Marginal Singularity, and the Benefits of Labels in Covariate-Shift

Samory Kpotufe[*] SAMORY@PRINCETON.EDU
ORFE, Princeton University

Guillaume Martinet[†] GGM2@PRINCETON.EDU
ORFE, Princeton University

Abstract

We present new minimax results that concisely capture the relative benefits of source and target labeled data, under covariate-shift. Namely, we show that, in general classification settings, the benefits of target labels are controlled by a *transfer-exponent* γ that encodes how *singular* Q is locally w.r.t. P, and interestingly allows situations where transfer did not seem possible under previous insights. In fact, our new minimax analysis – in terms of γ – reveals a *continuum of regimes* ranging from situations where target labels have little benefit, to regimes where target labels dramatically improve classification. We then show that a recently proposed semi-supervised procedure can be extended to adapt to unknown γ, and therefore requests target labels only when beneficial, while achieving nearly minimax transfer rates.

Keywords: Transfer learning, covariate-shift, nonparametric classification, nearest-neighbors.

Extended Abstract

Introduction

The goal in transfer learning is to improve prediction on a *target* distribution Q by harnessing labeled data coming from a *source* distribution P. Much of theoretical work in transfer learning concerns understanding the fundamental limitations of transfer, and in particular, proper ways of capturing *relatedness* between source P and target Q. Here we consider the common *covariate-shift* setting for classification, where it is assumed that conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ remain the same, while marginals P_X, Q_X are different but somewhat related.

We consider general nonparametric settings that capture a range of easy to difficult classification under Q, through standard smoothness and noise conditions (see e.g. **?**). Our aim is then to understand which relation between marginals P_X and Q_X control the rates of transfer, and in particular, control the relative benefits between source and target data in achieving low error under Q. A basic intuition, present in previous work, is that transfer is easiest when P assigns sufficient mass to regions of considerable Q-mass. Here, we formalize this intuition through a new *asymmetric* notion, the *transfer-exponent* γ, that parametrizes the behavior of ball-mass ratios $Q_X(B(x,r))/P_X(B(x,r))$ as a function of the radius r, namely, that these ratios behave like $r^{-\gamma}$. The notion of γ can be interpreted, roughly, as capturing how close to *singular* Q is with respect to P, as it shifts mass into regions of low P mass.

. * † Authors are listed in alphabetical order.

. Extended abstract. Full version appears as arXiv:1803.01833v2.

We show the pertinence of our parametrization by establishing tight minimax upper and lower bounds in terms of γ, under standard nonparametric conditions. The notion of γ is thus shown to encode a *continuum of regimes* between easy and hard transfer, and interestingly, reveals situations where transfer is possible (even at fast rates) despite P and Q seeming *unrelated* under previous notions of *relatedness*. As an example, γ remains well defined even when Q is singular w.r.t. P (e.g. Q puts mass on lower-dimensional structures) in which case common notions of density-ratio and information-theoretic divergences (KL or Renyi) fail to exist, and common extensions of total-variation can be too large to characterize transfer.

Finally, we show that a recently proposed semi-supervised procedure can be extended to adapt to unknown γ, and therefore requests target labels only when beneficial, while achieving nearly minimax transfer rates.

Related Work

Many insightful notions of *relatedness* are present in the literature on transfer and related problems.

A first line of work considers refinements of total-variation which encode changes in classification error from P to Q (restricted to a hypothesis class $\mathcal{H}$). The most common such measures are the so-called $d_{\mathcal{A}}$-divergence (**???**) and $\mathcal{Y}$-discrepancy (**???**). These notions are the first to capture – through *differences* in mass over space – the intuition that transfer is easiest when P has sufficient mass in regions of substantial Q-mass. Typical excess-error bounds on classifiers learned from source (and some or no target) data are of the form $o_p(1) + C \cdot \text{divergence}(P, Q)$. In other words, transfer seems impossible when these divergences are large; however, we show that there are ranges of reasonable situations ($0 \leq \gamma < \infty$) where fast transfer is possible even when such divergences are large. Furthermore, while such divergences are symmetric, the notion of γ is not, thus capturing the fact that transfer might be easy from P to Q but not from Q to P.

