Robust Domain Adaptation*

Yishay Mansour
Tel Aviv University
mansour@tau.ac.il

Mariano Schain
Tel Aviv University
marianos@tau.ac.il

Abstract

We derive a generalization bound for domain adaptation by using the properties of robust algorithms. Our new bound depends on λ-shift, a measure of prior knowledge regarding the similarity of source and target domain distributions. Based on the generalization bound, we design SVM variants for binary classification and regression domain adaptation algorithms.

Introduction

Learning algorithms are used to make decisions. The decision may be whether to classify an incoming email message as spam, to choose the next play in a game, to assist in the prognosis given a medical syndrome, and so on. Learning algorithms are expected to generalize from past examples (*training data*) to new unseen examples(*test data*). Indeed, fundamental results in the classical learning setting (Valiant 1984; Vapnik 1998) provide generalization bounds that relate the number of training observations to the expected performance of learning algorithms. However, an underlying assumption of those results is that the training observations available to the learning algorithm are drawn from the same distribution over which the algorithm is tested.

Nevertheless, in many practical situations (such as computer vision, speech recognition, and natural language processing) such assumption might not hold, either due to lack of control over the underlying environment or simply due to lack of labeled training examples. Domain adaptation addresses situations in which the nature of the training observations (the source *domain*) differs from that of the test observations (the target domain). In the classification setting for example, the domain may include the probability distribution of labeled examples and the labeling function.

In the domain adaptation setting addressed in this paper, the goal is to utilize labeled data from the source domain and unlabeled data from the target domain, to generate a good hypothesis for the target domain. As typical in many applications, abundance of unlabeled data from the target domain is assumed (in contrast to the associated cost of attaining labeled data).

It is natural to expect that the ability of the learning algorithm to generalize will depend significantly on the similarity of the source and target distributions. Naturally, if the two distribution are statistically indistinguishable (have a small total variation distance, i.e., a small L_1 norm distance), then simply learning with respect to the source labeled data would be a very beneficial strategy. However, one can get a good domain adaptation even in cases when the two distributions are statistically very far, or even have disjoint support. The d_A-distance (Ben-David et al. 2007; Blitzer et al. 2007; Ben-David et al. 2010) or the related discrepancy distance (Mansour, Mohri, and Rostamizadeh 2009a; 2009b) are similarity measures which are based on the hypothesis class, and can be used to derive generalization bounds. The nature of those generalization bounds is that they relate the observed error on the source domain to the expected error on the target domain. From this perspective they should be viewed more as studying what guarantee we can give, when we learn with respect to the source domain and later are tested with respect to the target domain, rather than giving constructive algorithmic tools.[1] (See (Mansour 2009) for more background on domain adaptation.)

Algorithmic Robustness, introduced by (Xu and Mannor 2010) as a different approach to generalization bounds, measures the sensitivity of an algorithm to changes in the training data. Consequently, the robustness level of an algorithm induces a partition of the input domain to multiple regions, and in each region the hypothesis of the robust algorithm has limited variation in its loss. The regions depend both on the input space and the label. The robustness of popular learning algorithms such as SVM was established in (Xu and Mannor 2010).

In this paper we address the domain adaptation problem by using the properties of robust algorithms. The main contributions of this paper are a new generalization bound and related classification and regression algorithms for domain adaptation. The generalization bound applies to the class

*This research was supported in part by the Google Interuniversity center for Electronic Markets and Auctions, by a grant from the Israel Science Foundation, by a grant from United States-Israel Binational Science Foundation (BSF), and by a grant from the Israeli Ministry of Science (MoS).

[1] Reweighing algorithms that optimize the discrepancy distance were presented in (Mansour, Mohri, and Rostamizadeh 2009a). However, in general, minimization of discrepancy distance does not insure an improved performance, since the reweighing might result in overfitting.

of robust algorithms. We also introduce λ-*shift*, a measure that encapsulates prior knowledge regarding the similarity of source and target domain distributions.

The most important property of robustness that we utilize is the limited variation in the loss within each region. This implies that a robust algorithm would have a limited loss variation within each region. Since the overall expected loss is an average of the losses in each region, bounding the loss in a region is our main tool to derive generalization bounds. The main difficulty that we encounter is that the regions guaranteed by the robustness depend on the label, which is not observable for the target distribution sample. We use the λ-shift to overcome this difficulty and derive parameterized generalization bounds for domain adaptation. Two interesting extreme cases are the pessimistic case (assuming that there is no relationship between the probability over labels across the source and target distributions) and the optimistic case (assuming that the probability over labels in the source and target distributions is identical, for any region of the input domain). This will lead us in the former case to pessimistic bounds, where we will use the worse case loss (over the labels) for each region, and to optimistic bounds in the latter.

