# Regularization 1

Saturday, 18 February 2023 9:54 AM

## Example: $L^2$-regularized Linear Regression

① $R(w; X, y) = \frac{1}{2}\|Xw - y\|^2$

Least Squares Formula: $\hat{w} = \underbrace{(X^TX)^{-1}}_{H} X^Ty$

② $\tilde{R}(w; X, y) = \frac{1}{2}\|Xw - y\|^2 + \underbrace{\frac{1}{2}\alpha\|w\|^2}_{\text{regularizer}}$

Set $\nabla_w \tilde{R} = 0 \Rightarrow X^T(X\tilde{w} - y) + \alpha\tilde{w} = 0$

$$(X^TX + \alpha I)\tilde{w} = \underset{M\times N}{X^T}\underset{N\times 1}{y} \in \mathbb{R}^M$$

$$\Rightarrow \tilde{w} = (H + \alpha I)^{-1}X^Ty$$

Observe that $H$ is symmetric positive definite, so,

$\Rightarrow$ eigenvectors: $\{u_1, \dots, u_m\}$ → orthonormal basis of $\mathbb{R}^m$

eigenvalues: $\{\lambda_1, \dots, \lambda_m\}$, $(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m)$

$$X^Ty = \sum_{i=1}^{M} \beta_i u_i \qquad (\beta_i = u_i^T(X^Ty))$$

Hence, ① $\hat{w} = H^{-1}X^Ty = H^{-1}\sum_{i=1}^{M}\beta_i u_i$

$$= \sum_{i=1}^{M}\beta_i \underbrace{H^{-1}u_i}_{u_i/\lambda_i}$$

$$= \sum_{i=1}^{M}\frac{\beta_i}{\lambda_i}\cdot u_i$$

② $\tilde{w} = (H+\alpha I)^{-1} X^T y = (H+\alpha I)^{-1} \sum_{i=1}^{m} \beta_i u_i$

$= \sum_{i=1}^{m} \beta_i \underbrace{(H+\alpha I)^{-1} u_i}_{u_i/(\lambda_i+\alpha)}$

$= \sum_{i=1}^{m} \frac{\beta_i}{\lambda_i+\alpha} \cdot u_i$

## Example: $L^1$ vs $L^2$ Regularization

$R(\theta) = \frac{1}{2} \sum_{i=1}^{m} \lambda_i (\theta_i - \theta_i^*)^2$ $\quad \theta^* = (\theta_1^*, \ldots, \theta_m^*) \in \mathbb{R}^m$ is given.

$L^2$-regularized Problem:

$\tilde{R}(\theta) = R(\theta) + \frac{1}{2}\alpha\|\theta\|^2 = \frac{1}{2} \sum_{i=1}^{m} \left[ \lambda_i (\theta_i - \theta_i^*)^2 + \alpha \theta_i^2 \right]$

$\nabla_\theta \tilde{R} = 0 \Rightarrow \lambda_i (\theta_i - \theta_i^*) + \alpha \theta_i = 0$

$\therefore \theta_i = \frac{\lambda_i}{\lambda_i + \alpha} \theta_i^*$

- $\theta_i \approx \theta_i^*$ if $\alpha << \lambda_i$
- $\theta_i \approx 0$ if $\alpha >> \lambda_i$

$L^1$-Regularized Problem:

$\tilde{R}(\theta) = R(\theta) + \alpha \sum_{i=1}^{m} |\theta_i|$

$= \sum_{i=1}^{m} \underbrace{\left[ \frac{1}{2} \lambda_i (\theta_i - \theta_i^*)^2 + \alpha |\theta_i| \right]}_{R_i(\theta_i)}$

Case 1: $\theta_i > 0$ attains minimum.

$R_i(\theta_i) = \frac{1}{2} \lambda_i (\theta_i - \theta_i^*)^2 + \alpha \theta_i$

$\nabla_{\theta_i} R_i = 0 \Rightarrow \lambda_i (\theta_i - \theta_i^*) + \alpha = 0$

$$\theta_i = \theta_i^* - \alpha/\lambda_i$$

(valid if $\theta_i^* > \alpha/\lambda_i$)

Case 2: $\theta_i < 0$ attains minimum

$$R_i(\theta_i) = \frac{1}{2}\lambda_i(\theta_i - \theta_i^*)^2 - \alpha\theta_i$$

$$\nabla_{\theta_i} R_i = 0 \Rightarrow \lambda_i(\theta_i - \theta_i^*) - \alpha = 0$$

$$\theta_i = \theta_i^* + \alpha/\lambda_i$$

(valid if $\theta_i^* < -\alpha/\lambda_i$)

Case 3: $\theta_i = 0$

valid if $-\alpha/\lambda_i \leq \theta_i^* \leq \alpha/\lambda_i$

Example: Early stopping for a linear regression

$$R(\theta) = \frac{1}{2}\lambda(\theta - \theta^*)^2 \Rightarrow \nabla R(\theta) = \lambda(\theta - \theta^*)$$

Consider G.D on this:

$$\theta_{k+1} = \theta_k - \varepsilon\lambda(\theta_k - \theta^*)$$

$$= (1 - \varepsilon\lambda)\theta_k + \varepsilon\lambda\theta^*$$

$$\Rightarrow \theta_k = (1-\varepsilon\lambda)\theta_{k-1} + \varepsilon\lambda\theta^*$$

$$= (1-\varepsilon\lambda)\left[(1-\varepsilon\lambda)\theta_{k-2} + \varepsilon\lambda\theta^*\right] + \varepsilon\lambda\theta^*$$

$$= \cdots$$

$$= (1-\varepsilon\lambda)^k\theta_0 + \left[1-(1-\varepsilon\lambda)^k\right]\theta^*$$

Stop at iteration $\tau$:

$$\hat{\theta} = \theta_\tau = (1-\varepsilon\lambda)^\tau\theta_0 + \left[1-(1-\varepsilon\lambda)^\tau\right]\theta^*$$

Consider a "variant" of $L^2$-regularization

$$\tilde{R}(\theta) = \frac{1}{2}\lambda(\theta-\theta^*)^2 + \frac{1}{2}\alpha(\theta-\theta_0)^2$$

$$\nabla\tilde{R}(\theta)=0 \Rightarrow \lambda(\tilde{\theta}-\theta^*)+\alpha(\tilde{\theta}-\theta_0)=0$$

$$\tilde{\theta} = \frac{\alpha}{\lambda+\alpha}\cdot\theta_0 + \left(1-\frac{\alpha}{\alpha+\lambda}\right)\theta^*$$

$\tilde{\theta}=\hat{\theta}$ if $\frac{\alpha}{\alpha+\lambda} = (1-\varepsilon\lambda)^\tau$ i.e. $\alpha = \frac{\lambda(1-\varepsilon\lambda)^\tau}{1-(1-\varepsilon\lambda)^\tau}$