Domain Adaptation for binary classification
- the covariate shift case

Shai Ben-David
University of Waterloo

Based on joint work with Ruth Urner

NIPS 2015 transfer learning workshop
Most of the statistical learning guarantees are based on assuming that the learning environment is unchanged throughout the learning process.
Most of the statistical learning guarantees are based on assuming that the learning environment is unchanged throughout the learning process.

Formally, it is common to assume that both the training and the test examples are generated i.i.d. by the same fixed probability distribution.
Most of the statistical learning guarantees are based on assuming that the learning environment is unchanged throughout the learning process.

Formally, it is common to assume that both the training and the test examples are generated i.i.d. by the same fixed probability distribution.

This is unrealistic for many ML applications.
Learning when Training and Test distributions differ

Examples:

- Spam filters – *train on email arriving at one address, test on a different mailbox.*
- Natural Language Processing tasks – *train on some content domains, test on others.*
- Medical diagnosis of some disease – *train on patient in one country and apply to patients in another.*
The focus of this talk

- Gain theoretical understanding of the circumstances under which domain adaptation is possible

- Examine algorithmic paradigms that will establish such learning.
Formalism

Domain: $\mathcal{X}$
Label set: $\{0, 1\}$
Source Distribution: $P_S$ over $\mathcal{X} \times \{0, 1\}$
Target Distribution: $P_T$ over $\mathcal{X} \times \{0, 1\}$

A DA-learner gets a labeled sample $S$ from the source and a (large) unlabeled sample $T$ from the target and outputs a label predictor

$$h : \mathcal{X} \rightarrow \{0, 1\}.$$

Goal: Learn a predictor with small target error

$$L_{P_T}(h) := \Pr_{(x,y) \sim P_T}[h(x) \neq y] \leq \epsilon.$$
Necessary assumptions for the success of DA learning

Clearly, the success of such learning relies on relationships between the source and target data distributions.

Relationship between the source and target labeling rules:

1. Covariate Shift - for all \( x \in X \), \( P_S(1|x) = P_T(1|x) \).

2. For some \( h : X \rightarrow \{0,1\} \) in a predetermined class \( H \), \( L_{P_T}(h) + L_{P_S}(h) \) is small.

3. \( L_{P_T}(h^*_S) + L_{P_S}(h^*_T) \) is small, where \( h_S \) and \( h_T \) are \( H \)-optimal classifiers of \( P_S \) and \( P_T \), respectively.

Let us focus on the strongest of those - **Covariate Shift**.
Example:

Let $X$ be the unit interval,

$P_S$ the uniform distribution over $[0, 0.5] \times \{0\}$

and $P_T$ the uniform distribution over $(0.5, 1] \times \{1\}$.

*(the common labeling rule is 0 for $x \leq 0.5$ and 1 for $x > 0.5$).*
Relationship between the marginal source and target distributions

Several notions of discrepancy between those marginal distributions have been examined.

Let us consider a strong such connection:

**Bounded point-wise weight ratio**

For some $c > 0$, for all $x$ in the support of $P_T$,

\[
\frac{P_S(x)}{P_T(x)} \geq c
\]

(For simplicity, we assume that $X$ is countable.)
Given a source generated labeled training sample \( \hat{S} = ((x_1, \ell_1), \ldots (x_m, \ell_m)) \),

1. Reweigh each \((x_i, \ell_i) \in \hat{S}\) by \(\tau_i = \frac{P_T(x_i)}{P_S(x_i)}\).

2. Feed the reweighed sample \(\hat{S}_r = (\tau_1(x_1, \ell_1), \ldots \tau_m(x_m, \ell_m))\) to a standard classification learner.

3. Apply the resulting classifier to the target task.
An observation using a point-wise weight ratio assumption

Under the above weight ratio assumption, for every $h \in \{0, 1\}^\mathcal{X}$

$$L_{PT}(h) \leq \frac{1}{c} L_{PS}(h).$$
An observation using a **point-wise** weight ratio assumption

Under the above weight ratio assumption, for every $h \in \{0, 1\}^\mathcal{X}$

$$L_{P_T}(h) \leq \frac{1}{c} L_{P_S}(h).$$

Thus, any algorithm that $(\epsilon, \delta)$-learns the source for **arbitrarily small** $\epsilon$ and $\delta$ also learns the target.
An observation using a point-wise weight ratio assumption

Under the above weight ratio assumption, for every $h \in \{0, 1\}^X$

$$L_{P_T}(h) \leq \frac{1}{c} L_{P_S}(h).$$

Thus, any algorithm that $(\epsilon, \delta)$-learns the source for arbitrarily small $\epsilon$ and $\delta$ also learns the target.

