}_{\substack{\text{worst risk}\\ \text{feasible}}}.$$

A New Trade-Off for Domain Adaptation

New domain adaptation theorem

For all ρ **on** $\mathcal{H}$:

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_{\mathcal{T}}(h) \;\leq\; \frac{1}{2}\,\overbrace{\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)}^{\substack{\text{target}\\\text{disagreement}}} + \overbrace{\beta_q(\mathcal{T}\|\mathcal{S})}^{\substack{\text{domain}\\\text{divergence}}} \times \Big[\overbrace{\mathbf{e}_{\mathcal{S}}(\rho)}^{\substack{\text{source}\\\text{joint error}}}\Big]^{1-\frac{1}{q}} + \overbrace{\eta_{\mathcal{T}\setminus\mathcal{S}}}^{\substack{\text{difference}\\\text{of supports}}} .$$

Breaks the **adaptation trade-off** into an atypical trade-off:

1. *Unlabeled* information $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ from the target domain;
2. *Labeled* information $\mathbf{e}_{\mathcal{S}}(\rho)$ from the source domain, weighted by the *source-target divergence* $\beta_q(\mathcal{T}\|\mathcal{S})$ (under the choice of parameter q);
3. *Worst feasible* target error $\eta_{\mathcal{T}\setminus\mathcal{S}}$ in regions where the source domain is uninformative;
   - ⇒ Non-estimable but constant term, does not depend on ρ;
   - ⇒ Should be reasonably small when adaptation is achievable.

Special Case

With $q \to \infty$

For all ρ on $\mathcal{H}$:

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_{\mathcal{T}}(h) \;\leq\; \tfrac{1}{2}\, \overbrace{\mathbf{d}_{\mathcal{T}_{\mathbf{x}}}(\rho)}^{\substack{\text{target}\\ \text{disagreement}}} + \overbrace{\underbrace{\beta_\infty(\mathcal{T}\|\mathcal{S})}_{= \sup\limits_{(\mathbf{x},y)} \frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}}}^{\substack{\text{domain}\\ \text{divergence}}} \times \overbrace{\Big[\mathbf{e}_{\mathcal{S}}(\rho)\Big]}^{\substack{\text{source}\\ \text{joint error}}} + \overbrace{\eta_{\mathcal{T}\setminus\mathcal{S}}}^{\substack{\text{difference}\\ \text{of supports}}} .$$

Linear trade-off between $\mathbf{d}_{\mathcal{T}_{\mathbf{x}}}(\rho)$ and $\mathbf{e}_{\mathcal{S}}(\rho)$:

- $\Rightarrow$ In the covariate shift setting, $\beta_\infty(\mathcal{T}\|\mathcal{S}) = \sup\limits_{\mathbf{x}} \frac{\mathcal{T}(\mathbf{x})}{\mathcal{S}(\mathbf{x})}$ can be estimated from learning samples;
- $\Rightarrow$ We consider $\beta_\infty(\mathcal{T}\|\mathcal{S})$ as a parameter to tune.

Generalization Bound

New PAC-Bayesian Domain Adaptation Theorem

For any prior π over $\mathcal{H}$, any $\delta \in (0,1]$, any real numbers $b > 1$ and $c > 1$, with a probability at least $1-\delta$ over the choices of $S \sim (\mathcal{S})^m$ and $T \sim (\mathcal{T}_{\mathbf{X}})^{m'}$, we have

$\forall \rho$ on $\mathcal{H}$,

$$\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{R}_{\mathcal{T}}(h) \leq \; c \times \frac{1}{2} \overbrace{\widehat{\mathbf{d}}_{T}(\rho)}^{\substack{\text{empirical}\\ \text{target disagreement}}} + b \times \beta_\infty(\mathcal{T}\|\mathcal{S}) \overbrace{\widehat{\mathbf{e}}_{S}(\rho)}^{\substack{\text{empirical}\\ \text{source joint error}}} + \eta_{\mathcal{T}\setminus\mathcal{S}} + \overbrace{O\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}\right)}^{\text{complexity term}}.$$

Learning algorithm for Linear Classifiers

As many PAC-Bayesian works (since Langford and Shawe-Taylor, 2002) :

- We consider the set $\mathcal{H}$ of all linear classifiers $h_{\mathbf{v}}$ in $\mathbf{X} := \mathbb{R}^d$:

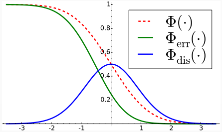
$$h_{\mathbf{v}}(\mathbf{x}) \;=\; \operatorname{sign}(\mathbf{v}\cdot\mathbf{x})\,.$$