From the algorithmic perspective we develop SVM variants for binary classification and regression domain adaptation algorithms. The algorithms are formulated as convex optimization programs where the optimized term is based on the generalization bound and the constraints are set to match the λ-shift level assumed. Specifically, the optimized term includes a weighted average (by the target domain distribution) of the bound on the loss in each region, and the constraints on the primal variables (the losses in each region) are the worst case average errors in the regions given the source domain empirical errors and the assumed λ-shift level. Finally, we use the dual representation of the convex optimization program to offer a reweighing interpretation of the resulting robust domain adaptation algorithms.

Model and Preliminaries

Let X be the input space and $f\colon X \to Y$ be the unknown target function (where the label set Y is $\{-1, 1\}$ in case of binary classification and a finite set $\{y_1, \ldots y_r\}$ otherwise). We also allow nondeterministic labeling. In that case the labeling is represented by the conditional probability of a label $y \in Y$ given $x \in X$. Let the input-label space be $Z = X \times Y$. Let H be the hypothesis class used to learn f.

In the domain adaptation setting we have two distributions. The source distribution is Q, from which we have access to labeled examples. The target distribution is P, from which we have only unlabeled examples. We denote by $S = \{(x_i, f(x_i))\}_{i=1..m}$ the set of m labeled sample examples drawn i.i.d. from Q. We denote by $T = \{t_j\}_{j=1..n}$ the set of n unlabeled examples drawn i.i.d. from P. Given $C \subset Z$ we define $S(C) = \frac{1}{m}|\{s \in S \cap C\}|$.

Let $l : Y \times Y \to [0, M]$ be a bounded non-negative loss function. We define the pointwise loss of a function h with respect to a labeled sample $z = (x, y)$

$$l(h, z) \triangleq l(h(x), y),$$

We also define the expected loss (w.r.t a probability distribution D) of a hypothesis $h \in H$ with respect to f,

$$\mathcal{L}_D(h, f) \triangleq E_{x\sim D}[l(h(x), f(x))],$$

for the case of deterministic labeling. When it is clear from the context we omit f and write $\mathcal{L}_D(h)$. Similarly, we define

$$\mathcal{L}_D(h) \triangleq E_{(x,y)\sim D}[l(h(x), y)],$$

for the case of nondeterministic labeling.

Now, for the distribution induced by a finite sample set $S \subset Z$ we have,

$$\mathcal{L}_S(h) \triangleq \frac{1}{|S|}\sum_{s\in S} l(h, s).$$

An *adaptation learning algorithm* uses the labeled sample set S (sampled from the source distribution Q) and the unlabeled sample set T (sampled from the target distribution P), to return a hypothesis $h \in H$. In the domain adaptation problem we are interested in the loss $\mathcal{L}_P(h)$ over the target distribution P.

Algorithmic Robustness

The robustness level of an algorithm A was introduced in (Xu and Mannor 2010) as a measure of its sensitivity to changes in the training data. Specifically, an algorithm A is $(K, \epsilon(S))$ robust if there is a partition of the input-label space $Z = X \times Y$ to K subsets such that the loss of the learned hypothesis h_S has $\epsilon(S)$-bounded variation in every region of the partition that contains a training sample:

Definition 1. *(Xu and Mannor 2010) Algorithm A is $(K, \epsilon(S))$-robust if Z can be partitioned to K disjoint sets $\{C_k\}_{i=1}^K$ such that $\forall s \in S$, for any $s, z \in C_k$, $|l(h_S, s) - l(h_S, z)| \leq \epsilon(S)$.*

Intuitively, since the error of h_S, the output of a $(K, \epsilon(S))$-robust algorithm, has $\epsilon(S)$ variation within each C_k, the empirical error of h_S is a good approximation for the expected error of h_S. Therefore, a robust algorithm that minimizes empirical error is expected to generalize well. Indeed, (Xu and Mannor 2010) prove this precise result, and bound the difference between the empirical error and the expected error of $(K, \epsilon(S))$-robust algorithms:

Theorem 1. *(Xu and Mannor 2010) If A is a $(K, \epsilon(S))$-robust algorithm then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$|\mathcal{L}_Q(h_S) - \mathcal{L}_S(h_S)| \leq \epsilon(S) + M\sqrt{\frac{2K \ln 2 + 2 \ln \frac{1}{\delta}}{|S|}}$$

Note the dependence of ϵ on the sample set S.[2] Indeed, the SVM algorithm, explored in our paper, is (K, ϵ)-robust, where ϵ does not depend on the training set S but only on its size m. In what follows, given a (K, ϵ)-robust algorithm, we assume that the associated partition of Z to K regions is of the following form: $Z = \cup_{i,j} X_i \times Y_j$, where the input space

[2]The parameter ϵ may also depend on K, and (Xu and Mannor 2010) provide a uniform bound for all K.

and output space partitions are $X = \cup_{i=1}^{K_x} X_i$, $Y = \cup_{j=1}^{K_y} Y_j$, and $K = K_x K_y$. This partition implies that the output hypothesis h_S of a (K, ϵ)-robust algorithm has at most an ϵ variation in the loss in each region $C_k = X_i \times Y_j$.

