ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Weighted Empirical Risk Minimization: Transfer Learning based on Importance Sampling

Robin Vogel[1,2], Mastane Achab[1], Stéphan Clémençon[1] and Charles Tillier[1]

1- Télécom Paris
19 Place Marguerite Perey, 91120 Palaiseau - France

2- IDEMIA
2 Place Samuel de Champlain, 92400 Courbevoie - France

Abstract. We consider statistical learning problems, when the distribution P' of the training observations $Z'_1, \ldots, Z'_n$ differs from the distribution P involved in the risk one seeks to minimize (referred to as the *test distribution*) but is still defined on the same measurable space as P and dominates it. In the unrealistic case where the likelihood ratio $\Phi(z) = dP/dP'(z)$ is known, one may straightforwardly extends the Empirical Risk Minimization (ERM) approach to this specific *transfer learning* setup using the same idea as that behind Importance Sampling, by minimizing a weighted version of the empirical risk functional computed from the 'biased' training data Z'_i with weights $\Phi(Z'_i)$. Although the *importance function* $\Phi(z)$ is generally unknown in practice, we show that, in various situations frequently encountered in practice, it takes a simple form and can be directly estimated from the Z'_i's and some auxiliary information on the statistical population P. By means of linearization techniques, we then prove that the generalization capacity of the approach aforementioned is preserved when plugging the resulting estimates of the $\Phi(Z'_i)$'s into the weighted empirical risk. Beyond these theoretical guarantees, numerical results provide strong empirical evidence of the relevance of the approach promoted in this article.

1 Introduction

Prediction problems are of major importance in statistical learning. The main paradigm of predictive learning is *Empirical Risk Minimization* (ERM in abbreviated form), see *e.g.* [8]. In the standard setup, Z is a random variable (r.v. in short) that takes its values in a feature space $\mathcal{Z}$ with distribution P, Θ is a parameter space and $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is a (measurable) loss function. The risk is then defined by: $\forall \theta \in \Theta$,

$$\mathcal{R}_P(\theta) = \mathbb{E}_P\left[\ell(\theta, Z)\right], \tag{1}$$

and more generally for any measure Q on $\mathcal{Z}$: $\mathcal{R}_Q(\theta) = \int_{\mathcal{Z}} \ell(\theta, z) dQ(z)$. In most practical situations, the distribution P involved in the definition of the risk is unknown and learning is based on the sole observation of an independent and identically distributed (i.i.d.) sample $Z_1, \ldots, Z_n$ drawn from P and the risk (1) must be replaced by an empirical counterpart (or a possibly smoothed/penalized version of it), typically:

$$\widehat{\mathcal{R}}_P(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i) = \mathcal{R}_{\widehat{P}_n}(\theta), \tag{2}$$

where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{Z_i}$ is the empirical measure of P and δ_z denotes the Dirac measure at any point z. The performance of minimizers of (2) can be studied by means

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

of concentration inequalities, quantifying the fluctuations of the maximal deviations $\sup_{\theta\in\Theta} |\widehat{\mathcal{R}}_P(\theta) - \mathcal{R}_P(\theta)|$ under various complexity assumptions for the functional class $\mathcal{F} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$ (*e.g.* VC dimension, metric entropies, Rademacher averages), see [5] for instance. Although the Big Data era features undeniable opportunities for machine-learning solutions, poor control of the data acquisition process puts practicioners at risk of jeopardizing the generalization ability of the algorithms. Bias selection issues in machine-learning are now the subject of much attention in the literature, see *e.g.* [3]. Machine learning algorithms trained with baised training data, *e.g.* in terms of gender or ethnicity, raise concerns about fairness in machine learning, see [14] for further details.

Throughout the article, we consider the case where the i.i.d. sample $Z'_1, \ldots, Z'_n$ available for training is not drawn from P but from another distribution P', with respect to which P is absolutely continuous, and the goal pursued is to set theoretical grounds for the application of ideas behind Importance Sampling (IS in short) methodology to extend the ERM approach to this learning setup. This problem is a very particular case of *Transfer Learning* (see *e.g.* [12]), a research area that encompasses general situations where the information/knowledge one would like to transfer may take a form in the *target* space very different from that in the *source* space.

