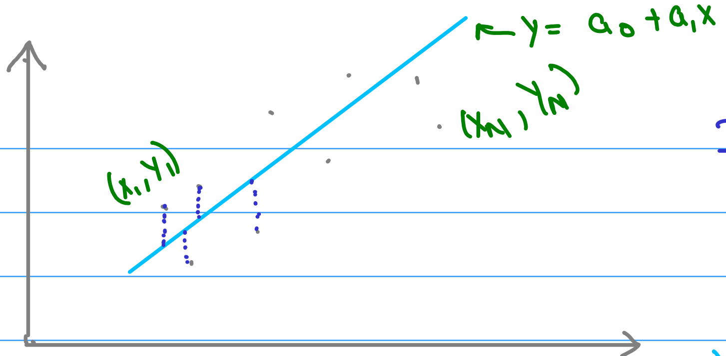
Linear Regression Derivation



In linear regression, we try to get the best fit line, which is a straight line.

Standard approach to solve linear regression

Same approach in all statistical learning →

1) Deriving closed form solution
2) Applying gradient descent

In linear regression, the model consists of linear function

$$y = \sum_j w_j x_j + b \quad \text{— } (1)$$

Loss function is defined as

$$\mathcal{L}(y,t) = \frac{1}{2}(y-t)^2 \quad \text{— } (2)$$

Considering (2) in (1), we get cost function

$$C(w_1, w_2, \dots, w_D, b) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y^{(i)}, t^{(i)})$$

$$\text{cost function} = \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2$$

$$= 1/2N \sum_{i=1}^{N}\Big(\underbrace{\sum_j w_j x_j^{(i)}} + b - t^{(i)}\Big)^2 \quad \text{— } (3)$$

— Here the choice of w_i & b is optimized to reduce C.

For a optimization problem, a good place to start is to compute the partial derivative of cost function

(3) ⇒ Applying chain rule

$$\delta C/\delta w_j = 1/N \sum_{i=1}^{N} x_j^{(i)}\Big(\sum_{j'} w_{j'} x_{j'}^{(i)} + b - t^{(i)}\Big)$$

$$\delta C/\delta b = 1/N \sum_{i=1}^{N}\Big(\sum_{j'} w_{j'} x_{j'}^{(i)} + b - t^{(i)}\Big)$$

The following can be rewritten as

$$\delta C / \delta w_j = (1/N) \sum_{i=1}^{N} x_j^{(i)} \left(y^{(i)} - t^{(i)} \right) \quad \text{—} (5)$$

$$\delta C / \delta b = 1/N \sum_{i=1}^{N} y^{(i)} - t^{(i)}$$

If a function is differentiable, then, $\delta f / \delta x_i$ is 0 at minimum.

$\delta f / x_i \rightarrow -ve$, increase x_i slightly
$\delta f / x_i \rightarrow +ve$, decrease x_i slightly

Critical point $\Rightarrow$ partial derivative is zero

So, equating (5), to 0, to get the critical point value,

$$\delta c / \delta w_j = 1/N \left(\sum_{i=1}^{N} x_j^{(i)} \left(y^{(i)} - t^{(i)} \right) \right) = 0 \quad \text{—} \underline{\text{Eq 6}}$$

Solving for above will provide the relevant weight

The above is direct method.

Gradient Descent Method

In gradient descent, the direction of steepest ascent of a f^n is determined. Gradient descent considers partial derivative of the variables.

$$\frac{\delta \mathcal{E}}{\delta w} = \begin{pmatrix} \frac{\delta \mathcal{E}}{\delta w_1} \\ \vdots \\ \delta \mathcal{E} / \delta w_D \end{pmatrix}$$

Using update parameter,

$$w \leftarrow w - \alpha \frac{\delta \mathcal{E}}{\delta w}$$