In fact, this “Importance Weighting”-based DA under Covariate Shift is quite common in more application oriented research ([Sugiyama and Mueller, 2005] and many works since).
A first drawback of the point-wise weight ratio assumption

The result may become meaningless if the learner cannot achieve zero error on the source task.

E.g., if \( \inf \{ L_{P_S}(h) : h \in H \} \).

(e.g. when the labeling rule is not deterministic or due to non-zero approximation error of the class of predictors, \( H \), that the algorithm considers).
A second drawback of the point-wise weight ratio assumption

A bound on the point-wise weight ratio is a rather strong assumption.

For example, consider some NLP task, where the domain set is the English dictionary.

If our source domain are *Biology research papers* and our target domain are *legal documents*, such a bound means that no word that occurs with some non-negligible frequency in legal documents, may have negligible frequency in Biology papers.
A second drawback of the point-wise weight ratio assumption

A bound on the point-wise weight ratio is a rather strong assumption.

For example, consider some NLP task, where the domain set is the English dictionary.

If our source domain are *Biology research papers* and our target domain are *legal documents*, such a bound means that no word that occurs with some non-negligible frequency in legal documents, may have negligible frequency in Biology papers.

Still, this assumption can be readily relaxed to allow failure on a set of instances of low target probability, as well as other relaxations.
In most DA applications, such a point-wise weight ratio needs to be deduced from empirical samples.
Learning the source-target weight ratios

In most DA applications, such a point-wise weight ratio needs to be deduced from empirical samples.

How expensive can such learning be, in terms of the required samples sizes?
Learning the source-target weight ratios

In most DA applications, such a point-wise weight ratio needs to be deduced from empirical samples.

How expensive can such learning be, in terms of the required samples sizes?

[B-D, Urner 2012] prove that under the assumption of bounded point-wise weight ratio (and covariate shift) any DA algorithm may fail badly as long as the number of training data $s + t$ is smaller than $\sqrt{|X|}$ (where $X$ is the domain set).
A third type of assumptions

So far, we have considered strong assumptions of the relationships of labeling rules and on the marginal distributions between the source and target tasks.
A third type of assumotions

So far, we have considered strong assumptions of the relationships of labeling rules and on the marginal distributions between the source and target tasks.

It seems that, if one wishes to have success guarantees fro DA learning, a third set of assumptions is needed:

Assumptions imposing ”well behavedness” of the target task
Probabilistic Lipschitzness ([Urner, Shalev-Shwartz, BD, 2011])

We say that $\ell : \mathcal{X} \to \mathbb{R}$ is $\phi$-Lipschitz with respect to a distribution $P_{\mathcal{X}}$ over $\mathcal{X}$ if, for all $\lambda > 0$:

$$\Pr_{x \sim P_{\mathcal{X}}} \left[ \Pr_{y \sim P_{\mathcal{X}}} \left[ |\ell(x) - \ell(y)| > 1/\lambda \|x - y\| \right] > 0 \right] \leq \phi(\lambda)$$

Essentially, the condition asserts that the boundaries between class-labels go through sparsely populated domain regions.
Probabilistic Lipschitzness ([Urner, Shalev-Shwartz, BD, 2011])

We say that \( \ell : \mathcal{X} \rightarrow \mathbb{R} \) is \( \phi \)-Lipschitz with respect to a distribution \( P_\mathcal{X} \) over \( \mathcal{X} \) if, for all \( \lambda > 0 \):

\[
\Pr_{x \sim P_\mathcal{X}} \left[ \Pr_{y \sim P_\mathcal{X}} \left[ |\ell(x) - \ell(y)| > \frac{1}{\lambda} \|x - y\| \right] > 0 \right] \leq \phi(\lambda)
\]

Essentially, the condition asserts that the boundaries between class-labels go through sparsely populated domain regions.

