Optimization Methods in Machine Learning Lecture 15: Optimization approaches to sparse regularized regression

Katya Scheinberg

Lehigh University

Spring 2016

Lasso

$$\begin{aligned} & \min \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \\ = & \min \quad \frac{1}{2}x^T A^T Ax - b^T Ax + \lambda\|x\|_1 \qquad \text{constant ignored} \end{aligned}$$

- $A^T A$ is a $n \times n$ matrix, where n is the number of feature in one sample.
- Two methods to solve this.
  - Coordinate descent method
  - First order method

Coordinate descent method

- choose one variable x_i and one column A_i.
- Let $\bar{x}$ and $\bar{A}$ be the fixed part.
- $$\begin{aligned} &\min_{x_i} \quad \frac{1}{2}\|A_i x_i + \bar{A}\bar{x} - b\|^2 + \lambda|x_i| \\ = &\min_{x_i} \quad \frac{1}{2}(A_i^T A_i)x_i^2 + (\bar{x}^T \bar{A}^T A_i)x_i - b^T A_i x_i + \lambda|x_i| \end{aligned}$$

Soft-thresholding operator

- Equivalent problem

$$\min_{\theta} \quad \frac{1}{2}(\theta - r)^2 + \lambda|\theta|$$

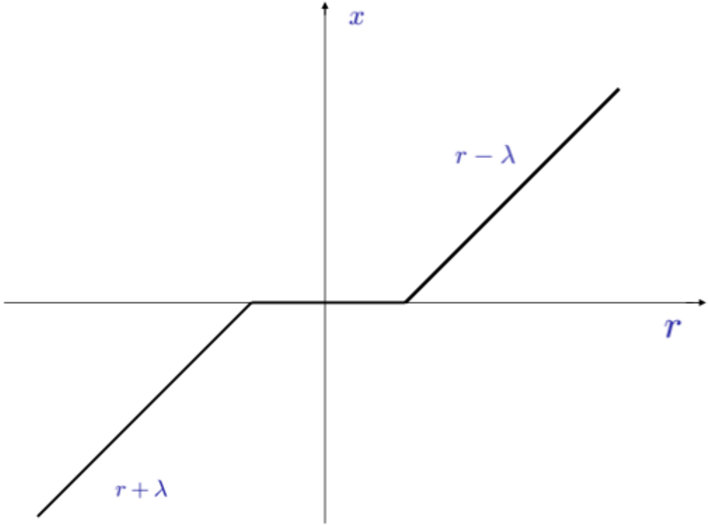
- Taking derivative

$$\nabla_\theta f(\theta) = \begin{cases} \theta - r - \lambda & \text{if } \theta < 0 \\ \theta - r + \lambda & \text{if } \theta > 0 \end{cases}$$

- Then we have

$$\theta^* = \begin{cases} 0 & \text{if } -\lambda \leq r \leq \lambda \\ r - \lambda & \text{if } r > \lambda \\ r + \lambda & \text{if } r < \lambda \end{cases}$$

Illustration



Specific step i

- On iteration i, we solve

$$\min \ \frac{1}{2}(x_i - r_i)^2 + \lambda |x_i|$$

- For r:

$$r_i = A_i^T(\bar{A}\bar{x} - b)/\|A_i\|$$

Algorithm

- Pick x_0
- For $k = 0, 1, 2, \ldots$
- compute $\bar{A}\bar{x}$, given Ax_i for given i
- compute $r_i = -A_i^T(\bar{A}\bar{x} - b)/\|A_i\|$
- compute $\mu_i = \lambda/\|A_i\|$
- Solve

$$\min_\theta \quad \frac{1}{2}(\theta - r_i)^2 + \mu_i|\theta|$$

- update

$$x_{k+1} = x_k + \theta e_i, \quad Ax_{k+1} = Ax_k + \theta A_i$$

- Avoid $n \times d$ operation. How? starting from 0 vector.

Computation cost

- $Ax_k - A_i x_i \approx \mathbb{O}(n)$
- $Ax_{k+1} = Ax_k + \theta A_i \approx \mathbb{O}(n)$

Problem approximation

- assume $f(x)$ is convex.

$$\min \qquad f(x) + \lambda \|x\|_1$$

Given x_k, linear approximation and bend it

$$Q(y) = f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2\mu} \|y - x_k\|^2 + \lambda \|y\|_1$$

- If there is no last $\lambda\|x\|_1$,

$$y^* = x_k - \mu_k \nabla f(x_k)$$

- Minimize Q(y)

$$\begin{aligned} \min_y \quad & f(x_k) + \frac{1}{2\mu_k} \|x_k - \mu_k \nabla f(x_k) - y\|^2 + \lambda \|y\|_1 \\ = \min_{y_i} \quad & \sum_i [\frac{1}{2\mu_k}((x_k)_i - \mu_k \nabla f(x_k) - y_i)^2 + \lambda |y_i|] \end{aligned}$$

- The last objective function is separable

Algorithm

- Pick y^1, μ_0
- for $k = 1, 2, ...,$ repeat
  - set $\mu_k = \mu_{k-1}$, compute

$$x^k = \text{argmin}_y f(y^k) + \frac{1}{2\mu_k}\|y^k - \mu_k \nabla f(y^k) - y\|^2 + \lambda\|y\|_1$$

  - Use line search backtracking, find μ_k such that

$$F(x^k) \leq Q_{\mu_k}(y^k, x^k)$$

where $F(x^k) = f(x^*) + \lambda|x^k|$
$Q_{y_k}(y, y_k) = f(y^k) + \nabla f(y_k)^T(y - y_k) + \frac{1}{2\mu}\|y - y_k\|^2 + \lambda\|y\|_1$

Extensions

Consider $g(y) = \lambda \sum_i \|y^i\|, \forall i$, then the problem

$$\min_{y} \quad \frac{1}{2}\|z - y\|^2 + \lambda \sum_i \|y^i\|$$

is equivalent to solve for each i separately because all variables y_i are independent, so

$$\min_{y_i} \quad \frac{1}{2}\|z^i - y^i\|^2 + \lambda\|y^i\|,$$

so

$$y^{i*} = \frac{r^i}{\|r^i\|}\max(0, \|r^i\| - \lambda)$$

Extensions

Consider one example that we have a group of identical features and want to study the effect,

$$Ax = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0.01 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = b,$$

where x_1, x_2 are identical.
Obviously, both $(1, 0, 1)^T$ and $(0.5, 0.5, 1.01)$ are solutions, however, if we solve the following problem,

$$\min_x \quad \frac{1}{2}\|Ax - b\|^2 + \lambda\|(x_1, x_2)^T\| + \lambda\|x_3\|,$$

we will find the unique optimal solution $(0.5, 0.5, 1.01)$.