Discriminative Elastic-Net Regularized Linear Regression

Zheng Zhang

I. A GENERAL FRAMEWORK OF ELASTIC-NET REGULARIZED LINEAR REGRESSION MODEL

To learn a compact and discriminative projection matrix, a general framework of elastic-net regularization based linear regression model is formulated as

$$\min_{\boldsymbol{D}} \phi(\boldsymbol{D}) + \lambda_1 \|\boldsymbol{D}\|_* + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2, \tag{1}$$

where λ_1 and λ_2 are the regularization parameters for balancing respective terms. The most straightforward regression loss function is $\phi(\boldsymbol{D}) = \|\boldsymbol{X^T D} - \boldsymbol{Y}\|_F^2$. For the above objective function (1), we have the following proposition.

A. Discriminative Elastic-net Regularized Linear Regression

By introducing the ε-dragging technique, a discriminative elastic-net regularized linear regression (DENLR) model is developed, and its objective function is formulated as

$$\min_{\boldsymbol{D}} \psi(\boldsymbol{D}) + \lambda_1 \|\boldsymbol{D}\|_* + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2, \tag{2}$$

where $\psi(\boldsymbol{D}) = \|\boldsymbol{X^T D} - \tilde{\boldsymbol{Y}}\|_F^2$ and $\tilde{\boldsymbol{Y}}$ is the relaxed regression target matrix.

To obtain an optimal $\tilde{\boldsymbol{Y}}$, an elaborate strategy is devised as follows. Let $\boldsymbol{E}$ be a constant matrix, and the i-th row and j-th column entry is defined as

$$\boldsymbol{E}_{ij} = \begin{cases} +1 & if \quad \boldsymbol{Y}_{ij} = 1 \\ -1 & if \quad \boldsymbol{Y}_{ij} = 0, \end{cases} \tag{3}$$

and then, we have $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{E} \odot \boldsymbol{M}$, where $\boldsymbol{M} \in \Re^{n \times c}$ is a learned nonnegative matrix. Thus, the proposed DENLR model (8) is rewritten as the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{D},\boldsymbol{M}} & \|\boldsymbol{X^T D} - (\boldsymbol{Y} + \boldsymbol{E} \odot \boldsymbol{M})\|_F^2 + \lambda_1 \|\boldsymbol{D}\|_* \\ & + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \quad s.t. \quad \boldsymbol{M} \geq 0. \end{aligned} \tag{4}$$

B. Marginalized Elastic-net Regularized Linear Regression

From problem (4), we can see that the relaxed target space of DENLR is subject to the bound that the regression results should be larger than 1 for true classes and smaller than 0 for false classes. However, this target space is still based on the zero-one label matrix $\boldsymbol{Y}$, which greatly confines the flexibility of the regression model. To this end, we propose to directly learn the regression targets from data, and a marginalized constraint is enforced to make the learned targets distinguishable. We consider the following marginalized elastic-net regularized linear regression (MENLR) problem:

$$\begin{aligned} \min_{\boldsymbol{D},\boldsymbol{R}} & \|\boldsymbol{X}^T\boldsymbol{D} - \boldsymbol{R}\|_F^2 + \lambda_1 \|\boldsymbol{D}\|_* + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \\ & s.t. \quad \boldsymbol{r}_{iy_i} - \max_{j \neq y_i} \boldsymbol{r}_{ij} \geq C, i = 1, \cdots, n, \end{aligned} \tag{5}$$

where $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n]^T \in \Re^{n \times c}$ is the learned regression targets, and C is a constant. Herein y_i denotes the index of the true class for the i-th sample $\boldsymbol{x}_i$. That is, if the i-th sample is from the m-th class (i.e. y_i=m), the value of the m-th element of the learned target vector $\boldsymbol{r}_i$, i.e. $\boldsymbol{r}_{im}$, should be bigger than the rest of the elements by a fixed margin of C. Similar to SVM, we simply set the marginal value between the true and the false classes to 1, i.e. $C = 1$. Apparently, the marginalized constraint makes the learned regression targets between the true and false classes separable by a fixed distance such that the proposed MENLR is more flexible and discriminative.