Weighted ERM (WERM). In this paper, we investigate conditions guaranteeing that values for the parameter θ that nearly minimize (1) can be obtained through minimization of a weighted version of the empirical risk based on the Z'_i's, namely

$$\widetilde{\mathcal{R}}_{w,n}(\theta) = \mathcal{R}_{\widetilde{P}_{w,n}}(\theta), \tag{3}$$

where $\widetilde{P}_{w,n} = (1/n)\sum_{i=1}^n w_i \delta_{Z'_i}$ and $w = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$ is a certain weight vector. Of course, ideal weights w^* are given by the likelihood function $\Phi(z) = (dP/dP')(z)$: $w_i^* = \Phi(Z'_i)$ for $i \in \{1, \ldots, n\}$. In this case, the quantity (3) is obviously an unbiased estimate of the true risk (1), i.e. $\mathbb{E}_{P'}[\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)] = \mathcal{R}_P(\theta)$, and generalization bounds for the $\mathcal{R}_P$-risk excess of minimizers of the empirical risk with ideal weights can be directly established by studying the concentration properties of the empirical process related to the Z'_i's and the class of functions $\{\Phi(\cdot)\ell(\theta, \cdot) : \theta \in \Theta\}$ (see section 2 below). However, the *importance function* Φ is unknown in general, just like distribution P. It is the major purpose of this article to show that, in far from uncommon situations, the (ideal) weights w_i^* can be estimated from the Z'_is combined with auxiliary information on the target population P. Such favorable cases include in particular classification problems where class probabilities in the test stage differ from those in the training step, risk minimization in stratified populations (see [2]), with strata statistically represented in a different manner in the test and training populations, positive-unlabeled learning (PU-learning, see *e.g.* [9]). Learning rate bounds for minimizers of the corresponding risk estimate are proved and, beyond these theoretical guarantees, the performance of the weighted ERM approach is supported by convincing numerical results.

2 Importance Sampling - Risk Minimization with Biased Data

Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the sup norm of any bounded function $h : \mathcal{Z} \to \mathbb{R}$ by $\|h\|_\infty$. We place ourselves in the framework

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

of statistical learning based on biased training data previously introduced. As a first go, we consider the unrealistic situation where the importance function Φ is known, insofar as we shall subsequently develop techniques aiming at mimicking the minimization of the ideally weighted empirical risk

$$\widetilde{\mathcal{R}}_{w^*,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_i^* \ell(\theta, Z_i'), \tag{4}$$

namely the (unbiased) Importance Sampling estimator of (1) based on the instrumental data $Z_1', \ldots, Z_n'$. The following result describes the performance of minimizers $\widetilde{\theta}_n^*$ of (4). Since the goal of this paper is to promote the main ideas of the approach rather than to state results with the highest level of generality due to space limitations, we assume throughout the article for simplicity that ℓ and Φ are both bounded functions. For $\sigma_1, \ldots, \sigma_n$ independent Rademacher random variables (*i.e.* symmetric $\{-1, 1\}$-valued r.v.'s), independent from the Z_i''s, we define the Rademacher average associated to the class of function $\mathcal{F}$ as $R_n'(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{\theta\in\Theta} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \ell(\theta, Z_i')\right|\right]$. This quantity can be bounded by metric entropy methods under appropriate complexity assumptions on the class $\mathcal{F}$, it is for instance of order $O_{\mathbb{P}}(1/\sqrt{n})$ when $\mathcal{F}$ is a VC major class with finite VC dimension, see *e.g.* [4].

Lemma 1. *With probability at least* $1-\delta$*, we have:* $\forall n \geq 1$*,*

$$\mathcal{R}_P(\widetilde{\theta}_n^*) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 4\|\Phi\|_\infty \mathbb{E}\left[R_n'(\mathcal{F})\right] + 2\|\Phi\|_\infty \sup_{(\theta,z)\in\Theta\times\mathcal{Z}} \ell(\theta, z)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Of course, when $P' = P$, we have $\Phi \equiv 1$ and the bound stated above simply describes the performance of standard empirical risk minimizers. Of course, the importance function Φ is generally unknown and must be estimated in practice. As illustrated by the elementary example below (related to binary classification, in the situation where the probability of occurence of a positive instance significantly differs in the training and test stages), in certain statistical learning problems with biased training distribution, Φ takes a simplistic form and can be easily estimated from the Z_i''s combined with auxiliary information on P.