λ-shift

Our main goal is to use the notion of robustness to overcome key difficulties in the domain adaptation setting. The most important difficultly is that we would like to learn with respect to a distribution P, from which we have only unlabeled samples, while the labeled samples are given with respect to a distribution Q.

The notion of robustness would guarantee that in every region $X_i \times Y_j$ the loss of the algorithm would be similar, up to ϵ, regardless of the distribution (source, or target) inside the region. However, a main difficulty still remains since the regions depend on the (unavailable) label of the target function. Therefore, our strategy is to consider the conditional distribution of the label in a given region X_i and the relation to its sampled value over the given labeled sample S. For a distribution σ over $Y = \{y_1, \ldots y_r\}$ (where $r = K_y$, the number of output labels) we denote the probability of y_v by σ^v and the total probability of the other labels by $\sigma^{-v} = 1 - \sigma^v$. We start with a definition of the λ-shift of a given distribution $\sigma \in \Delta_Y$:

Definition 2. *$\rho \in \Delta_Y$ is λ-shift w.r.t. to $\sigma \in \Delta_Y$, denoted $\rho \in \lambda(\sigma)$, if for all $y_v \in Y$ we have $\rho^v \leq \sigma^v + \lambda\sigma^{-v}$ and $\rho^v \geq \sigma^v(1-\lambda)$. If for some v we have $\rho^v = \sigma^v + \lambda\sigma^{-v}$ we say that ρ is strict-λ-shift w.r.t to σ*

A λ-shift therefore restricts the change of the probability of a label - the shift may be at most a λ portion of the probability of the other labels (in case of increase) or of the probability of the label (in case of decrease). To simplify notation, for $\rho \in \lambda(\sigma)$ we denote the upper bound of the probability ρ^v of a label y_v by $\overline{\lambda}^v(\sigma) \triangleq \sigma^v + \lambda(1-\sigma^v)$, and the lower bound on ρ^v by $\underline{\lambda}^v(\sigma) \triangleq \sigma^v(1-\lambda))$

For a non-negative function $l : Y \to \mathbb{R}_+$ we now consider its maximal possible average as a result of a λ-shift:

Definition 3.

$$E_{\lambda,\sigma}(l) \triangleq \max_{\rho \in \lambda(\sigma)} E_\rho l(y)$$

Since the maximum is achieved when ρ is strict-λ-shift to the label y_v of maximal value of l, we have the following:

$$E_{\lambda,\sigma}(l) = \max_v \{l(y_v)\overline{\lambda}^v(\sigma) + \sum_{v' \neq v} l(y_{v'})\underline{\lambda}^{v'}(\sigma)\}$$

Note that for the special case of no restriction (i.e. $1-$shift) we have $E_{1,\sigma}(l) = \max_j\{l(y_j)\}$ and for the special case of total restriction (i.e. $0-$shift) we have $E_{0,\sigma}(l) = E_\sigma(l)$.

To apply the above definitions to the domain adaptation problem first note that the labeled sample S induces in every region X_i a distribution σ_i on the labels: $\sigma_i^v \triangleq \frac{|S_{i,v}|}{|S_i|}$, where $|S_{i,v}|$ is the number of samples labeled y_v in region X_i and $|S_i|$ is the total number of samples in region X_i. Now, we say that the target distribution P is λ-shift of the source distribution Q w.r.t. a partition of the input space X, if in every region X_i the conditional target distribution on the labels $P(y|x \in X_i)$ is λ-shift w.r.t. the conditional source distribution on the labels $Q(y|x \in X_i)$.

We define for each region X_i a function that given a hypothesis h maps every possible label y_v to its maximal sampled empirical loss:

Definition 4.

$$l_i(h, y_v) \triangleq \begin{cases} \max_{s \in S \cap X_i \times y_v} l(h,s) & \text{if } S \cap X_i \times y_v \neq \phi \\ M & \text{otherwise} , \end{cases}$$

Now, for a fixed h, viewing $l_i(h, y)$ as a function of the label y (denote $l_i(h, y_v)$ by l_i^v) and restricting the target distribution in each region X_i to be λ-shift of the empirical σ_i we get that the average loss in region X_i is bounded by $E_{\lambda,\sigma_i}(l_i)$. Specifically, we bound the maximal average loss of a hypothesis h under the λ-shift assumption in region X_i, denoted $l_S^\lambda(h, X_i)$, by

$$l_S^\lambda(h, X_i) \leq \max_v \{l_i^v \overline{\lambda}^v(\sigma_i) + \sum_{v' \neq v} l_i^{v'} \underline{\lambda}^{v'}(\sigma_i)\} \quad (1)$$