C. Efficient MENLR

Based on the Theorem 1, we make an equivalent representation of MENLR as

$$\begin{aligned} & \min_{\boldsymbol{D},\boldsymbol{R}} \|\boldsymbol{X}^T\boldsymbol{D} - \boldsymbol{R}\|_F^2 + \frac{\lambda_1}{2}(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2) \\ & + \frac{\lambda_2}{2}\|\boldsymbol{D}\|_F^2 \quad s.t. \quad \boldsymbol{D} = \boldsymbol{AB}, \boldsymbol{r}_{iy_i} - \max_{j \neq y_i} \boldsymbol{r}_{ij} \geq C. \end{aligned} \tag{6}$$

D. Optimization of MENLR

It is easy to find that optimization of MENLR is very similar to the optimization procedures of DENLR, except for deducing the regression targets matrix $\boldsymbol{R}$.

Updating $\boldsymbol{A}$: Fix the other variables and update $\boldsymbol{A}$ by solving the following problem.

$$\begin{aligned} \boldsymbol{A}^+ &= \arg\min_{\boldsymbol{A}} \frac{\lambda_1}{2}\|\boldsymbol{A}\|_F^2 + < \boldsymbol{C}_1, \boldsymbol{D} - \boldsymbol{AB} > + \frac{\mu}{2}\|\boldsymbol{D} - \boldsymbol{AB}\|_F^2 \\ &= \arg\min_{\boldsymbol{A}} \frac{\lambda_1}{2}\|\boldsymbol{A}\|_F^2 + \frac{\mu}{2}\|\boldsymbol{D} - \boldsymbol{AB} + \frac{\boldsymbol{C}_1}{\mu}\|_F^2, \end{aligned} \tag{7}$$

where the rest terms irrelevant to $\boldsymbol{A}$ in $\mathcal{L}$ are viewed as constants and ignored in the loss since they make no differences in this particular procedure. The resulting problem (7) is a typical regularized least square problem, hence its solution is easily obtained as

$$\boldsymbol{A}^+ = (\boldsymbol{C}_1 + \mu\boldsymbol{D})\boldsymbol{B}^T(\lambda_1\boldsymbol{I} + \mu\boldsymbol{B}\boldsymbol{B}^T)^{-1}. \tag{8}$$

Z. Zhang is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail:darrenzz219@gmail.com).

Updating $\boldsymbol{B}$: The variable $\boldsymbol{B}$ plays a symmetric role to that of $\boldsymbol{A}$ in $\mathcal{L}$, hence the updating of $\boldsymbol{B}$ is performed in a symmetric way:

$$\begin{aligned}\boldsymbol{B}^{+} &= \arg\min_{\boldsymbol{B}} \frac{\lambda_1}{2}\|\boldsymbol{B}\|_F^2 + \langle \boldsymbol{C}_1, \boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\rangle + \frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}\|_F^2 \\ &= \arg\min_{\boldsymbol{B}} \frac{\lambda_1}{2}\|\boldsymbol{B}\|_F^2 + \frac{\mu}{2}\|\boldsymbol{D}-\boldsymbol{A}\boldsymbol{B}+\frac{\boldsymbol{C}_1}{\mu}\|_F^2.\end{aligned} \tag{9}$$

Similarly,

$$\boldsymbol{B}^{+} = (\lambda_1\boldsymbol{I} + \mu\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{C}_1+\mu\boldsymbol{D}). \tag{10}$$

