Student Name: LI JIAYI
Student ID: 1155188890

Regression Learning

Review

1. **Linear Regression**

   Linear Regression models the linear relationship between feature (explanatory) variables $\boldsymbol{x} \in \mathbb{R}^D$ to response $y \in \mathbb{R}$

   $$y = b + \boldsymbol{w}^\top \boldsymbol{x} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

   Denote the predicted value

   $$f_{\boldsymbol{w},b}(\boldsymbol{x}) = b + \boldsymbol{w}^\top \boldsymbol{x}$$

   Collect N linear independent sample $\boldsymbol{S} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$

   The goal is to choose the appropriate loss function to minimise the error.

   (a) **Mean Square Error**(MSE)

   $$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N} \varepsilon_n^2 = \frac{1}{N}\sum_{n=1}^{N} \left\{y_n - f_{\boldsymbol{w},b}(\boldsymbol{x}_n)\right\}^2$$

   define $\tilde{\boldsymbol{w}} = (b, \boldsymbol{w}^\top)^\top \in \mathbb{R}^{D+1}$ and $\tilde{\boldsymbol{X}} = (\boldsymbol{1}|\boldsymbol{X}) \in \mathbb{R}^{N\times(1+D)}$

   $$\mathcal{E} = \frac{1}{N}(\boldsymbol{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}})^\top(\boldsymbol{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}})$$

   $$\nabla_{\tilde{\boldsymbol{w}}}(\mathcal{E}) = \frac{1}{N}\Big(-2(\boldsymbol{y}^\top\tilde{\boldsymbol{X}})^\top + 2(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})\tilde{\boldsymbol{w}}\Big)$$

   (b) **Mean Absolute Error**(MAE)

   $$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N} |\varepsilon_n| = \frac{1}{N}\sum_{n=1}^{N} |y_n - f_{\boldsymbol{w},b}(\boldsymbol{x}_n)|$$

2. **Regularization**

   Regularization is a collective term that encompasses methods that force the learning algorithm to return a less complex model.

   In general, a loss function with a $\mathcal{L}^q$ regularizer

   $$\mathcal{L}^q: \quad \underset{\boldsymbol{w},b}{min}\left(\frac{1}{N}\sum_{n=1}^{N}\left\{y_n - f_{\boldsymbol{w},b}(\boldsymbol{x}_n)\right\}^2 + \lambda\sum_{d=1}^{D}|w_d|^q\right)$$

   (a) LASSO($\mathcal{L}^1$): good at feature selection, by identifying which features are essential for prediction or not, and the trained model possesses a higher explainable nature.

   (b) Ridge($\mathcal{L}^2$): Maximizes the performance of the model, and the underlying differentiability ensures the convenient use of various gradient descent method for parameter estimation.

   In practice, we can augment several Lq regularizers to loss function. *Elastic net* combines $\mathcal{L}^1$ and $\mathcal{L}^2$ regularizations with new penalty

   $$\lambda\sum_{d=1}^{D}(1-\alpha)\,|w_d| + \alpha w_d^2, \quad \alpha \in (0,1)$$