Group Sparsity and Optimization

Presenter: Zhe Wang

https://qdata.github.io/deep2Read

201909

Content

Background

Sparsity:

- Feature Selection
- Avoid Overfitting
- Prior Knowledge

Categories: Exponential family

- L_0: discontinuous, nonconvex
- L_1: continuous, non-smooth, convex, Laplacian prior
- L_2: smooth, convex, Gaussian prior
- Group Sparsity: structure involved, non-smooth (very sharp)

Model

Multi-tasks:

L models for L tasks.

For task j, training set: $\{x_i^j, y_i^j\}_{i=1}^{m_i}$, model: w^j, where $x_i^j, w_j \in \mathbb{R}^k$

Data fitting term: $\frac{1}{2}||Y^j - X^j w^j||_F^2$

Parameter matrix W, i_{th} row contains parameters for i_{th} model.

If sparsity is added to each model, it is Lasso.

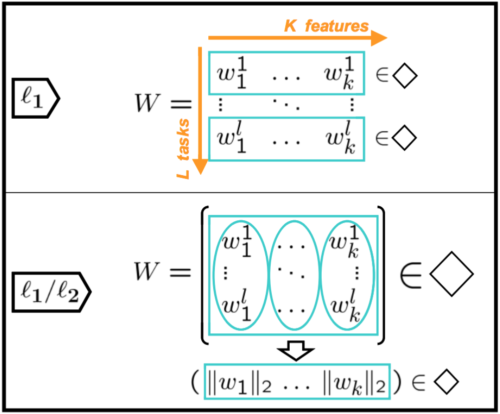


Figure 1: The ℓ_1/ℓ_2 vs the ℓ_1 regularization schemes .

Assumption

The parameters of same features have similar behavior.
They should be either all 0 or all nonzero.
Why feature selection?

Example

Suppose a polynomial regression, t_{th} feature is x^t.
$W_t = 0 \rightarrow x^t$ is not used for all models.

VI framework

Suppose $y_i^j = f^j(x_i^j) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$

$$p(y_i^j | x_i^j, w^j) \propto \exp(-\frac{[y_i^j - x_i^j w^j]^2}{2\sigma^2}) \tag{1}$$

Prior knowledge?

$$p(w_i | \delta_i) \propto \exp(-\delta ||w_i||_2) \tag{2}$$

Posterior distribution for W:

$$-logp(W|Y, X, \sigma, \delta) = \sum_{j=1}^{L} \frac{c}{2} ||Y^j - X^j w^j|| + \sum_{i=1}^{K} \delta ||w_i||_2 \tag{3}$$

Optimization

- Proximal Operator
- Alternating direction method of multipliers (ADMM)

Proximal Operator

Definition of Proximal Operator:

$$Prox_{\lambda,f}(y) = \arg\min_{x} \frac{1}{2}||y - x||_2^2 + \lambda f(x) \tag{4}$$

A generalized projection.
Suppose $I(x)$ is the indicator function on some convex set X:

$$I_X(x) = \begin{cases} 0, x \in X \\ \infty, otherwise \end{cases} \tag{5}$$

then, $Prox_I(y) = \arg\min_{x \in X} ||y - x||_2^2 = Proj_X(y)$

Closed form solution for l_1, l_2, nuclear norm, and group sparsity norm.

Proximal Operator

Optimization of $h(x) + \lambda f(x)$
$h(x)$: differentiable, convex function (data fitting term)
$f(x)$ is some kind of regularizer term.

$$
\begin{aligned}
h(x) &= h(x_k) + \nabla_h(x_k)(x - x_k) + \frac{1}{2t}||x - x_k||_F^2 \\
h(x) + \lambda f(x) &= h(x) + \frac{1}{2t}||x - (x_k - t\nabla_h(x_k))||_2^2 \\
&= Prox_{\lambda,h}(x_k - t\nabla_h(x_k))
\end{aligned} \tag{6}
$$

loss function:

$$\sum_{j=1}^{L} \frac{c}{2} ||Y^j - X^j w^j||_F^2 + \sum_{i=1}^{K} \delta ||w_i||_2 \tag{7}$$