Updating $\boldsymbol{R}$: By ignoring the constant terms independent of $\boldsymbol{R}$, minimizing (6) becomes the following optimization problem:

$$\min_{\boldsymbol{R}} \|\boldsymbol{H}-\boldsymbol{R}\|_F^2 \ s.t.\ \boldsymbol{r}_{iy_i} - \max_{j\neq y_i}\boldsymbol{r}_{ij} \geq 1, i=1,\cdots,n, \tag{11}$$

where $\boldsymbol{H}$=$\boldsymbol{X}^T\boldsymbol{D} \in \Re^{n\times c}$. Because problem (11) is a constrained quadratic programming problem, it can be decomposed into n independent subproblems. Suppose that the i-th sample $\boldsymbol{x}_i$ is from the mth-class, and then the i-th subproblem of (11) is

$$\min_{\boldsymbol{r}_i} \|\boldsymbol{h}_i - \boldsymbol{r}_i\|^2 \ s.t.\ \boldsymbol{r}_{im} - \max_{j\neq m}\boldsymbol{r}_{ij} \geq 1, \tag{12}$$

where $\boldsymbol{r}_i \in \Re^c$ and $\boldsymbol{h}_i \in \Re^c$ are the i-th row of $\boldsymbol{R}$ and $\boldsymbol{H}$, respectively. It should be noted that $\|\boldsymbol{h}_i - \boldsymbol{r}_i\|^2 = \sum_{j=1}^{c}(\boldsymbol{h}_{ij} - \boldsymbol{r}_{ij})^2$. To optimize problem (12), we introduce an auxiliary variable $\boldsymbol{\varphi} \in \Re^c$, and for the j-th entry, $\boldsymbol{\varphi}_j = \boldsymbol{r}_{ij} + 1 - \boldsymbol{r}_{im}$, where $\boldsymbol{\varphi}_j \leq 0$ indicates the optimal target, otherwise a unsatisfactory target. Assume that the optimal target for the true class $\boldsymbol{r}_{im}$ can be obtained by a modification of the regression result $\boldsymbol{h}_{im}$, i.e. $\boldsymbol{r}_{im} = \boldsymbol{h}_{im}+\zeta$, where ζ is a learning parameter. For the false class $\forall j \neq m$, we need $\boldsymbol{r}_{im} - \boldsymbol{r}_{ij} \geq 1$, and then the j-th subproblem of (12) is

$$\min_{\boldsymbol{r}_{ij}} (\boldsymbol{h}_{ij} - \boldsymbol{r}_{ij})_2^2 \ s.t.\ \boldsymbol{h}_{im} + \zeta - \boldsymbol{r}_{ij} \geq 1, \forall j \neq m, \tag{13}$$

which is a very simple quadratic programming problem. In this way, the optimal solution is $\boldsymbol{r}_{ij} = \boldsymbol{h}_{ij} + min(\zeta - \boldsymbol{\varphi}_j, 0)$, and the optimal solution of problem (13) is achieved by

$$\boldsymbol{r}_{ij} = \begin{cases} \boldsymbol{h}_{ij} + \zeta, & if\ j = m, \\ \boldsymbol{h}_{ij} + min(\zeta - \boldsymbol{\varphi}_j), & otherwise. \end{cases} \tag{14}$$

By substituting (14) into problem (12), we can obtain the following optimization problem:

$$\arg\min_{\zeta} \phi(\zeta) = \zeta^2 + \sum_{j\neq m}(min(\zeta - \boldsymbol{\varphi}_j))^2, \tag{15}$$

and its first-order derivation $\phi'(\zeta) = 2(\zeta + \sum_{j\neq m} min(\zeta - \boldsymbol{\varphi}_j))$. By setting $\phi'(\zeta) = 0$, we can achieve the optimal value of learning factor $\hat{\zeta}$. Specifically, let $\hat{\zeta}$ being the optimal solution that means $\phi^{'}(\hat{\zeta}) = 0$. It is easy to prove that $\phi^{'}(\cdot)$ is a monotone increasing piecewise function. Therefore, $\phi^{'}(\boldsymbol{\varphi}_j) > 0 \Leftrightarrow \boldsymbol{\varphi}_j > \hat{\zeta}$. Now, we have