Another prominent line or work, which has led to many practical procedures, considers so-called density-ratios f_Q/f_P or more generally, Radon-Nikodym derivatives dQ/dP, as a way to capture the similarity between P and Q (**??**). It is often assumed in such work that dQ/dP is bounded, which corresponds to assuming $\gamma = 0$. Typical excess-error bounds are dominated by the estimation rates for dQ/dP (see e.g. rates for α-Hölder dQ/dP, $\alpha \to 0$, in **?**), which unfortunately could be arbitrarily higher than the minimax rates we establish for the boundary case with $\gamma = 0$.

Finally, another line of work instead considers information-theoretic measures such as KL-divergence or Renyi divergence (**??**). In particular, such divergences are closer in spirit to our notion of transfer-exponent γ (viewing γ as roughly characterizing the log of mass-ratios between), but are also undefined in typical scenarios with structured data where Q_X might be singular w.r.t. P_X.

Result Overview

Our first results consider transfer settings where the learner has access to n_P labeled samples drawn from P and n_Q labeled samples drawn from Q, where typically $n_P \gg n_Q$. The label Y is assumed in $\{0, 1\}$, while the input X belongs to a compact metric space $\mathcal{X}$.

We work under common smoothness and low noise conditions, namely, we assume the *regression function* $\eta(x) \doteq E[Y|X = x]$ to be α–Hölder, and also that $Q_X\,(0 < |\eta(X) - 1/2| \leq t) \lesssim t^\beta$ (see e.g. **?**). A *transfer exponent* is then defined, roughly, as any quantity γ that satisfies:

$$\forall x, \forall \text{ small } r, \quad P_X(B(x, r)) \gtrsim Q_X(B(x, r)) \cdot r^\gamma.$$

Two main distributional regimes are considered, which capture the difficulty of vanilla classification under Q_X. The first regime, (DM) (for *doubling measure*), roughly assumes that Q_X behaves like a uniform measure on its support (this is the most common assumption in nonparametric classification, and is sometimes termed the *strong-density assumption*). The second regime, (BCN) (for *bounded covering number*), allows for general Q_X and is most difficult with slower rates. Both regimes introduce a parameter d that might be viewed as a notion of *dimension* of the marginal Q_X.

For exact definitions we refer the reader to the archived version of this work (**?**)).

Our minimax rates are then of the following form.

Theorem 1 (Sketch) *Call $\mathcal{T}_{(DM)}$ (resp. $\mathcal{T}_{(BCN)}$) the class of all the tuples (P, Q) under (DM) (resp. (BCN)) regime. Let $\mathcal{T} \in \{\mathcal{T}_{(DM)}, \mathcal{T}_{(BCN)}\}$, we have then:*

$$\inf_{\hat{h}} \sup_{(P,Q)\in\mathcal{T}} \mathbb{E}[\mathcal{E}_Q(\hat{h})] \asymp \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

where $\mathcal{E}_Q$ represents the excess error, the infimum is taken over all classifiers $\hat{h}$ learned on the data, the expectation is taken w.r.t. the data, $d_0 = 2 + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(DM)}$, and $d_0 = 2 + \beta + d/\alpha$ when $\mathcal{T} = \mathcal{T}_{(BCN)}$.

Our upper-bounds are established with a generic k-NN classifier defined over the combined source and target sample. In particular, our results imply new convergence rates of independent interest for vanilla k-NN under the BCN regime, which complements recent developments on vanilla k-NN (**????**). On the other hand, our lower-bounds are established over any learner with access to both source and target samples, and interestingly, which is also allowed access to infinite unlabeled source and target data (i.e., full knowledge of P_X and Q_X). In other words, the above rates cannot be improved (beyond constants) with access to unlabeled data, which is often an important consideration in practice given the cost of target labels (**??**).

Finally, we address semi-supervised situations where the learner has access to n_Q *unlabeled* target data, along with n_P labeled source data, and is allowed to request (as few as possible) target labels in order to improve classification (**???**). An early theoretical treatment of this can be found in (**?**), but which however considers a transfer setting with fixed marginal but varying conditionals (labeling functions). For our setting of covariate-shift, we build on a recent approach of **?** which constructs so-called k-$2k$ covers, to help limit label requests to regions of low P mass. In this work, we show a strategy for choosing k from data (building on so-called *Lepski's method* (**?**)), so as to nearly attain the above minimax rates with no a priori knowledge of distributional parameters, nor of γ. Furthermore, labeling complexity is shown to be controlled by unknown γ, i.e. the resulting approach requests labels only when *useful*, as controlled by γ and relative sample sizes n_P, n_Q.