This may be viewed as a formalization of the, often loosely stated, \textit{cluster assumption}. 
Relaxing the point-wise weight ratio assumption

Weight ratio for collection of sets

For some $\eta > 0$ we define the $\eta$-weight ratio of the source distribution and the target distribution with respect to $B \subset 2^B$ as

$$C_{B,\eta}(P^S_X, P^T_X) = \inf_{b \in B} \frac{P^S_X(b)}{P^T_X(b)}$$

We will add the assumption that $C_{B,\eta}(P^S_X, P^T_X) > 0$ w.r.t. the set $B$ of balls of some radius and $\eta$ some small constant.
Relaxing the point-wise weight ratio assumption

Weight ratio for collection of sets

For some \( \eta > 0 \) we define the \( \eta \)-weight ratio of the source distribution and the target distribution with respect to \( \mathcal{B} \subset 2^\mathcal{B} \) as

\[
C_{\mathcal{B}, \eta}(P^S_X, P^T_X) = \inf_{b \in \mathcal{B}} \frac{P^S_X(b)}{P^T_X(b)} \quad \text{subject to} \quad P^T_X(b) \geq \eta
\]

We will add the assumption that \( C_{\mathcal{B}, \eta}(P^S_X, P^T_X) > 0 \) w.r.t. the set \( \mathcal{B} \) of balls of some radius and \( \eta \) some small constant.

Such an assumption can be empirically verified from relatively small samples.
A very simple DA learner under those assumptions

Nearest Neighbor for domain adaptation

Algorithm:
Given a labeled sample $S$ from the source,

label each test point $t$ from the target by its nearest neighbor in $S$. 
Nearest-Neighbor learning guarantees

Theorem (BD, Urner, Shalev-Swartz 2012)

Let our domain $\mathcal{X} = [0, 1]^d$ be the unit cube in $\mathbb{R}^d$ and let $\mathcal{W}$ be the class of pairs $(P_S, P_T)$ of source and target distributions over $\mathcal{X} \times \{0, 1\}$ with $C_B(D_S, D_T) = C > 0$ satisfying the covariate shift assumption and their common labeling function $l : \mathcal{X} \to [0, 1]$ satisfying some Lipschitz property. Then, for all $\lambda$ we have

$$\mathbb{E}_{S \sim P_S^m}[\text{Err}_{P_T}(h_{NN})] \leq 2 \text{opt}(P_T) + \phi(\lambda) + 4\lambda \frac{\sqrt{d}}{C} \left( \frac{1}{m} \right)^{\frac{1}{d+1}}.$$
Theorem (BD, Urner, Shalev-Swartz 2012)

Let our domain $\mathcal{X} = [0, 1]^d$ be the unit cube in $\mathbb{R}^d$ and let $\mathcal{W}$ be the class of pairs $(P_S, P_T)$ of source and target distributions over $\mathcal{X} \times \{0, 1\}$ with $C_B(D_S, D_T) = C > 0$ satisfying the covariate shift assumption and their common labeling function $l : \mathcal{X} \to [0, 1]$ satisfying some Lipschitz property. Then, for all $\lambda$ we have

$$
\mathbb{E}_{S \sim P_S^m}[\text{Err}_{P_T}(h_{\text{NN}})] \leq 2\text{opt}(P_T) + \phi(\lambda) + 4\lambda \frac{\sqrt{d}}{C} \left( \frac{1}{m} \right)^{\frac{1}{d+1}}.
$$

[BD, Urner 2012], show almost matching lower bounds on the needed training sample sizes.
Utilizing the unlabeled target sample

A "proper learning" scenario:

Often, classes that are desirable for prediction (e.g. fast at prediction time or readily interpretable) require large amounts of labeled training data. On the other hand, there are methods that are easier to learn but output obscure predictors.
A "proper learning" scenario:

Often, classes that are desirable for prediction (e.g. fast at prediction time or readily interpretable) require large amounts of labeled training data. On the other hand, there are methods that are easier to learn but output obscure predictors.

**Step 1** Use the source labeled sample to label the target sample $T$ by $NN(S)$.

**Step 2** Feed that now-labeled sample to a fully supervised $H_T$-learner (where $HT$ is the desired target class of useful predictors).
The prior knowledge about the task that the learner has

Can one avoid exponential dependence of the source sample size on the domain dimension?

Another aspect determining a DA problem is prior domain knowledge available to the earner.

