Covariate Shift Adaptation in High-Dimensional and Divergent Distributions

Felipe Maia Polo[1,2] and Renato Vicente[1,3]

[1]University of São Paulo, Brazil

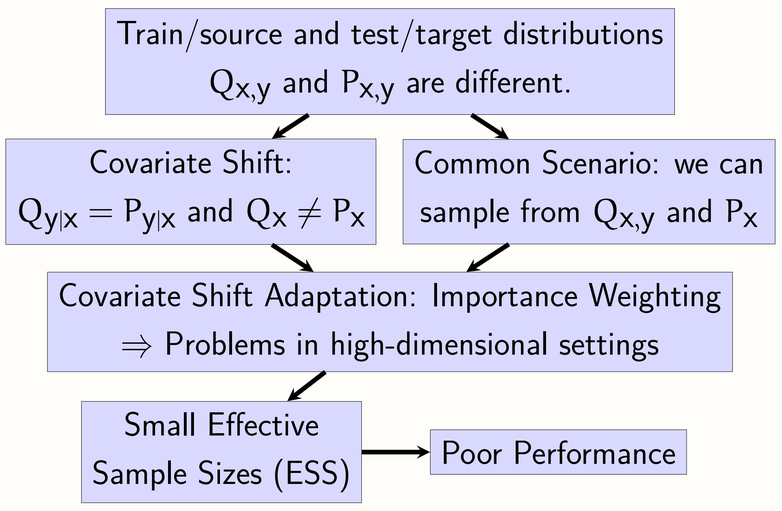
[2]Advanced Institute for Artificial Intelligence (AI2), Brazil

[3]Experian DataLab LatAm, Brazil

Full paper: `arxiv.org/abs/2010.01184`

Introduction

The Problem



Main Contributions

1. A new theoretical connection between ESS, high-dimensional settings and generalization bounds in the context of Covariate Shift Adaptation;
2. A simple, general and theoretically sound approach to combine feature selection and Covariate Shift Adaptation. Existing solutions to the same problem are not amenable to interpretability or not general in terms of suitable hypothesis classes.

Preliminaries

- We assume $Q_{y|x} = P_{y|x}$ and $Q_x \neq P_x$ with p.d.f.s p_x and q_x such that $\text{supp}(p_x) \subseteq \text{supp}(q_x)$.
- Given a hypothesis class $\mathcal{H}$ and a loss function L, our objective is to find a hypothesis $h^* \in \mathcal{H}$ that minimizes the risk R assessed in the target distribution.

Importance Weighting

- Rewriting the risk in terms of the source distribution:

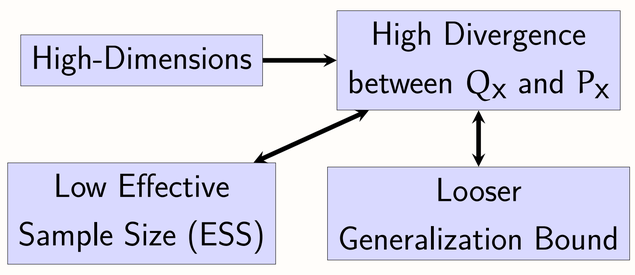
$$\begin{aligned} R(h) &= \mathbb{E}_{x\sim P_x}\mathbb{E}_{y|x}\left[L(h(x), y)\right] \\ &= \int \frac{p_x(\boldsymbol{x})}{q_x(\boldsymbol{x})} q_x(\boldsymbol{x}) \mathbb{E}_{y|x}\left[L(h(\boldsymbol{x}), y)\right] d\boldsymbol{x} \\ &= \mathbb{E}_{x\sim Q_x}\mathbb{E}_{y|x}\left[w(x)\cdot L(h(x), y)\right] \end{aligned}$$

- Having a fitted model $\hat{w}$ for w, and assuming $\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} Q_{x,y}$, we would like to find a hypothesis $h^{ERM}_{\hat{w}} \in \mathcal{H}$ that minimizes a weighted version of the regularized empirical risk:

$$\bar{R}_{\hat{w}} = \frac{1}{n}\sum_{i=1}^{n} \hat{w}(x_i)\cdot L(h(x_i), y_i) + \Omega(h)$$

Effective Sample Size (ESS) and Generalization Bounds

Main idea



Details

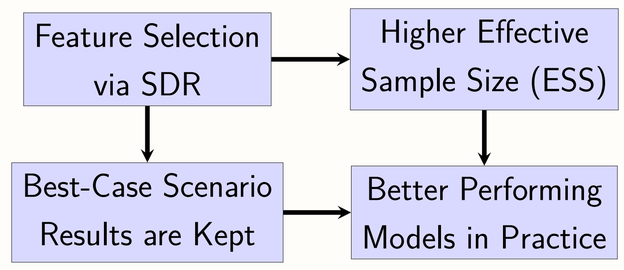
- Suppose we have a random sample $\{x_i\}_{i=1}^n \overset{iid}{\sim} Q_x$, and we define the weights $w_i \propto p_x(x_i)/q_x(x_i)$. It follows the most common formulation for the ESS:

