Localized Lasso for High-Dimensional Regression

Makoto Yamada[1,2], **Koh Takeuchi**[3], **Tomoharu Iwata**[3], **John Shawe-Taylor**[4], **Samuel Kaski**[5]
[1]RIKEN AIP, [2]JST, PRESTO, [3]NTT CS Labs, [4]UCL, [5]Aalto University
makoto.yamada@riken.jp, {takeuchi.koh,iwata.tomoharu}@lab.ntt.co.jp
j.shawe-taylor@ucl.ac.uk, samuel.kaski@aalto.fi

Abstract

We introduce the localized Lasso, which learns models that both are interpretable and have a high predictive power in problems with high dimensionality d and small sample size n. More specifically, we consider a function defined by local sparse models, one at each data point. We introduce sample-wise network regularization to borrow strength across the models, and sample-wise exclusive group sparsity (a.k.a., $\ell_{1,2}$ norm) to introduce diversity into the choice of feature sets in the local models. The local models are interpretable in terms of similarity of their sparsity patterns. The cost function is convex, and thus has a globally optimal solution. Moreover, we propose a simple yet efficient iterative least-squares based optimization procedure for the localized Lasso, which does not need a tuning parameter, and is guaranteed to converge to a globally optimal solution. The solution is empirically shown to outperform alternatives for both simulated and genomic personalized/precision medicine data.

Problem Formulation

Let us denote an input vector by $\boldsymbol{x} = [x^{(1)}, \ldots, x^{(d)}]^\top \in \mathbb{R}^d$ and the corresponding output value $y \in \mathbb{R}$. The set of samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ has been drawn i.i.d. from a joint probability density $p(\boldsymbol{x}, y)$. We further assume a graph $\boldsymbol{R} \in \mathbb{R}^{n\times n}$, where $[\boldsymbol{R}]_{i,j} = r_{ij} \geq 0$ is the coefficient that represents the relatedness between the sample pair $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_j, y_j)$. In this paper, we assume that $\boldsymbol{R} = \boldsymbol{R}^\top$ and the diagonal elements of $\boldsymbol{R}$ are zero.

The goal in this paper is to select multiple sets of features such that each set of features is locally associated with an individual data point or a cluster, from the training input-output samples and the graph information $\boldsymbol{R}$.

Proposed method

We employ the following linear model for each sample i:

$$y_i = \boldsymbol{w}_i^\top \boldsymbol{x}_i. \tag{1}$$

Here $\boldsymbol{w}_i \in \mathbb{R}^d$ contains the regression coefficients for sample $\boldsymbol{x}_i$ and $^\top$ denotes the transpose. Note that in regression problems the weight vectors are typically assumed to be equal, $\boldsymbol{w} = \boldsymbol{w}_1 = \ldots = \boldsymbol{w}_n$. Since we cannot assume the models to be based on the same features, and we want to interpret the support of the model for each sample, we use local models. The optimization problem of the *localized lasso* can be written as

$$\min_{\boldsymbol{W}} \; J(\boldsymbol{W}) = \sum_{i=1}^n (y_i - \boldsymbol{w}_i^\top \boldsymbol{x}_i)^2 + \lambda_1 \sum_{i,j=1}^n r_{ij} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 + \lambda_2 \sum_{i=1}^n \|\boldsymbol{w}_i\|_1^2, \tag{2}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the regularization parameters. By imposing the network regularization (second term) (Hallac, Leskovec, and Boyd 2015), we regularize the model parameters $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ to be similar if $r_{ij} > 0$. If λ_1 is large, we will effectively cluster the samples according to how similar the $\boldsymbol{w}_i$s are, that is, according to the prediction criteria in the local models. More specifically, when $\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2$ is small (possibly zero), we can regard the i-th sample and j-th sample to belong to the same cluster.

The third term is the $\ell_{1,2}$ regularizer (a.k.a., exclusive regularizer) (Kowalski 2009; Zhou, Jin, and Hoi 2010; Kong 2014). By imposing the $\ell_{1,2}$ regularizer, we can select a small number of elements within each $\boldsymbol{w}_i$.

Predicting for new test sample: For predicting on test sample $\boldsymbol{x}$, we use the estimated local models $\widehat{\boldsymbol{w}}_k$ which are linked to the input $\boldsymbol{x}$. More specifically, we solve the Weber problem (Hallac, Leskovec, and Boyd 2015)

$$\min_{\boldsymbol{w}} \quad \sum_{i=1}^n r'_i \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_i\|_2, \tag{3}$$

where $r'_i \geq 0$ is the link information between the test sample and the training sample $\boldsymbol{x}_i$. Since this problem is convex, we can solve it efficiently by an iterative update formula. If there is no link information available, we simply average all $\widehat{\boldsymbol{w}}_i$s to estimate $\widehat{\boldsymbol{w}}$, and then predict as $\widehat{y} = \widehat{\boldsymbol{w}}^\top \boldsymbol{x}$.

References

Hallac, D.; Leskovec, J.; and Boyd, S. 2015. Network lasso: Clustering and optimization in large graphs. In *KDD*.

Kong, D. e. a. 2014. Exclusive feature learning on arbitrary structures via ℓ_{12}-norm. In *NIPS*.

Kowalski, M. 2009. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis* 27(3):303–324.

Zhou, Y.; Jin, R.; and Hoi, S. C. 2010. Exclusive lasso for multi-task feature selection. In *AISTATS*.