Lecture 12: Regularized Regression

Course Teacher: Md. Shariful Islam Bhuyan

Regression

- *Multivariate linear regression*

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d = \mathbf{w}^{\mathrm{T}}\mathbf{x} \qquad \mathbf{w} \in \mathbb{R}^{\mathrm{d}}$$

$$L\big(y, f(\mathbf{x})\big) = \left(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}\right)^2$$

$$\mathbf{w}^* = \arg\min_f \mathcal{L}_{emp}(f) = \arg\min_{\mathbf{w}\in\mathbb{R}^{\mathrm{d}}} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{\mathrm{d}}} \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{d1} \\ 1 & x_{12} & x_{22} & \cdots & x_{d2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{dn} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \right\|^2 = \arg\min_{\mathbf{w}\in\mathbb{R}^{\mathrm{d}}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

Adaptive Basis Function

- What about this? Try it!

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2^3 + w_3\sqrt{x_3} + w_4 e^{-x_4} + w_5 \ln x_5 + \cdots + w_d \sin^{-1} x_d = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

- Still linear! Feature transformation. We do not know the origin.

$$f(\mathbf{x}) = w_0 + w_1 x_1 x_2$$

- Not linear!

$$f(\mathbf{x}) = \sum_{d=0}^{D} w_d \phi_d(\mathbf{x})$$

- Adaptive Basis Function Model

Analytical Solution

$$\frac{\partial \mathcal{L}_{emp}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = \frac{\partial}{\partial \mathbf{w}} (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{y}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right)(\mathbf{y} - \mathbf{X}\mathbf{w})$$

$$= \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}\right)$$

$$= \quad \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \left(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)^{\mathrm{T}} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}\right) = \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{2}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}\right)$$

$$= \frac{\partial}{\partial \mathbf{w}} \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{\partial}{\partial \mathbf{w}} \mathbf{2}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \frac{\partial}{\partial \mathbf{w}} \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

$$\frac{\partial}{\partial \mathbf{w}} \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = 2\mathbf{w}^{\mathbf{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} \qquad \left\{\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^{T}\mathbf{A}\mathbf{x} = \mathbf{x}^{T}(\mathbf{A} + \mathbf{A}^{T})\right\}$$

$$\frac{\partial}{\partial \mathbf{w}} \mathbf{w}^{\mathrm{T}} = \mathbf{I}_{d+1}$$

$$\frac{\partial \mathcal{L}_{emp}}{\partial \mathbf{w}} = -\mathbf{2}\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = 0, \qquad \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \mathbf{X}^{\mathrm{T}}\mathbf{y} = 0, \qquad \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y},$$

$$\mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Regularization

- Regularized risk function (Lagrange's multiplier)

$$\mathcal{L}_{reg}(h) = \mathcal{L}_{emp}(h) + \lambda \mathcal{R}(h)$$

- λ is a hyperparameter in this setting … scale conversion

$$h^*_{reg} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_{reg}(h)$$

$$\|\mathbf{w}\|_p = (|w_1|^p + |w_2|^p + \cdots + |w_d|^p)^{\frac{1}{p}} \quad p \geq 1$$

