Locally weighted regression through data imprecisiation

Shenzhen Lu, Eyke Hüllermeier

Department of Computer Science
University of Paderborn
E-Mail: lushenzhen@gmail.com, eyke@upb.de

Introduction

This extended abstract outlines the idea of realizing locally weighted learning and statistical inference within the framework of *superset learning*. More specifically, we propose an alternative to standard locally weighted linear regression, which is commonly used in statistics and machine learning. Our approach is based on replacing precisely observed output values by intervals—a process we refer to as "data imprecisiation". As will be explained in more detail later on, the influence of an observation can thus be controlled by the length of the corresponding interval.

Our approach builds on a generic framework for superset learning that we recently introduced in [5, 6], and that will be briefly recalled in the next section. The main purpose of this framework is to support the systematic development of methods for learning from imprecise or ambiguous data, namely, training data that is characterised in terms of sets of candidate values. Additionally, however, it can be used for learning from standard (precise) data, which is deliberately turned into imprecise data. In this way, different effects can be achieved, including the one already mentioned, namely the weighing of the influence of a training example on the overall result of the learning process: the more imprecise an observation is made, the less it will influence the model or prediction induced from the data.

In the next section, we recall our generic approach to superset learning based on generalized loss minimization. Then, we show how this approach can be used to develop an alternative method for locally weighted linear regression. Prior to concluding, we present some experimental results.

Superset Learning

Superset learning is a specific type of learning from weak supervision, in which the outcome (response) associated with a training instance is only characterized in terms of a subset of possible candidates. Thus, superset learning is somehow in-between supervised and semi-supervised learning, with the latter being a special case (in which supersets are singletons for the labeled examples and cover the entire output space for the unlabeled ones). There are numerous applications in which only partial information about outcomes is available [8]. Correspondingly, the superset learning problem has received increasing attention and has been studied by various authors in recent years, albeit under different names [4, 7, 9, 2].

Setting

Consider a standard setting of supervised learning with an input (instance) space $\mathcal{X}$ and an output space $\mathcal{Y}$. The goal is to learn a mapping from $\mathcal{X}$ to $\mathcal{Y}$ that captures, in one way or the other, the dependence of outputs (responses) on inputs (predictors). The learning problem essentially consists of choosing an optimal model (hypothesis) M^* from a given model space (hypothesis space) $\mathbf{M}$, based on a set of training data

$$\mathcal{D} = \Big\{ (\boldsymbol{x}_n, y_n) \Big\}_{n=1}^{N} \in (\mathcal{X} \times \mathcal{Y})^N \ . \tag{1}$$

More specifically, optimality typically refers to optimal prediction accuracy, i.e., a model is sought whose expected prediction loss or *risk*

$$\mathcal{R}(M) \, = \, \int L\big(y, M(\boldsymbol{x})\big) \, d\,\mathbf{P}(\boldsymbol{x}, y) \tag{2}$$

is minimal; here, $L : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$ is a loss function, and $\mathbf{P}$ is an (unknown) probability measure on $\mathcal{X} \times \mathcal{Y}$ modeling the underlying data generating process.

Here, we are interested in the case where output values $y_n \in \mathcal{Y}$ are not necessarily observed precisely; instead, only a superset $Y_n \subseteq \mathcal{Y}$ is observed, i.e., a subset Y_n such that $y_n \in Y_n$. Therefore, the learning algorithm does not have direct access to the (precise) data (1), but only to the (imprecise, ambiguous) observations

$$\mathcal{O} = \Big\{ (\boldsymbol{x}_n, Y_n) \Big\}_{n=1}^{N} \in (\mathcal{X} \times 2^{\mathcal{Y}})^N \ . \tag{3}$$

Generalized Loss Minimization

Recall the principle of *empirical risk minimization* (ERM): A model M^* is sought that minimizes the *empirical risk*

$$\mathcal{R}_{emp}(M) = \frac{1}{N}\sum_{n=1}^{N} L\big(y_n, M(\boldsymbol{x}_n)\big) \ , \tag{4}$$

i.e., the average loss on the training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. The empirical risk (4) serves as a surrogate of the true risk (2). In order to avoid the problem of possibly *overfitting* the data, not (4) itself is typically minimized but a *regularized* version thereof.