Binary classification with varying class probabilities. The flagship problem in supervised learning corresponds to the simplest situation, where $Z = (X, Y)$, Y being a binary variable valued in $\{-1, +1\}$ say, and the r.v. X takes its values in a measurable space $\mathcal{X}$ and models some information hopefully useful to predict Y. The parameter space Θ is a set $\mathcal{G}$ of measurable mappings (*i.e.* classifiers) $g : \mathcal{X} \to \{-1, +1\}$ and the loss function is given by $\ell(g, (x, y)) = \mathbb{I}\{g(x) \neq y\}$ for all g in $\mathcal{G}$ and any $(x, y) \in \mathcal{X}\times\{-1, +1\}$. The distribution P of the random pair (X, Y) can be either described by X's marginal distribution $\mu(dx)$ and the posterior probability $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ or else by the triplet (p, F_+, F_-) where $p = \mathbb{P}\{Y = +1\}$ and $F_\sigma(dx)$ is X's conditional distribution given $Y = \sigma 1$ with $\sigma \in \{-, +\}$. It is very common that the fraction of positive instances in the training dataset is significantly lower than the rate p expected in the test stage, supposed to be known here. We thus consider the case where the distribution P' of the training data $(X_1', Y_1'), \ldots, (X_n', Y_n')$ is described by the triplet (p', F_+, F_-) with $p' < p$.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

The likelihood function takes the simple following form

$$\Phi(x,y) = \mathbb{I}\{y=+1\}\frac{p}{p'} + \mathbb{I}\{y=-1\}\frac{1-p}{1-p'} \stackrel{def}{=} \phi(y),$$

which reveals that it depends on the label y solely, and the ideally weighted empirical risk process is

$$\widetilde{\mathcal{R}}_{w^*,n}(g) = \frac{p}{p'}\frac{1}{n}\sum_{i:Y_i'=1}\mathbb{I}\{g(X_i')=-1\} + \frac{1-p}{1-p'}\frac{1}{n}\sum_{i:Y_i'=-1}\mathbb{I}\{g(X_i')=+1\}. \tag{5}$$

In general the theoretical rate p' is unknown and one replaces (5) with

$$\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) = \frac{p}{n'_+}\sum_{i:Y_i'=1}\mathbb{I}\{g(X_i')=-1\} + \frac{1-p}{n'_-}\sum_{i:Y_i'=-1}\mathbb{I}\{g(X_i')=+1\}, \tag{6}$$

where $n'_+ = \sum_{i=1}^n \mathbb{I}\{Y_i'=+1\} = n - n'_-$, $\widehat{w}_i^* = \widehat{\phi}(Y_i')$ and $\widehat{\phi}(y) = \mathbb{I}\{y=+1\}np/n'_+ + \mathbb{I}\{y=-1\}n(1-p)/n'_-$. The stochastic process above is not a standard empirical process but a collection of sums of two ratios of basic averages.

Theorem 1. *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $p' \in (\varepsilon,\ 1-\varepsilon)$. Let $\widetilde{g}_n$ be any minimizer of $\widetilde{\mathcal{R}}_{\widehat{w}^*,n}$ over class $\mathcal{G}$. For any $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$\mathcal{R}_P(\widetilde{g}_n) - \inf_{g\in\mathcal{G}}\mathcal{R}_P(g) \leq \frac{2\max(p,1-p)}{\varepsilon}\left(2\mathbb{E}[R_n'(\mathcal{G})] + \sqrt{\frac{2\log(2/\delta)}{n}}\right) + \frac{4}{\varepsilon^2}\sqrt{\frac{\log(4/\delta)}{2n}},$$

as soon as $n \geq 2\log(4/\delta)/\varepsilon^2$; where $R_n'(\mathcal{G}) = (1/n)\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}|\sum_{i=1}^n \sigma_i\mathbb{I}\{g(X_i')\neq Y_i'\}|]$.

We now briefly introduce more general transfer learning problems where WERM can be applied by following the same steps as above: i) express the likelihood ratio $\Phi = dP/dP'$, ii) approximate Φ by $\widehat{\Phi}$ based on samples from P' combined with side information on P, and iii) apply WERM with weights given by $\widehat{\Phi}$. We only describe step i) below, due to space limitations.

Learning from biased stratified data. Consider a general mixture model, where distributions P and P' are stratified over $K \geq 1$ strata. Namely, $Z = (X,S)$ and $Z' = (X',S')$ with auxiliary random variables S and S' (the strata) valued in $\{1,\ \ldots,\ K\}$. We place ourselves in a *stratum-shift* context, assuming that the conditional distribution of X given $S = k$ is the same as that of X' given $S' = k$, denoted by $F_k(dx)$, for any $k \in \{1,\ \ldots,\ K\}$. However, stratum probabilities $p_k = \mathbb{P}(S=k)$ and $p'_k = \mathbb{P}(S'=k)$ may possibly be different. In this setup, the likelihood function depends only on the strata and can be expressed in a very simple form, as follows:

$$\frac{dP}{dP'}(x,s) = \sum_{k=1}^K \mathbb{I}\{s=k\}\frac{p_k}{p'_k} \stackrel{def}{=} \phi(s).$$

