ON THE HARDNESS OF DOMAIN ADAPTATION

Tamas Madarasz & Michael Rabadi

April 15, 2015

Questions

- DA under the covariate shift assumption can succeed if we are given/can estimate the weight ratios. But what happens if we are not given these ratios and some (all?) of the points in the **source** sample are not found in the **target** sample?

- Can we find satisfying error guarantees without prior knowledge about the **source** but strong prior knowledge about the **target** task?

- Can we trade the 'expensive' labeled examples from the **source** distribution for unlabeled **target**-generated data.

Assumptions

- Covariate shift
- Pointwise density ratio between source and target distributions is bounded below by 0.5
- Realizable hypothesis class with smallest possible VC dimension

Definitions I

Solvability:

Let $\mathscr{W}$ be a class of triples (P_S, P_T, l) of source and target distributions over some domain $\mathscr{X}$ and a labeling function l.

The DA learner $\mathscr{A}$ **(ϵ, δ,m, n)-solves** DA for the class $\mathscr{W}$, if for
$\forall$ triples $(P_S, P_T, l) \in \mathscr{W}$,
labeled sample S of size m i.i.d. from P_S,
unlabeled sample T of size n from P_T ,

with probability at least $1 - \delta$ $\mathscr{A}$ outputs a function h with $Err_l^T(h) \leq \epsilon$.

Definitions II

d_A distance

$\mathbf{d}_{H\triangle H}$ distance: $d_{H\triangle H}(P_S, P_T) = \sup_{A \in H\triangle H} |P_T(A) - P_S(A)|$

where $H\triangle H = \{h_1 \triangle h_2 | h_1, h_2 \in H\}$,

$h_1 \triangle h_2 = \{x \in \mathscr{X} | h_1(x) \neq h_2(x)\}$

Definitions III

Weight Ratio: For $\mathscr{B} \subseteq 2^{\mathscr{X}}$, source and target distributions P_S and P_T

$$C_{\mathscr{B}}(P_S, P_T) = \inf_{b \in \mathscr{B}, P_T(b) \neq 0} \frac{P_S(b)}{P_T(b)}$$

We use $C(P_S, P_T)$ if $\mathscr{B}$ is the collection of all sets that are P_S and P_T -measurable.

We will bound the wight ratio by $C(P_S, P_T) \geq 1/2$.

Theorem I

Theorem 1: Let $\mathscr{X}$ be a finite domain, and $\epsilon + \delta < \frac{1}{2}$.

No algorithm can (ϵ, δ, s, t)-solve the DA problem for the class $\mathscr{W}$ of triples (P_S, P_T, l) with

- $C(P_S, P_T) \geq 1/2$
- $d_{H_{1,0} \triangle H_{1,0}}(P_S, P_T) = 0$
- $opt_T^l(H_{1,0}) = 0$

if $s + t < \sqrt{(1 - 2(\epsilon + \delta))|\mathscr{X}|} - 2$.

We prove this theorem by reducing the related **Left/Right Problem** (Kelly et al. 2010) to Domain Adaptation under specific assumptions.

Left/Right Problem

The Left/Right problem: Given three finite samples, L, R and M of points from some domain set $\mathscr{X}$, with

- L i.i.d. sample from some distribution P,
- R an i.i.d. sample from some distribution Q over $\mathscr{X}$,

If M is an i.i.d. sample generated by one of these two probability distributions, can we tell which one it is?

Left/Right Problem II

An algorithm $(\boldsymbol{\delta}, \mathsf{l}, \mathsf{r}, \mathsf{m})$-**solves** the L/R problem if, given samples L, R and M of sizes l, r and m respectively, it gives the correct answer with probability at least $1 - \delta$.

Left/Right Problem III

More formally

Let $\mathscr{W}_n^{uni} = \{(U_A, U_B, U_C) : A \cup B = \{1, ...n\}, A \cap B = \oslash, |A| = |B|$, and $C = A$ or $C = B\}$, where U_Y denotes the uniform distribution over the set Y.

Lemma 1: For any given sample sizes l for L, r for R and m for M and any $0 < \gamma < 1/2$, if $k = max\{l, r\} + m$, then for $n > (k+1)^2/(1-2\gamma)$ no algorithm has probability of success greater than $1 - \gamma$ over the class $\mathscr{W}_n^{uni}$. .

Left/Right Problem IV

Idea for Proof (Batu et al. 2010): The L/R problem is permutation invariant (doesn't depend on permutations of X). Find large enough n that with high probability no element is repeated more than once across the three samples.

Proposition 1: Let X be a finite domain of size n. For every $0 < \delta < 1$, with probability $> (1 - \delta)$, an i.i.d. sample of size at most $\sqrt{\delta n} - \delta$ uniformly drawn over X , contains no repeated elements.

Left/Right Problem V

Permutation invariance

For multi-sets L, R, M of size $\leq n$, from distributions P or Q over some domain X the **fingerprint F** is defined as $\{C_{i,j,k} | 1 \leq i, j, k \leq n\}$, where $C_{i,j,k}$ is the number of elements of X , that appear exactly i times in L, j times in R and k times in M.

Proposition 2: If there exists an algorithm A for testing some permutation-invariant property of distributions, then there exists an algorithm for that same task that gets as input only the fingerprints of the samples that A takes and enjoys the same guarantee on its probability of success.