Note that a distribution P can be a 0-shift of Q, even if they have disjoint support. What will be important for us is that due to the robustness the loss of the algorithm in any region $X_i \times Y_v$ will be almost the same. Therefore, the major issue would be how to weigh the losses w.r.t the different labels. The λ-shift captures this issue very nicely. Assuming $\lambda = 1$ may be interpreted as a pessimistic assumption, where there is no restriction on the weights of the labels. Assuming $\lambda = 0$ represents an optimistic assumption for which in every region X_i the target distribution assigns the same probability to the samples as the source distribution. In general a $\lambda \in (0, 1)$ represent a tradeoff between the two extremes.

Adaptation Bounds using Robustness

We now prove the following generalization bound for $\mathcal{L}_P(h_S)$, where h_S is the output hypothesis of a (K, ϵ)-robust learning algorithm A which is given a set of labeled samples S and a set of unlabeled samples T of size n.

Theorem 2. *For a (K, ϵ)-robust algorithm A and the related partition of $Z = X \times Y$, if P is λ-shift of Q w.r.t. the partition of X then $\forall \delta > 0$, with probability at least $1 - \delta$, $\forall h \in H$:*

$$\mathcal{L}_P(h) \leq \epsilon + M\sqrt{\frac{2K \ln 2 + 2 \ln \frac{1}{\delta}}{n}} + \sum_{i=1}^{K_x} T(X_i) l_S^\lambda(h, X_i) \quad (2)$$

Proof. The loss of h w.r.t. P is,

$$\mathcal{L}_P(h) = \sum_{k=1}^{K} (P(C_k) - T(C_k))\mathcal{L}_{P|C_k}(h) + \sum_{i=1}^{K} T(C_k) l_{P|C_k}(h)$$

Now, for the second sum above we have

$$\sum_{i=1}^{K} T(C_k) l_{P|C_k}(h) = \sum_{i=1}^{K_x}\sum_{j=1}^{K_y} T(X_i \times Y_j)\mathcal{L}_{P|X_i\times Y_j}(h) = \sum_{i=1}^{K_x} T(X_i)\sum_{j=1}^{K_y} T(Y_j|X_i)\mathcal{L}_{P|X_i\times Y_j}(h)$$

By the robustness property, the loss of h in any region $X_i \times Y_j$ is at most ϵ away from the sampled loss at that region, so we may replace $\mathcal{L}_{P|X_i\times Y_j}(h)$ above with $\mathcal{L}_{T|X_i\times Y_j}(h) + \epsilon$. Also, since P is λ-shift of Q w.r.t. the given partition of X, in every region X_i we have that with probability at least $1-\delta$ the empirical target sample T is $(\lambda+\epsilon)$-shift of the empirical source sample S (for a sample size that depends polynomialy on $\frac{1}{\epsilon}$ and $\log\frac{1}{\delta}$). We therefore get

$$\sum_{i=1}^{K} T(C_k) l_{P|C_k}(h) \le \sum_{i=1}^{K_x} T(X_i) l_S^{\lambda}(h, X_i) + \epsilon$$

Finally, from the bounded loss property we have $\mathcal{L}_{P|C_k}(h) \le M$. Furthermore, as T is sampled from P, by the Bretagnolle-Huber-Carol inequality (as in the proof of Theorem 3 in (Xu and Mannor 2010)) we have that with probability $> 1-\delta$,

$$\sum_{i=1}^{K} |P(C_k) - T(C_k)| \le \sqrt{\frac{2K\ln 2 + 2\ln\frac{1}{\delta}}{n}},$$

which completes the proof. □

Note that although the target sample probability $T(X_i \times y_j)$ of a label y_j in a region X_i is not available, given the hypothesis h and the partition $\{X_i\}_{i=1}^{K_x}$, the last term of the bound $\sum_{i=1}^{K_x} T(X_i) l_S^{\lambda}(h, X_i)$ can be evaluated from the sample sets S and T.

Robust Domain Adaptation SVM for Classification

We consider the classification problem for which the label set $Y = \{1, -1\}$. To simplify notation we set $S_i^+ = S_{i,1}$, $S_i^- = S_{i,-1}$, and $\sigma_i = \sigma_i^1$ the empirical probability of label 1 in region X_i. Using the notation $l_i^+ = l_i(h, 1)$ and $l_i^- = l_i(h, -1)$, the bound (1) on $l_S^{\lambda}(h, X_i)$ for the general case is $\max\{l_i^+(\sigma_i + \lambda(1-\sigma_i)) + l_i^-(1-\sigma_i)(1-\lambda), l_i^+\sigma_i(1-\lambda) + l_i^-((1-\sigma_i)+\lambda\sigma_i)\}$, for the optimistic case $l_i^+\sigma_i + l_i^-(1-\sigma_i)$, and for the pessimistic case $\max\{l_i^+, l_i^-\}$.