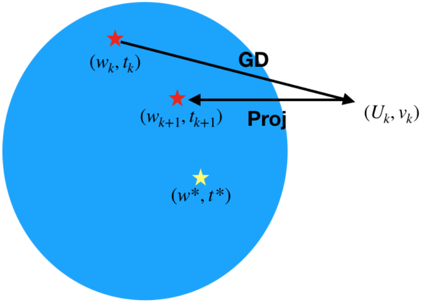
$$loss(W) + ||W||_{2,1} \tag{8}$$

Reformulation:

$$loss(W) + \rho \sum_{i=1}^{k} t_i \tag{9}$$
$$s.t. \ ||w_i||_2 < t_i (Feasible \ region : D)$$

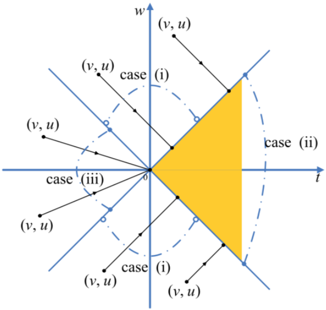
Basic Idea:

- one step gradient descent on $loss(W)$
- project the current solution back to feasible region


(w_k, t_k)
GD
(w_{k+1}, t_{k+1})
Proj
(U_k, v_k)
(w^*, t^*)

Projection

$$(W,t) = \arg\min_{(W,t)\in D} \frac{1}{2}||W-U||_F^2 + \frac{1}{2}||t-v||^2 = Prox_{I_D}(U,v) \qquad (10)$$



$$
W_i = \begin{cases} \frac{||U_i|| + v_i}{2||U_i||} U_i, & ||U_i|| > |v_i| \\ U_i, & ||U_i|| \leqslant v_i \\ 0, & ||U_i|| \leqslant -v_i \end{cases} \qquad t_i = \begin{cases} \frac{||U_i|| + v_i}{2}, & ||U_i|| > |v_i| \\ v_i, & ||U_i|| \leqslant v_i \\ 0, & ||U_i|| \leqslant -v_i \end{cases} \tag{11}
$$

Second Reformulation

$$\arg\min_{W} loss(W), \tag{12}$$

$$s.t. \quad ||W||_{2,1} < c \tag{13}$$

Lagrange Multiplier Framework:
Primal:

$$\arg\min_{W}\max_{\lambda \geqslant 0} loss(W) + \lambda(||W||_{2,1} - c) \tag{14}$$

Dual

$$\arg\max_{\lambda \geqslant 0}\min_{W} loss(W) + \lambda(||W||_{2,1} - c) \tag{15}$$

KKT Condition

Karush-Kuhn-Tucker

- $||W^*||_{2,1} < c$
- $\lambda^* > 0$
- $\lambda^*(||W^*||_{2,1} - c) = 0$
- $\nabla_W loss(W) + \nabla_W \lambda^*(||W^*||_{2,1} - c) = 0$

Suppose the current dual variable is λ^*, and current U:

$$W_i = Prox_{l_2}(U_i) = \frac{1}{2}||W_i - U_i||^2 + \lambda^*||W_i|| \tag{16}$$

Some useful facts for norm:

- For L_q norm, conjugate function: Indicator of unit ball of L_q norm, with $1/p + 1/q = 1$
- $w = Prox_{\lambda, L_p}(w) + \lambda Prox_{IL_q}(w/\lambda)$

Example:

- $Prox_{L_2}$, the dual norm L_2, the conjugate function:

$$\begin{cases} 0, ||x||_2 < 1 \\ \infty, otherwise \end{cases} \tag{17}$$

-

$$Prox_{L_q}(x) = \begin{cases} x, ||x||_2 < 1 \\ x/||x||_2, otherwise. \end{cases} \tag{18}$$

