Lecture 14: Regularized Regression

Course Teacher: Md. Shariful Islam Bhuyan

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

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

- Norm computes a measure of non-zero distance

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

- L0 norm: cardinality function, tricky due to zero

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

L1 and L2 Regularization

- Lasso or L1 norm: *sparsity*

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

- Ridge or L2 norm: *favor small co-efficient*

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

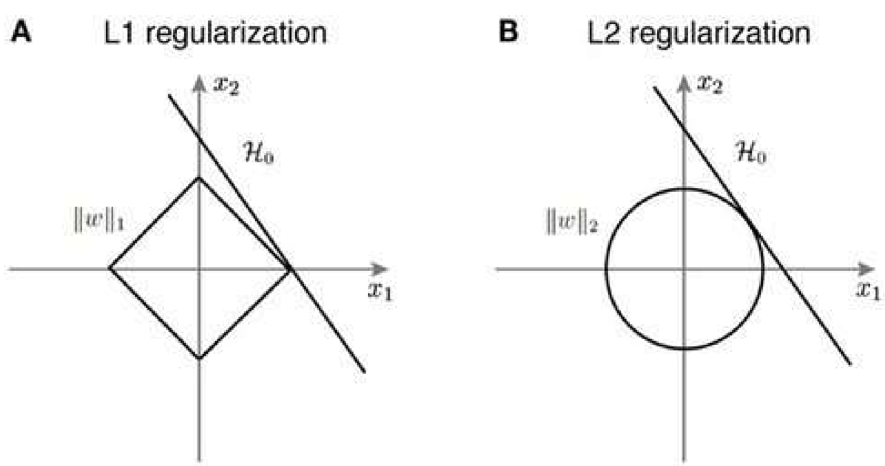
- Elastic net: *affine combination of L1 and L2*

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

- For 2-D constraint

$$|w_1|^2 + |w_2|^2 \le c$$

Sparsity L1 vs. L2 Regularization



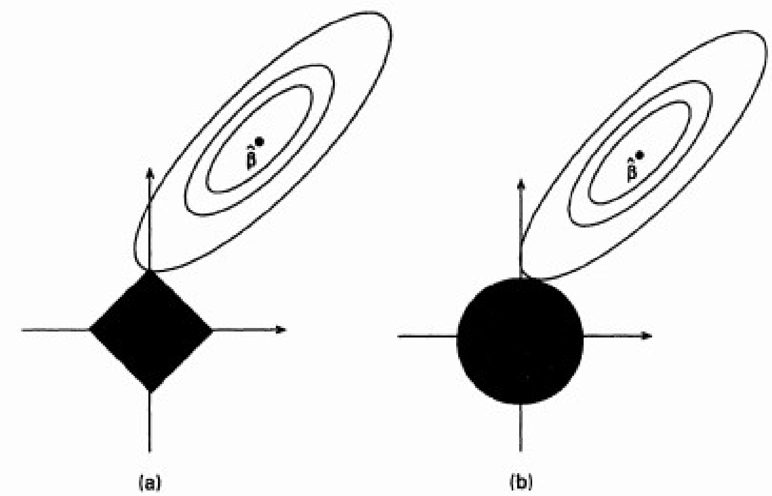


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

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