Stochastic Density Ratio Estimation and Its Application to Feature Selection*

Igor Braga†
igorab@icmc.usp.br
Advisor: Prof. Maria Carolina Monard†. *Co-advisor*: Prof. Vladimir Vapnik‡

Abstract

In this work, we deal with a relatively new statistical tool in machine learning: the estimation of the ratio of two probability densities, or density ratio estimation for short. As a side piece of research that gained its own traction, we also tackle the task of parameter selection in learning algorithms based on kernel methods.

1 Density Ratio Estimation

The estimation of the ratio of two probability densities $r(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}$ is a statistical inference problem that finds useful applications in machine learning. Several approaches have been proposed and studied for the *direct* solution of the density ratio estimation problem, that is, to estimate the density ratio without going through density estimation [Sugiyama *et al.*, 2011, and references therein]. By avoiding taking the ratio of two estimated densities, we avoid a dangerous source of error propagation.

Next, we introduce situations where density ratio estimation naturally arises.

Covariate-shift adaptation. Under the hood, most supervised learning algorithms apply the so-called *Empirical Risk Minimization* — ERM — principle, which selects a function f_n^* from a given set of functions $\mathcal{F}$ that minimizes the average of a loss function $L : R \times R \mapsto R$ over a given set of training points $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. Formally:

$$f_n^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(f(\boldsymbol{x}_i), y_i).$$

Assuming that each $(\boldsymbol{x}_i, y_i)$ is independently and identically distributed (*i.i.d.*) according to a distribution $P_{\boldsymbol{x},y}$ and the set of functions $\mathcal{F}$ has limited *capacity* (*e.g.* VC-dimension), with increasing n, the average loss of f_n^* converges with probability 1 to the smallest possible expected loss that can be achieved by a function $f \in \mathcal{F}$ [Vapnik, 1998]. This key result of Statistical Learning Theory justifies the use of the ERM principle when test and training points are sampled from the same distribution $P_{\boldsymbol{x}}$.

Now, suppose that test and training features (the so-called covariates) follow different distributions $P_{\boldsymbol{x}}^1$ and $P_{\boldsymbol{x}}^2$, and that such distributions admit densities p_1 and p_2 (resp.) such that $p_1(\boldsymbol{x}) > 0$ implies $p_2(\boldsymbol{x}) > 0$. Then, the following *modified* ERM principle

$$f_n^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(\boldsymbol{x}_i)}{p_2(\boldsymbol{x}_i)} L(f(\boldsymbol{x}_i), y_i) \qquad (1)$$

guarantees the convergence to the smallest expected loss on the *test set* [Sugiyama *et al.*, 2011]. However, before applying it, the density ratios $\frac{p_1(\boldsymbol{x}_i)}{p_2(\boldsymbol{x}_i)}$ need to be estimated from data.

Mutual information estimation and feature selection. Irrelevant features may degrade the performance of a learning algorithm. The problem of feature selection asks for criteria to tell relevant and irrelevant features apart. One such criteria is the information theoretic concept of mutual information.

In one of its forms, mutual information is calculated as

$$I(\boldsymbol{X}, Y) = \int_Y \int_{\boldsymbol{X}} p(\boldsymbol{x}, y) \log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)} \, d\boldsymbol{x} \, dy.$$

Intuitively, mutual information is a way of comparing the densities $p(\boldsymbol{x}, y)$ and $p(\boldsymbol{x})p(y)$, the former associated to the actual distribution of the data and the later associated to the distribution that assumes independence between the feature vector $\boldsymbol{x}$ and the target variable y. The estimation of the value $I(\boldsymbol{X}, Y)$ depends on the estimation of the density ratio $r(\boldsymbol{x}, y) = \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})p(y)}$. The closer $r(\boldsymbol{x}, y)$ is to the unity value for pairs $(\boldsymbol{x}, y)$ distributed according to $P_{\boldsymbol{x},y}$, the more the feature vector and the target variable are independent of each other, in which case $I(\boldsymbol{X}, Y) = 0$.

2 Parameter Selection in Kernel Methods

In supervised learning, there is an essential trade-off between training set error and the capacity of the given set of functions $\mathcal{F}$: one can always be minimized at the expense of the other. Learning algorithms usually make it possible to explore this dilemma through a set of parameters. To obtain the

*This work is supported by grant #2009/17773-7, São Paulo Research Foundation (FAPESP).

†Institute of Mathematics and Computer Science, University of São Paulo, São Carlos–SP, Brazil.

‡Data Science Institute, Columbia University, New York–NY, USA.

best performance in practice, one needs to investigate candidate solutions in this parameter space.