Robustness of SVM (See (Xu and Mannor 2010)) implies the existence of a partition $X = \cup_{i=1}^{K} X_i$ for which (2) holds. Given the labeled sample set S and the unlabeled set T, our algorithms selects a hyperplane $h \in H$ that minimizes the generalization bound with an additional appropriate regularization term. We present a robust adaptation algorithm, a general scheme for the λ-shift case in which we assume that the target distribution is λ-shift of the source distribution w.r.t. the partition of X. We then consider two special cases: an *optimistic* variation in which we assume that in every region X_i the probability of each label is the same in the source and target distributions (i.e., 0-shift), and a *pessimistic* variation in which no relationship is assumed between the probability of the labels in the source and target distributions (i.e., 1-shift). We also use the notation $T_i = T(X_i)$ for the T-sampled probability of region X_i.

Note that robustness of SVM implies that $l(h, s)$ varies at most ϵ over $s \in S_i^+$ (and similarly over $s \in S_i^-$). For SVM we use the hinge loss, $l(h,(x,y)) \triangleq \max\{0, 1-yh(x)\}$. For a separating hyperplane $h_{w,b}(x) = w^t x + b$ we have $l(h_{w,b},(x,y)) = \max\{0, 1-y(w^t x+b)\}$.

λ-shift SVM Adaptation

We assume that for some given $\lambda \in [0,1]$, the target distribution P is λ-shift of the source distribution Q w.r.t the partition of the domain X. We define a quadratic optimization program that finds the best separating hyperplane $h_{w,b} = w^t x + b$ in the sense that the related set of losses l_i (the primal variables, together with w and b) minimizes the worst case bound (2)[3]. In addition to the usual SVM constraints on the losses $l_i \ge l(h_{w,b}, s)$ for each sample $s \in S$ (where $l(\cdot,\cdot)$ is the hinge loss), we want to constrain the losses to satisfy $l_i \ge l_S^{\lambda}(h, X_i)$ for each region X_i (thereby minimizing l_i implies that we minimize $l_S^{\lambda}(h, X_i)$). We achieve the latter condition by using a lower bound on l_i which upper bounds $l_S^{\lambda}(h, X_i)$. Using a tradeoff parameter C results in the following convex quadratic program:

$$\min_{w,b,l_1,l_2,\ldots,l_K} C\sum_{i=1}^{K} T_i l_i + \frac{1}{2}\|w\|^2 \tag{3}$$

subject to

$$l_i^+ \ge 1 - (w^t x_j + b) \quad (x_j, 1) \in S_i^+ \tag{4}$$

$$l_i^- \ge 1 + (w^t x_j + b) \quad (x_j, -1) \in S_i^- \tag{5}$$

$$l_i \ge l_i^+(\sigma_i + \lambda(1-\sigma_i)) + l_i^-(1-\sigma_i)(1-\lambda) \tag{6}$$

$$l_i \ge l_i^+\sigma_i(1-\lambda) + l_i^-((1-\sigma_i) + \lambda\sigma_i) \tag{7}$$

$$l_i \ge 0,\ \ l_i^+ \ge 0,\ \ l_i^- \ge 0 \tag{8}$$

Note the first two constraints: for each sample $(x_j, y_j) \in S,\ \ j = 1\ldots m$ we have a constraint regarding one of the two primal variables l_i^+ or l_i^- (depending on the value of y_j) where i is the index of the region X_i to which x_j belongs. The other constraints are for each region $i = 1\ldots K$.

To find the dual representation of this problem we introduce the dual variables $\alpha_1, \ldots, \alpha_m,\ \ \beta_1^+, \ldots, \beta_K^+$,$\beta_1^-, \ldots, \beta_K^-$, $r_1 \ldots, r_K$, $s_1^+, \ldots, s_K^+$, and $s_1^-, \ldots, s_K^-$ The variables α_j pertain to the first or second constraint above

[3] Actually, minimize the last term of (2), which is the only part of the bound that depends on the hypothesis h

depending on the label y_j. The variables β_i^+ and β_i^- pertain to the third and fourth constraint respectively, the variables r_i, s_i^+ and s_i^- pertain to the last constraint. The Lagrangian is,

$$\begin{aligned}
L(w,b,l,\alpha,\beta,r) &= C\sum_{i=1}^{K} T_i l_i + \frac{1}{2}\|w\|^2 \\
&+ \sum_{(x_j,1)\in S_i^+} \alpha_j(1-(w^t x_j+b)-l_i^+) \\
&+ \sum_{(x_j,-1)\in S_i^-} \alpha_j(1+(w^t x_j+b)-l_i^-) \\
&+ \sum_{i=1}^{K} \beta_i^+(l_i^+(\sigma_i+\lambda(1-\sigma_i)) \\
&\quad + l_i^-((1-\sigma_i)-\lambda(1-\sigma_i))-l_i) \\
&+ \sum_{i=1}^{K} \beta_i^-(l_i^+(\sigma_i-\lambda\sigma_i)+l_i^-((1-\sigma_i)+\lambda\sigma_i)-l_i) \\
&- \sum_{i=1}^{K} r_i l_i - \sum_{i=1}^{K} s_i^+ l_i^+ - \sum_{i=1}^{K} r_i^- l_i^-
\end{aligned}$$