So, for objective function:

$$W_i = Prox_{l_2}(U_i) = \frac{1}{2}||W_i - U_i||^2 + \lambda^*||W_i|| \tag{19}$$

$$W_i^* = \begin{cases} (1 - \frac{\lambda^*}{||U_i||})U_i, if\, \lambda^* > 0, ||U_i|| > lambda^* \\ 0, \lambda^* > 0, ||U_i|| < \lambda^* \\ U_i, \lambda^* = 0 \end{cases} \tag{20}$$

Convergence Rate

Convergence rate: $O(1/k)$

Nesterov acceleration version: $O(1/k^2)$

ADMM

In single task setting: Solutions contain some group sparsity structure. Suppose $x \in \mathbb{R}^n$, $\{x_{g_i} \in \mathbb{R}^{n_i} : i = 1, 2, \cdots, s\}$ be the group structure for x

$$||x||_{2,1} = \sum_{i=1}^{s} w_i ||x_{g_i}||_2 \\ s.t. \quad Ax = b \tag{21}$$

Main idea:

- Introduce some auxiliary variable
- Split the big optimization problem into some subproblems
- Optimize each subproblems alternatively.

Procedures

Introduce new variables:

$$\min_{x,z} ||z||_{w,2,1} = \sum_{i=1}^{2} w_i ||z_{g_i}||_2 \tag{22}$$

$$s.t. \quad z = x, Ax = b \tag{23}$$

Augmented Lagrangian problem:

$$\min_{x,z} ||z||_{w,2,1} - \lambda_1^T (z - x) + \frac{\beta_1}{2} ||z - x||_2^2 - \lambda_2^T (Ax - b) + \frac{\beta_2}{2} ||Ax - b||_2^2, \tag{24}$$

where λ_1, λ_2 are multipliers and β_1, β_2 are penalty parameters.

x-subproblem:

$$\min_x \lambda_1^T x + \frac{\beta_1}{2}||z - x||_2^2 - \lambda_2^T Ax + \frac{\beta_2}{2}||Ax - b||_2^2, \tag{25}$$

$$\min_x \frac{1}{2}x^T(\beta_1 I + \beta_2 A^T A)x - (\beta_1 z - \lambda_1 + \beta_2 A^T b + A^T \lambda_2)^T x, \tag{26}$$

Strongly convex, reduces to linear system:

$$(\beta_1 I + \beta_2 A^T A)x = (\beta_1 z - \lambda_1 + \beta_2 A^T b + A^T \lambda_2). \tag{27}$$

z-subproblem:

$$\min_{z} \sum_{i=1}^{s} [w_i ||z_{g_i}||_2 + \frac{\beta_1}{2} ||z_{g_i} - x_{g_i} - \frac{1}{\beta_1}(\lambda_1)_{g_i}||_2^2] \tag{28}$$

multipliers update: gradient ascent

$$\begin{aligned} \lambda_1 &= \lambda_1 - \gamma_1 \beta_1 (z - x) \\ \lambda_2 &= \lambda_2 - \gamma_2 \beta_2 (Ax - b) \end{aligned} \tag{29}$$

Algorithm 1: Primal-Based ADM for Group Sparsity

Initialize $z \in \mathbb{R}^n$, $\lambda_1 \in \mathbb{R}^n$, $\lambda_2 \in \mathbb{R}^m$, $\beta_1, \beta_2 > 0$ and $\gamma_1, \gamma_2 > 0$;
while *stopping criterion is not met* **do**
    $x \leftarrow (\beta_1 I + \beta_2 A^T A)^{-1}(\beta_1 z - \lambda_1 + \beta_2 A^T b + A^T \lambda_2)$;
    $z \leftarrow Shrink(x + \frac{1}{\beta_1}\lambda_1, \frac{1}{\beta_1} w)$ (group-wise);
    $\lambda_1 \leftarrow \lambda_1 - \gamma_1 \beta_1 (z - x)$;
    $\lambda_2 \leftarrow \lambda_2 - \gamma_2 \beta_2 (Ax - b)$;