$$\begin{aligned}\arg\min_{\hat{\zeta}} \frac{1}{2}\phi'(\hat{\zeta}) &= \hat{\zeta} + \sum_{j\neq m} \min(\hat{\zeta} - \boldsymbol{\varphi}_j) \\ &= \hat{\zeta} + \sum_{j\neq m}(\hat{\zeta} - \boldsymbol{\varphi}_j)\Pi(\boldsymbol{\varphi}_j > \hat{\zeta}) \\ &= \hat{\zeta} + \sum_{j\neq m}(\hat{\zeta} - \boldsymbol{\varphi}_j)\Pi(\phi^{'}(\boldsymbol{\varphi}_j) > 0)\end{aligned} \tag{16}$$

where $\Pi(\cdot)$ is the indicator operator. Hence, by setting $\phi^{'}(\hat{\zeta}) = 0$, we have the optimal solution of ζ, that is,

$$\hat{\zeta} = \frac{\sum_{j\neq m}\boldsymbol{\varphi}_j\Pi(\phi'(\boldsymbol{\varphi}_j) > 0)}{1 + \sum_{j\neq m}\boldsymbol{\varphi}_j\Pi(\phi'(\boldsymbol{\varphi}_j) > 0)}, \tag{17}$$

The detailed process of learning the optimal solution of the i-th row of $\boldsymbol{R}$ is given in Algorithm 2.

Updating $\boldsymbol{D}$: Fix the other variables and update $\boldsymbol{D}$, the optimal solution of $\boldsymbol{D}$ is computed as

$$\boldsymbol{D}^{+} = (2\boldsymbol{X}\boldsymbol{X}^T + \lambda_2\boldsymbol{I} + \mu\boldsymbol{I})^{-1}(2\boldsymbol{X}\boldsymbol{R} + \mu\boldsymbol{A}\boldsymbol{B} - \boldsymbol{C}_1). \tag{18}$$

Algorithm 2. Solving Problem (12)

Input: $\boldsymbol{r} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_c]^T \in \Re^c$, the true class index m.
Initialization: $\forall j, \boldsymbol{\varphi}_j = \boldsymbol{h}_{ij} + 1 - \boldsymbol{h}_{im}$, $\zeta = 0$, $iter = 0$.
for $j \neq m$ **do**
  if $\psi'(\boldsymbol{\varphi}_j) > 0$ **then** $\zeta = \zeta + \boldsymbol{\varphi}_j, iter = iter + 1$ **end**
end
Define $\zeta = \zeta/(1 + iter)$, and then update $\boldsymbol{r}_j$ by Eqn.(14).
Output: Marginalized target vector $\boldsymbol{r}_i$.

Algorithm 3. Optimization of MENLR by Exact ALM

Require: Feature Matrix $\boldsymbol{X}$; Label Matrix $\boldsymbol{Y}$; Parameters λ_1, λ_2.
Initialization: $\boldsymbol{T} = \boldsymbol{Y}$, $\boldsymbol{D} \in \Re^{d\times c}$, $\boldsymbol{A} \in \Re^{d\times r}$, $\boldsymbol{B} \in \Re^{r\times c}$, $\lambda_1 > 0$, $\lambda_2 > 0$, $\boldsymbol{C}_1 \in \Re^{d\times c}$, $\mu > 0$.
While not converged **do**
  While not converged **do**
    Step 1. Update $\boldsymbol{A}$ by using (8);
    Step 2. Update $\boldsymbol{B}$ by using (10);
    Step 3. Update $\boldsymbol{D}$ by using (18);
    Step 4. Update $\boldsymbol{R}$ row-by-row by using Algorithm 2;
  End While
  Step 5. Update the Lagrange multipliers $\boldsymbol{C}_1$ by
    $\boldsymbol{C}_1 = \boldsymbol{C}_1 + \mu(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{B})$.
End While
Output: Projection matrix $\boldsymbol{D}$