Applying the KKT conditions, and simplifying, we get the following dual program:

$$\max_{\alpha_1\ldots\alpha_m} \{\sum_{j=1}^{m}\alpha_j - \frac{1}{2}\|\sum_{j=1}^{m}\alpha_j y_j x_j\|^2\} \tag{9}$$

subject to

$$A_i^+ \le (\sigma_i+\lambda(1-\sigma_i))CT_i \qquad i=1..K \tag{10}$$
$$A_i^- \le (1-\sigma_i(1-\lambda))CT_i \qquad i=1..K \tag{11}$$
$$\sum_{j=1}^{m} y_j\alpha_j = 0 \tag{12}$$
$$\alpha_j \ge 0 \qquad j=1..m \tag{13}$$
$$A_i^+ = \sum_{x_j\in S_i^+}\alpha_j,\ \ A_i^- = \sum_{x_j\in S_i^-}\alpha_j, \qquad i=1..K \tag{14}$$

The primal solution is related to the dual solution by $w = \sum_{j=1}^{m}\alpha_j y_j x_j$, and b is recovered from primal constraints corresponding to dual variables satisfying $A_i^+ + A_i^- < CT_i$. The conditions of the dual program may be interpreted as reweighing of the samples of S. The constraints above imply that $A_i^+ + A_i^- \le CT_i$. Therefore, the total weight of the samples in region X_i is bounded (up to a tradeoff parameter C) by the weight of region X_i as sampled from the target distribution T. Furthermore, in this general case, within each region X_i the total weights of positive labeled samples A_i^+ (or total weight of negative labeled samples A_i^-), is at most a λ-shift of the empirical positive (or negative, respectively) weight of the region.

We now proceed to consider the two special cases, the optimistic case ($\lambda = 0$) and the pessimistic case ($\lambda = 1$).

Optimistic SVM Adaptation

In this variation we assume that P is 0-shift of Q [4]. Setting $\lambda = 0$ in (4) - (8) we get a slightly simplified primal problem whose dual is (10) - (14) with λ set to 0.

For a reweighing interpretation of the dual variables α_j (pertaining to the sample (x_j, y_j) in the primal solution $w = \sum_{j=1}^{m}\alpha_j y_j x_j$) note that at most σ_i portion of the weight allocated to region X_i is allocated to positive samples $(x_j, 1)$ and at most $1-\sigma_i$ portion of the weight is allocated to negative samples $(x_j, -1)$. Note that this may differ from the naive reweighting approach that assigns the weight $\alpha_j = C\frac{|T_i|}{|S_i|}$ to every sample $(x_j, y_j) \in S_i$. This is because the naive reweighting satisfies (10) and (11) with equality, and is not restricted by (12).

Pessimistic SVM Adaptation

In this variation we make no assumptions on P (i.e., P is 1-shift of Q). Setting $\lambda = 1$ simplifies (4) - (8) and the resulting dual program is (10) - (14) with λ set to 1:

$$\max_{\alpha_1,\ldots,\alpha_m} \{\sum_{j=1}^{m}\alpha_j - \frac{1}{2}\|\sum_{j=1}^{m}\alpha_j y_j x_j\|^2\} \tag{15}$$

subject to

$$A_i = \sum_{x_j\in S_i}\alpha_j \le CT_i \qquad i=1..K \tag{16}$$
$$\sum_{j=1}^{m} y_j\alpha_j = 0 \tag{17}$$
$$\alpha_j \ge 0 \qquad j=1..m \tag{18}$$

Again, the primal solution is related to the dual solution by $w = \sum_{j=1}^{m}\alpha_j y_j x_j$ and the dual variables may be interpreted as reweighing of the samples of S: The weight A_i, the total weight of the samples in region X_i, is bounded by the weight of region X_i in the set T. In this *pessimistic* variation there is no restriction on A_i^+ or A_i^- and the weight of region X_i is fully allocated to the region samples with the highest loss. This is natural since the support of the target distribution in every region might only include points of such worst-case loss.

