Foundation of Machine Learning (IT 582)

Autumn 2022

Pritam Anand

Least Squares Loss function and Normal Noise

Given the training set $\{(x_i, y_i) : x_i \in \mathbf{R}^n, y_i \in \mathbf{R} \text{ for } i = 1, 2, ..l, \}$, the regularized Least Squares Regression model solves the optimization problem

$$\min_{(w)} \ \frac{\lambda}{2} w^T w + \frac{1}{2} \sum_{i=1}^{l} (y_i - (w^T \phi(x_i)))^2 \tag{1}$$

where $\phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_1(x) \\ .... \\ ..... \\ \phi_m(x) \end{bmatrix}$ and $\phi_j(x)$ for $j = 1, 2, ..., m$, are m different basis functions.

In next few lectures, we shall discuss the significance and limitation of the Least Squares loss function used in the Least Squares Regression models. The main advantage of the Least Square loss function is that the solution of the Least Squares-based regression model can be easily obtained.

As we have discussed earlier, the Least Squares loss function is a smooth and convex loss function. It makes the optimization problem (1) convex and unconstrained optimization problem. For the solution of the optimization problem (1), we need to set the gradient of its objective function with respect to w equal to zero which finally requires the solution of a system of equations.

Now, we shall talk about the statistical significance of the Least Squares loss function in the maximum likelihood sense. Our basic assumption behind the regression model is

$$y_i = f(x_i) + \epsilon_i, \tag{2}$$

where, $E(\epsilon_i) = 0$ and the true estimate of $f(x_i)$ is $E(y/x_i)$. In our Least Squares Regression model, we estimate $f(x_i) = w^T \phi(x_i)$. Let us consider the assumption (which is very relevant in view of the Central Limit Theorem) that the conditional distribution y/x_i is Normal with mean $E(y/x_i)$ and variance σ. Since our estimate for $E(y/x_i)$ is $w^T \phi(x_i)$ therefore, we can write for a given value of x_i, $y \sim N(w^T \phi(x_i), \sigma)$. We can also obtain that $\epsilon_i = y_i - w^T \phi(x_i) \sim N(0, \sigma)$.

In our regression problem, a selected sample $T = \{(x_i, y_i) : x_i \in \mathbf{R}^n, y_i \in \mathbf{R} \text{ for } i = 1, 2, ..l, \}$ is only provided from the actual population. Let us attempt to obtain the optimal value of parameter $w \in \mathbb{R}^m$ by maximizing the probability

of obtaining the sample T in the maximal likelihood sense. With the assumption that (x_i, y_i) are independent and identically distributed random variables, we maximize the probability of obtaining the response values $(y_1, y_2, ...y_l)$ by assuming that $y/x_i \sim N(w^T\phi(x_i), \sigma)$. For this, we should consider the optimization problem

$$\max\ p(y_1, y_2, ..., y_l)/(x_1, x_2, ..., x_l)) = \max_w \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi}\sigma} e^{-} \frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2},$$

which is equivalent to the following problem

$$\begin{aligned}\max_w log\ (\prod_{i=1}^{l} \frac{1}{\sqrt{2\pi}\sigma} e^{-} \frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2}) &= \max_w\ (\sum_{i=1}^{l} log(\frac{1}{\sqrt{2\pi}\sigma} e^{-} \frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2})) \\ &= \max_w\ \textstyle\sum_{i=1}^{l} log(\frac{1}{\sqrt{2\pi}\sigma}) - \sum_{i=1}^{l} \frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2}. \end{aligned} \tag{3}$$

Since σ is constant in the optimization problem (3), therefore, we can easily reduce it as

$$\begin{aligned}&\max_w - \sum_{i=1}^{l} (y_i - w^T\phi(x_i))^2. \\ &= \min_w \sum_{i=1}^{l} (y_i - w^T\phi(x_i))^2. \end{aligned} \tag{4}$$

We can note that the optimization problem (4) is minimizing the Least Square Loss function. Now, we can realize that if the conditional distribution y/x is normal, then the maximization of the probability of obtaining the sample T is equivalent to the minimization of the Least Squares Loss function. **In other words, if the noise is normally distributed, then the Least Square loss function is the optimal loss function to be used in the maximal likelihood sense.**

In the next lecture, we shall talk about the limitations of the Least Squares loss function.

Thanks
. Pritam.