Left/Right Problem VI

Proof of Lemma 1 Let $\delta = 1 - 2\gamma$. Then with probability $> (1 - \delta)$ the input to the Left/Right problem over $\mathscr{W}_n^{uni}$ has no repeated elements, and the fingerprint F has

$C_{1,0,0} = l, C_{0,1,0} = r, C_{0,0,1} = m$, and $C_{i,j,k} = 0$ for all other i, j and k, *independently* of M coming from U_A or U_B.

Let A be some algorithm, $p = P(\text{A outputs } U_A \mid F)$

If $p \geq 0.5$, then P(A errs) $> (1 - \delta)/2$ for all triples where C is equal to B.

If $p \leq 0.5$, then P(A errs) $> (1 - \delta)/2$ on all triples where C is equal to A.

But $(1 - \delta)/2 = \gamma$,thus, no algorithm can (γ, l, r, m)-solve the Left/Right problem for the class $\mathscr{W}_n^{uni}$.

Left/Right Problem to DA

How do we get from DA to the L/R problem?

For n, let $\mathscr{W}_n^{DA}$ be the class of triples (P_S, P_T, l), where

- $\mathscr{X}$ is a finite set of size n,
- $P_S \sim U(\mathscr{X})$
- P_T is uniform over some subset U of X of size n/2
- $l(x) = 1$ for $x \in U$ and $l(x) = 0$ for $x \in \mathscr{X} \setminus U$.

Notably $C(P_S, P_T) = 1/2$.

Lemma 2. For $n \in \mathscr{N}$ and an algorithm A that can (ϵ, δ, s, t)-solve DA for $\mathscr{W}_n^{DA}$ given that the Target is realizable by $H_{1,0}$, we can construct an algorithm that $(\epsilon + \delta, s, s, t + 1)$-solves the Left/Right problem on $\mathscr{W}_n^{uni}$.

Left/Right Problem to DA II

Proof

Given sample $L = \{l_1, l_2, ..., l_s\}$ and $R = \{r_1, r_2, ..., r_s\}$, sample M of size $t + 1$ for L/R in $\mathscr{W}_n^{uni}$

Let $T = M \backslash \{p\}, p \in M$ chosen uniformly at random.

Construct S from $L \times \{0\} \cup R \times \{1\}$ by successively flipping an unbiased coin, and choosing the next element from $L \times \{0\} \cup R \times \{1\}$ accordingly.

Then P_T of this DA problem has marginal either U_A or U_B and the labeling function of this Domain Adaptation instance is $l(x) = 0$ if $x \in A$ and $l(x) = 1$ if $x \in B$.

If $\mathscr{A}$ outputs h on input S and T, then Left/Right problem outputs U_A if $h(p) = 0$, U_B if $h(p) = 1$, and $(\epsilon + \delta, s, s, t + 1)$-solves the Left/Right problem.

Theorem 1 concluded

Thus if an algorithm A that can (ϵ, δ, s, t)-solve DA by Lemma 1 we have $|X| \leq (s + (t + 1) + 1)^2/(1 - 2(\epsilon + \delta))$, thus

$s + t \geq \sqrt{(1 - 2(\epsilon + \delta))|\mathscr{X}|} - 2$ completing the proof of Theorem 1.

Corollary:

Theorem 2:

Let $\mathscr{X} = [0, 1]^d$, ϵ, $\delta > 0$ s.t $\epsilon + \delta < 1/2$.

Let $\lambda > 1$ and let $\mathscr{W}$ be the set of triples (P_S, P_T, l) of distributions over $\mathscr{X}$, with covariance shift and realizability as before, and λ-Lipschitz labeling function l.

Then no DA-learner can (s, t, ϵ, δ)-solve the DA-problem for the class $\mathscr{W}$ unless $s + t \geq \sqrt{(\lambda + 1)^d(1 - 2(\epsilon + \delta))} - 2$.

Theorem 2

Proof:

Let $G \subseteq \mathscr{X}$ be the points of a grid in $[0,1]^d$ with distance $1/\lambda$. Then $|G| = (\lambda+1)^d$, and any labeling function $l : G \to \{0,1\}$ is λ-Lipschitz.

As G is finite, the bound follows from Theorem 1.

Citations

Ben-David, S., and Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In Algorithmic Learning Theory (pp. 139-153). Springer Berlin Heidelberg.

Bernstein, S. (1946). The theory of Probabilities.

Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. In Advances in neural information processing systems (pp. 442-450).

Jin, J., Wilson, J., Nobel, A. (2014). Jensen's and Holder's Inequalities.

Kelly, B.G., Tularak, T., Wagner, A.B., Viswanath, P.: Universal hypothesis testing in the learning-limited regime.(2010) IEEE, ISIT.

Algorithm

Algorithm A Input An i.i.d. sample S from P_S labeled by l, an unlabeled i.i.d. sample T from P_T, and a margin parameter γ.

Step 1 Partition the domain $[0,1]^d$ into a collection B of boxes (axis-aligned rectangles) with sidelength $(\gamma/\sqrt{d})$.

Step 2 Obtain sample $S'$by removing every point in S, which is sitting in a box that is not hit by T.

Step 3 Output an ERM classifier from H for the sample S'