Regularization :

Good for avoiding $\begin{cases} \text{Data overfitting} \\ \text{multi\_collinearity} \end{cases}$

LS :

$$S(\underline{\theta}) = \sum_{i=1}^{n} (y_i - \underline{x}_i^T \underline{\theta})^2$$

$$= (\underline{y} - X\underline{\theta})^T (\underline{y} - X\underline{\theta})$$

Modified cost function :

Positive Regularizer

$$\tilde{S}(\underline{\theta}) = \underbrace{S(\underline{\theta})}_{\text{LS cost}} + \lambda \cdot \underbrace{r(\underline{\theta})}_{\text{regularization term}}$$

$$r(\underline{\theta}) = \begin{cases} \|\underline{\theta}\|_2^2 & \rightarrow \text{Ridge regression} \\ |\underline{\theta}|_1 & \rightarrow \text{Lasso regression} \\ L_p\text{-norm of } \underline{\theta} & \rightarrow \text{general form} \end{cases}$$

Example 1: Ridge regression

$$\tilde{S}(\underline{\theta}) = (\underline{y} - X\underline{\theta})^T (\underline{y} - X\underline{\theta}) + \frac{\lambda}{2} \|\underline{\theta}\|_2^2$$

Note that: $\|\underline{\theta}\|_2^2 = \sum \theta_i^2$.

$$\hat{\underline{\theta}}_{R} = \underset{\underline{\theta}}{\arg\min}\ \tilde{S}(\underline{\theta})$$

"Ridge"

Solving $\nabla_{\underline{\theta}} \tilde{S}(\underline{\theta}) = \underline{0}$, yields

$$\Rightarrow -X^T(\underline{y} - X\underline{\theta}) + \lambda \cdot \underline{\theta} = \underline{0}$$

$$\Rightarrow (X^T X + \lambda I_p)\,\underline{\theta} = X^T \underline{y}$$

$$\Rightarrow \hat{\underline{\theta}}_R = (X^T X + \lambda I_p)^{-1} X^T \underline{y}$$

When $\lambda = 0$, $\hat{\theta}_R = \hat{\theta}_{LS} = (X^T X)^{-1} X^T \underline{y}$

Example 2: Lasso regression

The cost function is:

$$\tilde{S}(\underline{\theta}) = S(\underline{\theta}) + \lambda \cdot |\theta|,$$

we often need to resort to numerical algorithms.

The solution $\hat{\theta}_L$ is sparse, which is favorable in many applications!

"$\underline{L}$asso"