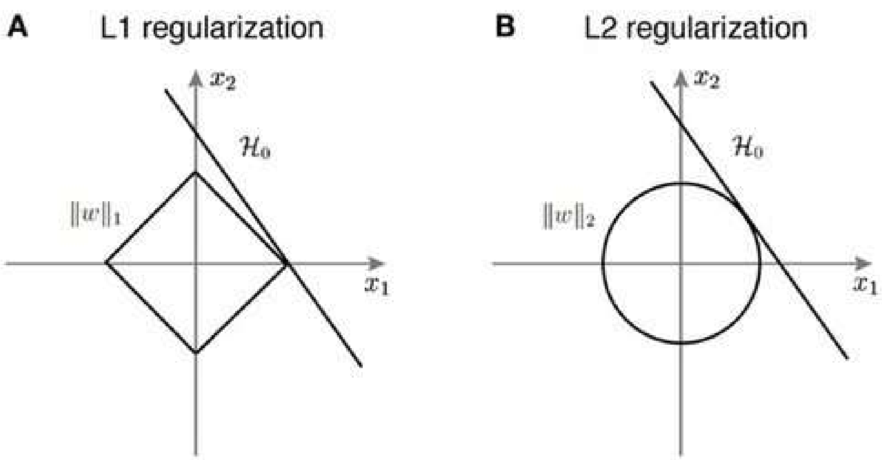
$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^{\mathrm{d}}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|_0$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^{\mathrm{d}}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|_1 \quad \text{(Lasso)}$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^{\mathrm{d}}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|_2^2 \quad \text{(Ridge)}$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^{\mathrm{d}}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda(\alpha \|\mathbf{w}\|_1 + (1-\alpha)\|\mathbf{w}\|_2), \alpha \in [0,1] \quad \text{(Elastic net)}$$

Sparsity L1 vs. L2 Regularization



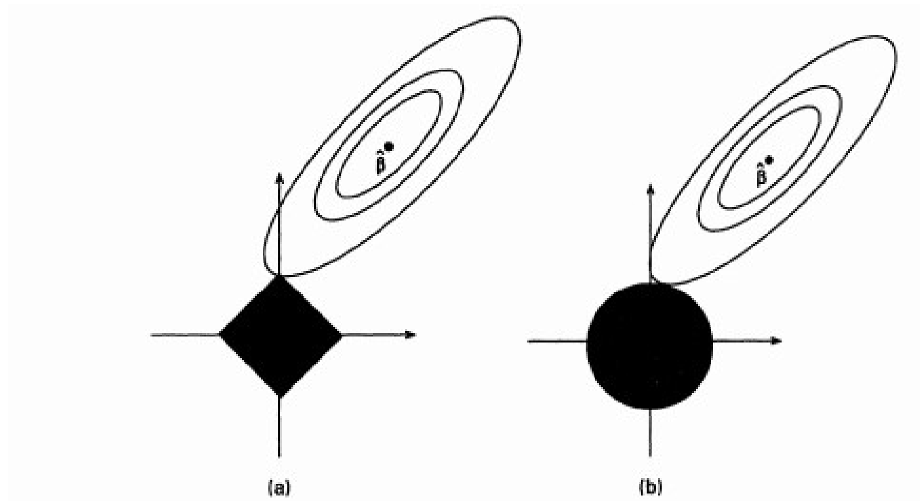


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

Analytical Solution

$$\frac{\partial \mathcal{L}_{emp}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\partial}{\partial \mathbf{w}} \lambda \|\mathbf{w}\|_2^2 = -\mathbf{2}\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 0$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}_{d+1}\right)\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}_{d+1}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathcal{L} = \|\mathbf{y} - \mathbf{X}\mathbf{w}\| = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)^2$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = -2 \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)\mathbf{x}_i$$

$$\frac{\partial}{\partial \mathbf{w}} \|\mathbf{w}\|_2^2 = \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \frac{\partial}{\partial w_1} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{bmatrix} \sum_{i=0}^{d} w_i^2 = 2\mathbf{w}$$

$$\mathcal{L}_{reg} = \|\mathbf{y} - \mathbf{X}\mathbf{w}\| + \lambda \|\mathbf{w}\|_2^2$$

$$\frac{\partial \mathcal{L}_{reg}}{\partial \mathbf{w}} = -2 \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)\mathbf{x}_i + 2\lambda\mathbf{w}$$

Numerical Solution: Gradient Descent

- Start with an initial value of

$$\mathbf{w} = \mathbf{w}^{(0)}$$

- Update $\mathbf{w}$ by moving along the gradient of the loss function $\mathcal{L}$

$$\mathbf{w}^{(\mathrm{t})} = \mathbf{w}^{(\mathrm{t}-1)} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}}$$

- Repeat until converge