Linear Regression

Shumin An

March 2019

1 Introduction

Assume our dataset:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

$$X = (x_1, x_2, \cdots, x_N)^T, Y = (y_1, y_2, \cdots, y_N)^T$$

Linear Regression is very intuitive. Imagine a two-dimension space, the goal is to find a line which can precisely separate all the points in the space. First comes to out mind is to calculate the distance to each point and solve a optimization problem.

2 Least Square Method

We could use L2-norm to define the loss function. So the loss function here means the sum of the distance to each point.

$$L(w) = \sum_{i=1}^{N} ||w^T x_i - y_i||_2^2 \tag{1}$$

Further we could write it in the matrix format. (In this way, we could throw away the Σ, so it becomes more convenient to compute the derivatives)

$$\begin{aligned} L(w) &= (w^T x_1 - y_1, \cdots, w^T x_N - y_N) \cdot (w^T x_1 - y_1, \cdots, w^T x_N - y_N)^T \\ &= (w^T X^T - Y^T) \cdot (Xw - Y) = w^T X^T X w - Y^T X w - w^T X^T Y + Y^T Y \\ &= w^T X^T X w - 2 w^T X^T Y + Y^T Y \end{aligned}$$

We need to minimize the loss function:

$$\begin{aligned} \hat{w} = \underset{w}{argmin}\, L(w) &\longrightarrow \frac{\partial}{\partial w} L(w) = 0 \\ &\longrightarrow 2X^T X \hat{w} - 2X^T Y = 0 \\ &\longrightarrow \hat{w} = (X^T X)^{-1} X^T Y = X^+ Y \end{aligned}$$

3 MLE

Actually, the Least Square Method solve the problem from geometric perspective. From statistical way, we could have MLE as follows. Basically, it's hard to let all the sample points fall on the line $w^T x$, so assume there's a bias obeys Gaussian Distribution, here we have: $y = w^T x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$, $y \sim \mathcal{N}(w^T x, \sigma^2)$ .Then we can compute the MLE, the result is same as Least Square Method:

$$\begin{aligned} L(w) = \log p(Y|X, w) &= \log \prod_{i=1}^{N} p(y_i|x_i, w) \\ &= \sum_{i=1}^{N} \log(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}) \\ &= \sum_{i=1}^{N} (\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - w^T x_i)^2}{2\sigma^2}) \\ \underset{w}{argmax}\, L(w) &= \underset{w}{argmin} \sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{2\sigma^2} \end{aligned}$$

The calculation next is same as LSE part.

4 MAP

Assume the prior distribution: $w \sim \mathcal{N}(0, \sigma_0^2)$ So we have:

$$\begin{aligned} \hat{w} = \underset{w}{argmax}\, p(w|Y) &= \underset{w}{argmax}\, p(Y|w)p(w) \\ &= \underset{w}{argmax} \log p(Y|w)p(w) \\ &= \underset{w}{argmax} (\log p(Y|w) + \log p(w)) \\ &= \underset{w}{argmax} \sum_{i=1}^{N} [-(\frac{(y_i - w^T x_i)}{\sigma^2}) - \frac{||w||_2^2}{\sigma_0^2}] \\ &= \underset{w}{argmin} \sum_{i=1}^{N} [(y_i - w^T x_i)^2 + \frac{\sigma^2}{\sigma_0^2} w^T w] \end{aligned}$$

The hyperparameter σ_0 is related to the regularization, which is really fantastic when I saw it in the first time!.

5 Regularization

We always heard about the words "over-fitting". And here's what I think, when we compute the partial derivative, we need to compute the inverse matrix

of $X^T X$, so what if it dosen't have a inverse matrix. We know that only non-singular matrix has its inverse matrix. So if the number of samples(rows) much smaller than features(columns), the $X^T X$ matrix may not be non-singular.

There are two classical regularization form:

$$\begin{aligned} L1 &: \underset{w}{argmin}\, L(w) + \lambda ||w||_1 \\ L2 &: \underset{w}{argmin}\, L(w) + \lambda ||w||_2^2 \end{aligned} \tag{2}$$

And $\hat{w}$ for L2 regularization:

$$\begin{aligned} \hat{w} = \underset{w}{argmin}\, L(w) + \lambda w^T w &\longrightarrow \frac{\partial}{\partial w} L(w) + 2\lambda w = 0 \\ &\longrightarrow 2X^T X\hat{w} - 2X^T Y + 2\lambda\hat{w} = 0 \\ &\longrightarrow \hat{w} = (X^T X + \lambda \mathbb{I})^{-1} X^T Y \end{aligned}$$