This research topic stemmed from the main research topic on density ratio estimation and feature selection. Eventually, we realized the initial results fitted a broader purpose, which led us to further investigate the topic. This way, we investigate in this work the parameter selection stage of Support Vector Machines — SVM — in classification and the Regularized Least Squares — RLS — method in regression. Both learning algorithms belong to the category of kernel methods.

In general, SVM have two parameters to be selected: the generalization parameter C and the kernel function k. The current practice in choosing these parameters leaves few alternatives. One either

- spends a lot of computational time using a comprehensive set of candidates that includes the best ones; or
- resorts to default parameters of the implementation of choice and risks achieving poor classification results.

Both alternatives are obviously unattractive. In order to avoid having to choose between these two alternatives in the feature selection experiments carried out in this work, it was important to investigate new alternatives that explore the gap between these two standard choices.

The RLS method has also two parameters: the regularizing constant γ and the kernel function k. In the case of RLS, we are interested in the performance of parameter selection when training sets are small. The combination of the squared loss function and the regression task is a complicating factor for parameter selection using small training sets. Contrary to SVM, the computational time spent by parameter selection procedures is not an issue. What is relevant is the risk of overfitting due to the cross-validation parameter selection procedure. Therefore, it is important to investigate alternative candidate evaluation procedures. Another important trait of the parameter selection problem in RLS is that we face a similar problem for selecting parameters of the density ratio estimation methods proposed in this work.

3 Contributions

The contributions in this work can be grouped into three categories:

1. Density ratio estimation;
2. Mutual information estimation and feature selection;
3. Parameter selection for SVM and RLS.

Regarding the first category, the approach to density ratio estimation taken in this work is based on unexplored ideas of searching the solution of a stochastic integral equation defining the ratio function. In this integral equation, we find the so-called *empirical cumulative distribution functions*. To the best of our knowledge, there is no attempt to use these functions in the literature of density ratio estimation. The proposed methods based on this approach outperform previous methods, with the advantage that their computational cost is no greater than previous methods [Vapnik *et al.*, 2014].

Using one of the proposed methods of density ratio estimation, we have developed a new mutual information estimation method, which, in turn, is employed in feature selection in classification tasks. Regarding mutual information estimation alone, the new estimator outperforms previous state-of-the-art methods [Braga, 2013]. In feature selection, the resulting algorithm provides results that are comparable to the best ones, while outperforming the popular Relief-F feature selection algorithm [Braga, 2014].

Regarding parameter selection for SVM, we have proposed and evaluated easy-to-use and economic procedures that provide reasonable results [Braga *et al.*, 2013]. In addition, we proposed a new kernel, namely the *min* kernel, which mixes the advantages of linear and non-linear kernels, resulting in fast, easily applicable, and universal SVM classification.

In RLS, we have investigated several parameter selection methods that were shown to perform well for other regression methods. Unfortunately, no method is able to consistently outperform cross-validation, though we find situations where some alternative methods perform comparably well. Regarding the kernel function, we proposed the use of the additive INK-splines kernel instead of RBF or the multiplicative INK-splines kernel. The proposed kernel function clearly outperforms the other ones in the small sample size regime [Braga and Monard, 2013; 2015].

References

[Braga and Monard, 2013] Igor Braga and Maria Carolina Monard. Statistical and heuristic model selection in regularized least-squares. In *BRACIS '13: Proceedings of the 2013 Brazilian Conference on Intelligent Systems*, pages 231–236, 2013.

[Braga and Monard, 2015] Igor Braga and Maria Carolina Monard. Improving the kernel regularized least squares method for small-sample regression (in print). *Neurocomputing*, 2015.

[Braga *et al.*, 2013] Igor Braga, Lais P. do Carmo, Caio Csar Benatti, and Maria Carolina Monard. A note on parameter selection for support vector machines. In *MICAI '13: Proceedings of the 2013 Mexican International Conference on Artificial Intelligence*, pages 233–244, 2013.

[Braga, 2013] Igor Braga. The constructive density-ratio approach to mutual information estimation: An experimental comparison. In *KDMile '13: Proceedings of 2013 Symposium on Knowledge Discovery, Mining and Learning*, pages 1–4, 2013.

[Braga, 2014] Igor Braga. A constructive density-ratio approach to mutual information estimation: experiments in feature selection. *Journal of Information and Data Management*, 5(1):134–143, 2014.

[Sugiyama *et al.*, 2011] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2011.

[Vapnik *et al.*, 2014] Vladimir Vapnik, Igor Braga, and Rauf Izmailov. A constructive setting for the problem of density ratio estimation. In *SDM '14: Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 434–442, 2014.

[Vapnik, 1998] Vladimir Naumovich Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.