In [6], we developed a generic approach to superset learning that can be seen as a generalization of empirical risk minimization. More specifically, this approach is based on the minimization of the empirical risk with respect to the generalized loss function or *optimistic superset loss* (OSL)

$$L^* : 2^{\mathcal{Y}} \times \mathcal{Y} \longrightarrow \mathbb{R}_+, \ (Y, \hat{y}) \mapsto \min \big\{ L(y, \hat{y}) \,|\, y \in Y \big\} \tag{5}$$

instead of the original loss L. Thus, each candidate model $M \in \mathbf{M}$ is evaluated in terms of

$$\overline{\mathcal{R}}_{emp}(M) = \frac{1}{N}\sum_{n=1}^{N} L^*\big(Y_n, M(\boldsymbol{x}_n)\big) \ , \tag{6}$$

and an optimal model M^* is one that minimizes (6). The choice of the minimum as an aggregation of the possible true losses in (5) is motivated by the goal of "data disambiguation", i.e., of finding the most plausible instantiations $y_n^* \in Y_n$ of the ambiguous observations Y_n. For details of this approach, we refer to [**?**].

Interestingly, several existing machine learning methods are recovered as special cases of our framework, i.e., for specific combinations of output space, loss function and imprecisiation of the data. For example, support vector regression [10] is obtained as a generalisation of standard regression with L_1 loss if precise output values $y_n \in \mathbb{R}$ are replaced by interval-valued data $Y_n = [y_n - \epsilon, y_n + \epsilon]$, i.e., ϵ-intervals around the original data points; in fact, our generalized loss then corresponds to the ϵ-insensitive loss function used in support vector regression.

Locally Weighted Linear Regression

Obviously, the OSL L^* is a relaxation of the original loss L in the sense that $L^* \leq L$. More specifically, the larger the set Y, the smaller the loss:

$$Y \supset Y' \;\Rightarrow\; \forall\, \hat{y} \in \mathcal{Y}:\; L^*(Y, \hat{y}) \leq L^*(Y', \hat{y})$$

Thus, the loss $L(y, \hat{y})$ incurred for a prediction $\hat{y}$ can be weakened by replacing the original observation y with a subset around y, and the larger the subset, the smaller the loss. This observation is on the basis of our idea of realizing locally weighted inference within the framework of superset learning.

In particular, our framework suggests natural approaches to locally weighted linear and support vector regression that deviate from the standard approaches [1, 3]. In standard locally weighted regression, a prediction $\hat{y}$ for a query instance $\boldsymbol{x}$ takes the form $\hat{y} = \boldsymbol{x}^\top \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is obtained by minimizing a sum of weighted losses:

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) L\left(y_n, \boldsymbol{x}_n^\top \boldsymbol{\beta}\right) \quad , \tag{7}$$

where $k(\boldsymbol{x}, \cdot)$ is a kernel function that assigns large weights to instances $\boldsymbol{x}_n$ close to $\boldsymbol{x}$ and smaller weights to instances farther away; moreover, L is a loss function such as absolute or squared difference.

Instead of *weighing* each individual loss $L\left(y_n, \hat{y}_n\right)$ in terms of a constant factor $c_n = k(\boldsymbol{x}, \boldsymbol{x}_n)$, our approach suggests another modification, namely, a specific kind of “stretching” of the loss function around the observed outcome y_n, which is achieved by the OSL (5) if y_n is replaced by a superset $Y_n \ni y_n$. This superset reasonably takes the form of an interval $[y_n - \delta_n, y_n + \delta_n]$, where the length δ_n plays the role of the weight $c_n = k(\boldsymbol{x}, \boldsymbol{x}_n)$ in the original approach. Thus, our method finds the generalized empirical risk minimizer

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{n=1}^{N} L^*\left(Y_n, \boldsymbol{x}_n^\top \boldsymbol{\beta}\right) \quad . \tag{8}$$

This is accomplished by an iterative algorithm that alternates between two steps:

- Given a current *instantiation* of the set-valued data, i.e., values $y_n^* \in Y_n$, $n = 1, \ldots, N$, a parameter vector $\boldsymbol{\beta}^*$ is fit to this data using standard linear regression (the first instantiation is initialized with the original data y_n, i.e., the midpoints of the intervals Y_n).