Robust Domain Adaptation for Regression

In the regression setting the label set Y and the domain X are each a bounded convex subset of $\mathbb{R}$. The classification loss at a sample $z_j = (x_j, y_j)$ is $l(h, z_j) = (h(x_j) - y_j)^2$. Robustness of regression algorithms (e.g. Lasso, see (Xu and Mannor 2010)) implies that we may assume a partition $Y = \cup_{v=1}^{K_y} Y_v$ of the label range for which (2) holds, and we define the sample subsets $S_i^v \triangleq S \cap X_i \times Y_v$ and

[4] Note that this is not equivalent to assuming that $Q = P$. The source and target distributions might substantially differ and still have the same probability in each region X_i, and even more importantly, they can arbitrarily differ in the probability that they assign to different regions X_i.

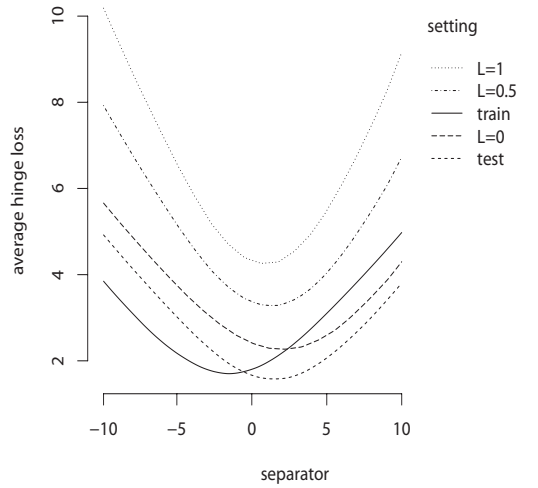

Figure 1: Separators performance w.r.t. experiment data

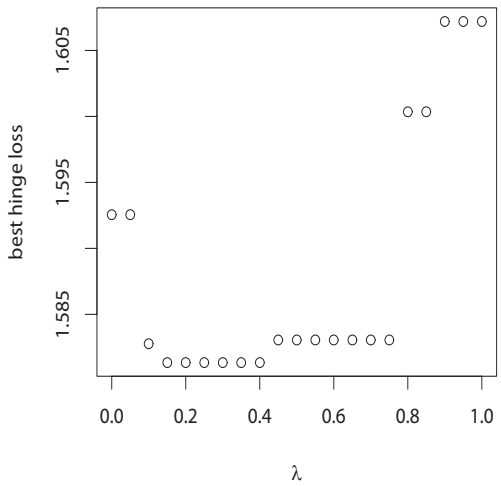

Figure 2: Performance of λ-shift SVM optimal separator

$S^v \triangleq S \cap X \times Y_v$. As before, we assume that the target distribution is λ-shift of the empirical distribution in every region X_i. We use the notation σ_i^v for the empirical probability (in sample set S) of label v in region X_i, and $l_i^v = l_i(h, v)$ for the maximal loss of hypothesis h in $X_i \times Y_v$. To solve the domain adaptation problem in this setting, in addition to the usual constraints on the losses $l_i^v \geq l(h_{w,b}, s)$ for each sample $s \in S_i^v$, we want to constrain the losses to satisfy $l_i \geq l_S^\lambda(h, X_i)$ for each region X_i (thereby minimizing l_i implies that we minimize $l_S^\lambda(h, X_i)$). As before, we achieve the latter condition by using a lower bound on l_i which upper bounds $l_S^\lambda(h, X_i)$ by (1). The algorithm selects among all linear functions $h_{w,b}(x) = w^t x + b$ the one that minimizes the generalization bound (2) with an additional appropriate regularization term. We assume that for each region X_i the target probability distribution on the labels ρ_i is λ-shift of the empirical distribution σ_i. To simplify nitation we denote the upper bound of ρ_i^v, the probability of label y_v in region X_i by $\overline{\lambda}_i^v \triangleq \overline{\lambda}^v(\sigma_i)$, and the lower bound on ρ_i^v by $\underline{\lambda}_i^v \triangleq \underline{\lambda}^v(\sigma_i)$. Finally, using a tradeoff parameter C results in the following convex quadratic program for Ridge Regression[5]:

$$\min_{w,b,l_1,l_2,\ldots,l_K} C \sum_{i=1}^{K} T_i l_i^2 + \frac{1}{2}\|w\|_2^2 \qquad (19)$$

subject to

$$l_i^v \geq y_j - (w^t x_j + b) \qquad (x_j, y_j) \in S_i^v \qquad (20)$$

$$l_i^v \geq (w^t x_j + b) - y_j \qquad (x_j, y_j) \in S_i^v \qquad (21)$$

$$l_i \geq l_i^v \overline{\lambda}_i^v + \sum_{v' \neq v} l_i^{v'} \underline{\lambda}_i^{v'} \qquad (22)$$

Note that we have a constraint on the primal variable l_i for each $i = 1 \ldots K_X$ and $v = 1 \ldots K_Y$. Note also that for the pessimistic case ($\lambda = 1$) the last constraint above simplifies to $l_i \geq l_i^v$ and for the optimistic case ($\lambda = 0$) the last constraint above simplifies to $l_i \geq \sum_v \sigma_i^v l_i^v$.