Positive-Unlabeled Learning. Relaxing the *stratum-shift* assumption made in the previous subsection, the importance function becomes more complex and writes: $\Phi(x,s) =$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

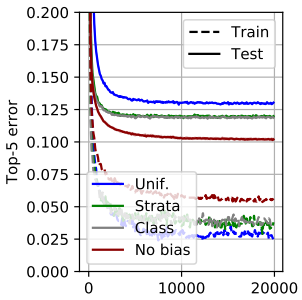
$\frac{dP}{dP'}(x,s) = \sum_{k=1}^{K} \mathbb{I}\{s=k\}\frac{p_k}{p'_k}\frac{dF_k}{dF'_k}(x)$, where F_k and F'_k are respectively the conditional distributions of X given $S = k$ and of X' given $S' = k$. The Positive-Unlabeled (PU) learning problem, which has recently been the subject of much attention (see *e.g.* [9], [**?**], [**?**]), provides a typical example of this situation. The testing and training distributions P and P' are respectively described by the triplets (p, F_+, F_-) and (q, F_+, F), where $F = pF_+ + (1-p)F_-$ is the marginal distribution of X. Hence, the objective pursued is to solve a binary classification task, based on the sole observation of a training sample pooling data with positive labels and unlabeled data, q denoting the theoretical fraction of positive data among the dataset. As noticed in [9] (see also [**?**], [**?**]), the likelihood/importance function can be expressed in a simple manner, as follows:

$$\forall (x,y) \in \mathcal{X}\times\{-1,\ +1\},\quad \Phi(x,y) = \frac{p}{q}\mathbb{I}\{y=+1\} + \frac{1}{1-q}\mathbb{I}\{y=-1\} - \frac{p}{1-q}\frac{dF_+}{dF}(x)\mathbb{I}\{y=-1\}.$$

3 Numerical Experiments

This section illustrates the impact of reweighting by the likelihood ratio on classification performances, as a special case of the general strategy presented in Section 2.

We present an experiment that leverages the structure of the well-known ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset to illustrate the *learning from biased stratified data* setting. The code of the experiments can be found at `https://tinyurl.com/qwthb9q` (Google drive). The challenge consists in learning a classifier from 1.3 million training images spread out over 1,000 classes. Performance is evaluated using the validation dataset of 50,000 images of ILSVRC as our test dataset. ImageNet is an image database organized according to the WordNet hierarchy, which groups nouns in sets of related words called synsets. In that context, images are examples of very precise nouns, e.g. *flamingo*, which are contained in a larger synset, e.g. *bird*. The impact of reweighting in presence of strata bias is illustrated on the ILSVRC classification problem with broad significance synsets for strata. To do this, we encode the data using deep neural networks. Specifically our encoding is



Dynamics for the top-5 error.

Reweighting	miss rate	top-5 error
Unif. $\widehat{\Phi} = 1$	0.344	0.130
Strata $\widehat{\Phi}$	**0.329**	**0.120**
Class $\widehat{\Phi}$	0.328	0.119
No bias	0.297	0.102

Table of results for a linear model.

Fig. 1: Results of the strata reweighting experiment with ImageNet for a linear model.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

the flattened output of the last convolutional layer of the network ResNet50 introduced in [11]. It was trained for classification on the training dataset of ILSVRC. A total of 33 strata are derived from a list of high-level categories provided by ImageNet at `http://www.image-net.org/about-stats`. By default, strata probabilities p_k and p'_k for $1 \leq k \leq K$ are equivalent between training and testing datasets, meaning that reweighting by Φ would have little to no effect. Leveraging the abundance of train data, we discarded elements of the train set in each strata in a way that induces a parameterized strata biais.

We report better performance when reweighting using the strata information, compared to the case where the strata information is ignored (Unif.). For comparison, we added two reference experiments: one which reweights the train instances by the class probabilities (Class), which we do not know in a stratified population experiment, and one with more data and no strata bias (No bias) because it uses all of the ILSVRC train data. Tests with a multilayer perceptron (MLP) did not perform as well, see the code.

References

[1] J. Bekker and J. Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. *CoRR*.

[2] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, 2016.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : a survey of some recent advances. *ESAIM: Probability and Statistics*, 2005.

[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[5] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[6] M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *NIPS*, 2014.

[7] M. C. du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML*, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, 2015.

[9] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NIPS*, 2017.

[10] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[11] M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: A flexible approach for fair classification. *JMLR*, 2019.