- Then, given this solution, an improved instantiation is determined by replacing the current values y_n^* with those that appear most plausible under this solution:

$$y_n^* \leftarrow \arg\min_{y \in Y_n} L(y, \boldsymbol{x}_n^\top \boldsymbol{\beta}^*)$$

For this algorithm, convergence to the optimal solution (8) can be proved formally. The values y_n^* eventually found serve as a (hypothetical) *disambiguation* of the set-valued data Y_n, $n = 1, \ldots, N$.

Experiments

In a first experimental study, we compared our interval-based approach to locally weighted linear regression with the conventional one on a number of UCI benchmark data sets. The loss function L in (7) was initialised with the standard squared error loss, and the Gaussian kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\lambda^2 \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$ was used. To assure maximal comparability of the two approaches, we used a similar function, namely $\exp(\lambda^2 \|\boldsymbol{x} - \boldsymbol{x}'\|^2) - 1$, to specify the width of intervals in our method. Prediction accuracy (mean squared error) was estimated by means of a 10-fold cross validation, and the hyper-parameter λ was selected in an internal (5-fold) cross validation.

Table 1 shows the results in terms of average error ± standard deviation. As can be seen, the two methods perform more or less on a par; at least, there are no statistically significant differences between them.

Table 1: Experiment result on real-world data (with standardized outputs).

data set	locally weighted	interval-based
Breast cancer Wisconsin	0.1399 ± 0.4781	0.1432 ± 0.4621
Red wine quality	0.6083 ± 0.9737	0.5936 ± 0.9601
White wine quality	0.6456 ± 1.3481	0.6291 ± 1.1302
Community violence pred.	0.0361 ± 0.0735	0.0419 ± 0.0704
Combined cycle power plant	0.0612 ± 0.0944	0.0648 ± 0.0933
Parkinsons telemonitoring	0.3765 ± 0.9295	0.3596 ± 0.6646
Physicochemical properties of protein tertiary structure	0.1011 ± 0.3913	0.1582 ± 0.7911

Concluding Remarks

Overall, the results are quite promising, suggesting that our method based on superset learning provides a viable alternative to standard locally weighted learning. This provides a strong motivation for investigating this idea in more detail, not only from an empirical but also from a theoretical and algorithmic point of view.

Currently, we are elaborating on locally weighted linear regression with L_1 instead of L_2 loss, a case for which our framework suggests an alternative approach to locally weighted support vector regression. Of course, going beyond local regression, the same framework can also be applied to generalize any other type of instance weighing in machine learning. Thus, building on the results so far, there are various interesting lines of research to be explored in future work.

References

[1] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 1979.

[2] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

[3] E.E. Elattar, J.Y. Goulermas, and Q.H. Wu. Electric load forecasting based on locally weighted support vector regression. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 40(4):438–447, 2010.

[4] Y. Grandvalet. Logistic regression for partial labels. In *IPMU–02, Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1935–1941, Annecy, France, 2002.

[5] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

[6] E. Hüllermeier and W. Cheng. Superset learning based on generalized loss minimization. In *Proceedings ECML/PKDD–2015, European Conference on Machine Learning and Knowledge Discovery in Databases*, Porto, Portugal, 2015.

[7] R. Jin and Z. Ghahramani. Learning with multiple labels. In *16th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2002.

[8] L.P. Liu and T.G. Dietterich. A conditional multinomial mixture model for superset label learning. In *Proc. NIPS*, 2012.

[9] N. Nguyen and R. Caruana. Classification with partial labels. In *Proc. KDD 2008, 14th Int. Conf. on Knowledge Discovery and Data Mining*, Las Vegas, USA, 2008.

[10] B. Schölkopf and AJ. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2001.