- Let $\rho_{\mathbf{w}}$ on $\mathcal{H}$ be a Gaussian distribution centered on $\mathbf{w}$ (with $\Sigma = \mathbf{I}_d$):

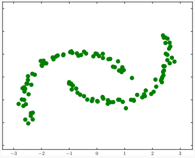
$$h_{\mathbf{w}}(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbf{E}}_{\mathbf{v}\sim\rho_{\mathbf{w}}} h_{\mathbf{v}}(\mathbf{x})\right].$$

Given $T = \{\mathbf{x}_i\}_{i=1}^{m'}$ and $S = \{(\mathbf{x}_j, y_j)\}_{j=1}^{m}$, find $\mathbf{w} \in \mathbb{R}^d$ that minimizes:

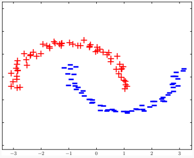
$$C \times \underbrace{\widehat{\mathbf{d}}_T(\rho_{\mathbf{w}})}_{\frac{1}{m'}\sum_i \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)} + B \times \underbrace{\widehat{\mathbf{e}}_S(\rho_{\mathbf{w}})}_{\frac{1}{m}\sum_j \Phi_{\mathrm{err}}\left(y_j\frac{\mathbf{w}\cdot\mathbf{x}_j}{\|\mathbf{x}_j\|}\right)} + \underbrace{\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})}_{\frac{1}{2}\|\mathbf{w}\|^2}.$$



Margin $y\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}$ losses



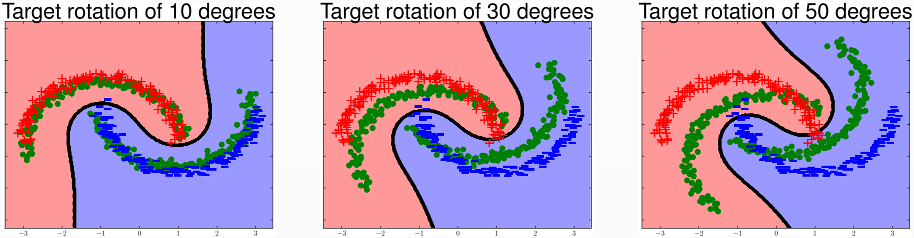
Low density region on target



Classification accuracy on source

Toy Experiment

- RBF kernel
- $B = 1$
- $C = 1$



Empirical results on Amazon Dataset

- Linear kernel
- Hyper-parameter selection by reverse cross-validation

	SVM	DASVM	CODA	**ICML2013**	**ICML2016**
books→DVDs	*0.179*	0.193	0.181	0.183	**0.178**
books→electro	0.290	*0.226*	0.232	0.263	**0.212**
books→kitchen	0.251	**0.179**	0.215	0.229	*0.194*
DVDs→books	0.203	0.202	0.217	*0.197*	**0.186**
DVDs→electro	0.269	**0.186**	*0.214*	0.241	0.245
DVDs→kitchen	0.232	0.183	*0.181*	0.186	**0.175**
electro→books	0.287	0.305	0.275	**0.232**	*0.240*
electro→DVDs	0.267	**0.214**	0.239	*0.221*	0.256
electro→kitchen	*0.129*	0.149	0.134	0.141	**0.123**
kitchen→books	0.267	0.259	*0.247*	*0.247*	**0.236**
kitchen→DVDs	0.253	**0.198**	0.238	0.233	*0.225*
kitchen→electro	0.149	0.157	0.153	**0.129**	*0.131*
Average	0.231	*0.204*	0.210	0.208	**0.200**

Conclusion

Highlights

- We introduced a new domain adaptation trade-off, relying on:
  – the target disagreement $\mathbf{d}_{\mathcal{T}_{\mathbf{X}}}$;
  – the source joint error $\mathbf{e}_{\mathcal{S}}$;
  $\Rightarrow$ Weighted by the domain divergence $\beta_q(\mathcal{T}\|\mathcal{S})$.
- We designed a learning algorithm minimizing a PAC-Bayesian guarantee.

Future Work

Explore the covariate-shift setting:

- Estimate the domain divergence $\beta_q(\mathcal{T}\|\mathcal{S})$
  $\Rightarrow$ Could motivate an *instance reweighting approach.*
- Estimate the "area" covered by the unknown term $\eta_{\mathcal{T}\setminus\mathcal{S}}$
  $\Rightarrow$ Could be reduced by *learning a new representation*.

Poster Tuesday morning — Thank you!