Can knowledge of some class of predictors, $H_T$ that has low approximation error w.r.t. the target data distribution help reduce required sample sizes?
Learning with extremely good target class (realizability)

Assume

- $P_T$ is realizable by $H$ with margin $\lambda$ (but $P_S$ is not necessarily)
- $P$ satisfies Probabilistic Lipschitzness with $\phi(\lambda) \leq \epsilon$
- Source and target satisfy weight ratio w.r.t. $H \Delta H \cap C$

Then we can $\epsilon$-learn the target distribution with a reweighing algorithm if

$$|S| \geq O \left( \frac{1}{\epsilon} \right) \quad \text{and} \quad |T| \geq O \left( \frac{1}{\lambda^d \epsilon} \right)$$
Learning with extremely good target class (realizability)

Assume

- $P_T$ is realizable by $H$ with margin $\lambda$ (but $P_S$ is not necessarily)
- $P$ satisfies Probabilistic Lipschitzness with $\phi(\lambda) \leq \epsilon$
- Source and target satisfy weight ratio w.r.t. $H\Delta H \cap C$

Then we can $\epsilon$-learn the target distribution with a reweighing algorithm if

$$|S| \geq O\left(\frac{1}{\epsilon}\right) \quad \text{and} \quad |T| \geq O\left(\frac{1}{\lambda^d \epsilon}\right)$$

However, the realizability assumption is unrealistically strong, and, regretfully, necessary for such a result [U+FFFĐ]
Yet another well-behavedness assumption

**Concentratedness of the marginal distribution**

We say that a probability distribution $P$ over $\mathcal{X}$ is $(\alpha, \beta)$-concentrated with respect to $\mathcal{B} \subseteq 2^{\mathcal{X}}$ if

$$P \left[ \bigcup \{ b \in \mathcal{B} : P_X(b) \geq \alpha \} \right] > 1 - \beta$$

In other words, upper bound the probability weight of light-weight balls.
Yet another well-behavedness assumption

Concentratedness of the marginal distribution
We say that a probability distribution $P$ over $\mathcal{X}$ is $(\alpha, \beta)$-concentrated with respect to $B \subseteq 2^{\mathcal{X}}$ if

$$P \left[ \bigcup \{ b \in B : P_{\mathcal{X}}(b) \geq \alpha \} \right] > 1 - \beta$$

In other words, upper bound the probability weight of light-weight balls.

This is the same spirit as assuming some low intrinsic dimension of the data.
Learnability result

Assume

- Concentration of the marginal
  \[ P^T \left[ \bigcup \{ b \in B : P^T(b) \geq \epsilon/2 \} \right] > 1 - \epsilon. \]

- Margin w.r.t. \( H \)
  There exist some \( h^* \in H \) that has \((2\rho, \epsilon/2)\)-margin w.r.t. \( P^T \chi \) and \( \text{Err}_{P^T}(h^*) \leq \text{opt}_{P^T}(H) + \epsilon. \)

- Weight-ratio for cells
  \( \mathcal{B}, \epsilon/4(P^S, P^T) \geq C. \)

Then \( \text{ERM}_{\hat{H}}(w(T, B, \eta)(S)) \) is an \( 3\epsilon \)-successful DA learner for \( \mathcal{W} \), for sample sizes

\[ |S| = O \left( \frac{1}{C\epsilon} \right) \quad \text{and} \quad |T| \geq O \left( \frac{1}{\epsilon^4} \right) \]
Addressing under-coverage of the target by the source distribution

Throughout this talk, we have assumed that the source distribution does not miss any regions in which the target has significant weight.

Can one do DA learning when this assumption is relaxed?
Addressing under-coverage of the target by the source distribution

Throughout this talk, we have assumed that the source distribution does not miss any regions in which the target has significant weight.

Can one do DA learning when this assumption is relaxed?

In a recent paper Berlind and Urner address this scenario by proposing the addition of an active learning component (actively querying labels of the target sample points).

They propose an algorithm that is town to handle such scenarios, provably outperforming each of DA with respect to only source labels and Active Learning with respect to only target samples.
Summary

We investigated which assumptions allow which kind of replacement of “perfect” data by “proxy” data:

- For some algorithms labeled source data suffices:
  - Learners that achieve arbitrary small error on the source (source realizability) and have access to very large target samples.
  - Nearest Neighbor w.r.t very large source labeled samples.

- There are scenarios where (unlabeled) target data is provably necessary and beneficial:
  - Proper DA-learning.
  - When there is prior knowledge about a class of predictors that do well on the target task.
Many open questions

- Which assumptions make sense in practice?
- Are there adaptive algorithms that can guaranteed to succeed based on (more) realistic assumptions?
- Analyze the utility of (relatively few) labeled target-generated examples.