To find the dual representation of the Ridge regression problem in the general case case we introduce a dual variables α_j^+ associated with the first constraint above, α_j^- associated with the second, and B_i^v associated with the last one. Setting the partial derivatives (according to the primal variables w, b, l_i^v, and l_i) of the resulting Lagrangian to 0 we get the following relations: $w = \sum_j (\alpha_j^+ - \alpha_j^-) x_j$, $\sum_j \alpha_j^+ = \sum_j \alpha_j^-$, $B_i = 2CT_i l_i$, and $\sum_{(x_j,y_j) \in S_i^v} (\alpha_j^+ + \alpha_j^-) = \lambda_i^v B_i^v + \overline{\lambda}_i^v B_i^{\overline{v}}$, where $B_i = \sum_v B_i^v$ and $B_i^{\overline{v}} = \sum_{v' \neq v} B_i^{v'}$. Using the above relations we get the following dual problem:

$$\max_{\alpha, B} -\frac{1}{2}\|\sum_j (\alpha_j^+ - \alpha_j^-) x_j\| + \sum_j (\alpha_j^+ - \alpha_j^-) y_j - \frac{1}{4C} \sum_i \frac{B_i^2}{T_i}$$

subject to

$$\begin{aligned} \sum_j \alpha_j^+ &= \sum_j \alpha_j^- , \\ \sum_{(x_j,y_j) \in S_i^v} (\alpha_j^+ + \alpha_j^-) &= \lambda_i^v B_i^v + \overline{\lambda}_i^v B_i^{\overline{v}} \\ \alpha_j^+ &\geq 0 \\ \alpha_j^- &\geq 0 \end{aligned}$$

The primal solution is related to the dual solution by $w = \sum_j (\alpha_j^+ - \alpha_j^-) x_j$. Note also that $\alpha_j^+ \alpha_j^- = 0$.

Experiment

To illustrate the ability of performing the domain adaptation task by using our methods we considered a synthetic one dimensional binary classification problem. We run the λ-shift domain adaptation SVM on a synthetic data set containing train and test samples from significantly different domains.

[5] We may similarly solve for Lasso Regression: replacing (19) with $\min_{w,b,l_1,l_2,\ldots,l_K} C \sum_{i=1}^{K} T_i l_i^2 + \|w\|_1$

The experiment confirmed that for several values of λ (not necessarily 0 or 1) the test error of the optimal (with respect to the train set) linear separator may be improved by using the separator returned by our algorithm.

Figure 1 shows the resulting loss levels of the linear separators. The labeled train samples are a mixture of three Gaussians, centered at -5, 0, and 5, producing positive, negative, and positive labels respectively [6]. Standard deviation is 5 for the Gaussians generating the positive labels and 3 for those generating negative labels. In the source domain the probabilities of generating a sample for the first, second, or third Gaussians are 0.4, 0.5, and 0.1, respectively, while in the target domain the probabilities are 0.1, 0.5, and 0.4. The upper curves ($L = 1$, $L = 0.5$, and $L = 0$) correspond to the bounds L_S^λ on the average loss of the separator as computed by our λ-shift SVM for $\lambda = 1$, 0.5, and 0.

Now, the best linear separator for the train set will incur significantly higher loss on the test set. However, the best linear separator produced by a λ-shift SVM (corresponding to the lowest points of each of the three upper curves of figure 1) may be closer to the optimal linear separator of the test set, and therefore perform better in the target domain[7]. Indeed, as figure 2 shows, running the λ-shift SVM (on the same data sets) with λ ranging between 0.2 and 0.4 results in separators having loss that is comparable to the loss of the best test-set separator.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of Representations for Domain Adaptation. *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning* 79(1-2):151–175.

Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Wortman, J. 2007. Learning Bounds for Domain Adaptation. *Advances in Neural Information Processing Systems*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009a. Domain adaptation: Learning bounds and algorithms. In *COLT*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009b. Multiple source adaptation and the renyi divergence. In *UAI*.

Mansour, Y. 2009. Learning and domain adaptation. In *ALT*, 4–6.

Valiant, L. G. 1984. *A theory of the learnable*. ACM Press New York, NY, USA.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York: Wiley-Interscience.

Xu, H., and Mannor, S. 2010. Robustness and generalization. In *COLT*, 503–515.

[6] Note that the positives and negatives are not linearly separable

[7] Note that the precise loss values of the λ-shift loss curve (loss bounds calculated by the λ-shift SVM) are not important, the value of interest is the specific separator that achieves minimal loss on the curve!