$$ESS_n := \frac{(\sum_{i=1}^n w_i)^2}{n\sum_{i=1}^n w_i^2}$$

- Assuming $0 < \mathbb{E}_{x\sim Q_x}\left[w(x)^2\right] < \infty$, ESS_n converges almost surely to $1/d_2(P_x||Q_x)$. $\mathbf{d_2}$ being the exponential of the Rényi Divergence of order 2;
- Another important result is that, given two joint probability distributions P_{x_1,x_2} and Q_{x_1,x_2} over $\mathcal{X} \subseteq \mathbb{R}^d$, $d_2(P_{x_1,x_2}||Q_{x_1,x_2}) \geq d_2(P_{x_1}||Q_{x_1})$. That is, the Rényi Divergence (and its exponential) does not decrease with the number of variables (dimensions);
- A smaller $d_2(P_x||Q_x)$ leads to a tighter generalization bound of an importance weighted learning algorithm. In consequence, the rationale behind using ESS as an heuristic for diagnosis of Covariate Shift Adaptation becomes clearer.

Variable Selection for Covariate Shift Adaptation

Main idea



Details

- We propose a feature selection approach prior to covariate shift correction;
- Working with a good subset of features potentially enables a greater ESS and better generalization;
- Right after feature selection, the covariate shift adaptation is carried via importance weighting using off-the-shelf methods for density ratio estimation.

Variable Selection via Sufficient Dimensionality Reduction (SDR)

- Given a set of features x and a target variable y, the objective of SDR is to find a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d'}$, with $d' < d$ and $\boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{I}_{d'}$, such that $y \perp\!\!\!\perp x \mid \boldsymbol{M}^\top x$. That is, $\boldsymbol{M}^\top x$ is sufficient for y;
- We focus in the case where the matrix $\boldsymbol{M}$ is sparse and each column of it is given by zeros, except for one entry set as 1 to create a feature selector;
- Sufficient Dimensionality Reduction can be faced using the concept of Mutual Information;
- **Theorem**: Consider the random quantities $x = (x_1, x_2)$ and y with joint distribution $Q_{x,y}$. Then $I(y;x) \geq I(y;x_1)$ and $I(y;x) = I(y;x_1)$ iff $y \perp\!\!\!\perp x|x_1$;

Details

- On the selection step, we fix how many features we want to keep and then try to find the proper sized subset of features that maximizes the Mutual Information of features and labels. We adopt a selection greedy strategy known as "*Forward Selection*".
- An efficient alternative for estimating Mutual Information to perform feature selection for regression and classification problems is the use of Gaussian Mixture Models (GMMs) as density models. Other methods can be used though;

Experiments

- We used 10 popular datasets in the literature for regression and classification tasks (20 Bootstrap experiments in total);
- We caused Covariate Shift artificially and then tried to correct it with some strategies;
- We compare our approach ("Sel" - Selection and estimated weights) with three other ones: (i) whole set of features and no weighting method; (ii) whole set of features and use of 'true' weights; (iii) whole set of features and estimated weights (used the "logistic regression method");
- In approaches that used a weighting function, we used the density ratio as a weight ("Raw") and also a flattened version of it ("Adapt.");

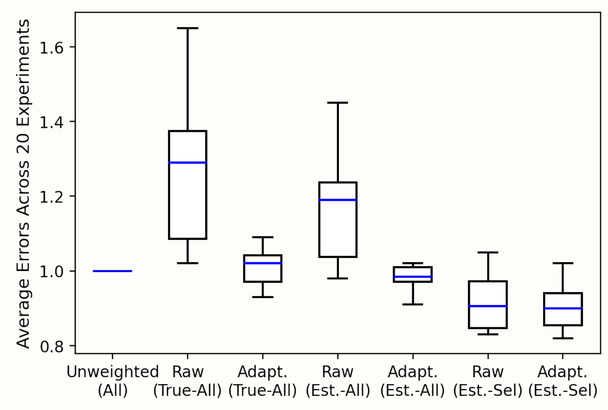


Figure 1: The average errors across all the 20 experiments. All the results are relative to the "Unweighted" benchmark. We have four basic scenarios: (i) whole set of features and no weighting method; (ii) whole set of features and use of 'true' weights; (iii) whole set of features and estimated weights; (iv) selected features and estimated weights. In the last three scenarios, we use both raw weights and their flatter version ("Adapt."). Our approach ("Sel") gives the best results.

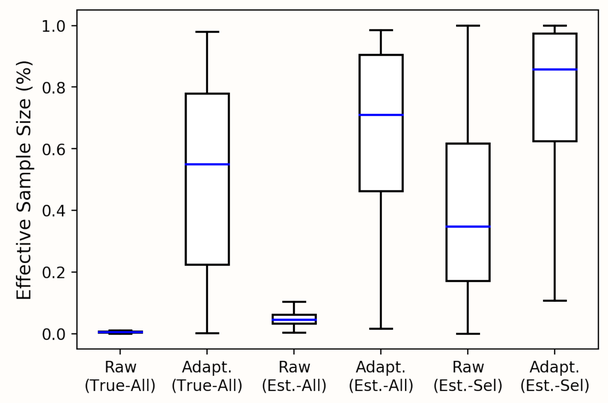


Figure 2: Effective Sample Size distributions across all experiments. Notice higher ESSs can be achieved by a prior feature selection stage. We use both raw weights and their flatter version ("Adapt.") in a combination of scenarios which includes all/selected features and true/estimated weights. Our approach ("Sel") gives the best results.