09/09/2024

Lecture: Linear Regression & Optimization

Pg 6. $\mathcal{L}(y,t) = \frac{1}{2}(y-t)^2$

i = index of the sample/vector (there are N samples in the training set)

Cost $\mathcal{J}(w,b)$ = average of the sum of all $\mathcal{L}$'s.

$$= \frac{1}{N}\sum_{i=1}^{N} \boxed{\mathcal{L}_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(y^i - t^i)^2$$

loss calculated by the ground truth and the prediction of the i^{th} vector

j = index of the feature (regarding one sample)

$$= \frac{1}{2N}\sum_{i}^{N}(w^T x^i + b - t^i)^2$$

$$= \frac{1}{2N}\sum_{i}^{N}\Big(\sum_{j}^{D}\boxed{w_j}\,x^i_j + b - \boxed{t^i}\Big)^2$$

w_j: Weight of the j^{th} feature within a vector/sample

x^i_j: Value of the j^{th} feature in the i^{th} sample/vector in the training set

t^i: ground truth/target of the ith vector

Pg 10.

$$X = \begin{bmatrix} \boxed{\begin{matrix}1\\1\\\vdots\\1\end{matrix}} & \begin{matrix}[x^{(1)}]^T\\ [x^{(2)}]^T\\ \vdots \\ [x^{(N)}]^T\end{matrix}\end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} b \\ w_1 \\ w_2 \\ \vdots \\ w^N \end{bmatrix} \in \mathbb{R}^{D+1}$$

Why we do this: Incorporate the bias b into the manipulation of matrix of all samples by essentially creating a new "dummy"/"placeholder" feature consisting of 1's. This allows the weight vector to incorporate b into it, for better vectorization.

Essentially, $y = w^T X + b\mathbf{1} = Xw + b\mathbf{1}$ becomes $y = Xw$, to get rid of b.

We want to get rid of b for better vectorization.

P.g.11 "To show that z^* minimizes $f(z)$, show that $\forall z$, $f(z) \geq f(z^*)$"

this line refers to the loss function (f is a analogy to $\mathcal{L}$). To show that a particular z^* (* means we are talking about a particular one / optimal one), in which z^* is an analogy to parameters $\vec{w}$ and b, minimizes $f/\mathcal{L}$, you need to show that for all combinations of (w, b), or for all possible z, $f(z^*)$ is the smallest (that z^* minimizes the loss function).

P.g.12. ①

$$\frac{\partial y}{\partial w_j} = \frac{\partial}{\partial w_j}\Big(\sum_{\underline{j'}}^{D} w_{j'} x_{j'} + b\Big)$$

→ looping through all dimensions

$$= \frac{\partial}{\partial w_j}(w_0x_0 + w_1x_1 + \cdots + w_jx_j + \cdots + w_Dx_D + b)$$

$$= \frac{\partial}{\partial w_j}(\cancel{w_0}x_0)^{0} + \cdots + \frac{\partial}{\partial w_j}(\cancel{w_j}x_j)^{1 \cdot x_j} + \cdots + \frac{\partial}{\partial w_j}(\cancel{w_0}x_0)^{0} + \frac{\partial}{\partial w_j}\cancel{b}^{0}$$

$$= x_j.$$

②

$$\frac{\partial y}{\partial b} = \frac{\partial}{\partial b}\Big(\sum_{j'}^{D} w_{j'} x_{j'} + b\Big) = \frac{\partial}{\partial b}(\cancel{w_0x_0})^{0} + \cdots + \frac{\partial}{\partial b}(\cancel{w_D}x_D)^{0} + \frac{\partial}{\partial b}b^{1} = 1$$

③

$$\frac{\partial \mathcal{L}}{\partial w_j} = \frac{\partial \mathcal{L}}{\partial y} \cdot \underbrace{\boxed{\frac{\partial y}{\partial w_j}}}_{}^{x_j} = \frac{d}{dy}\Big(\frac{1}{2}(y-t)^2\Big)x_j = (y-t)x_j$$

④

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial b} = \frac{d}{dy}\Big(\frac{1}{2}(y-t)^2\Big) \cdot 1 = (y-t)$$

☆ Remark: we calculate EQ ① and EQ ② to calculate EQ ③ & EQ ④, as ③ & ④ relate to how w_j & b affect the loss.

☆ Remark: We apply the chain rule to get to the expanded form of ③ & ④ because we cannot do $\frac{\partial \mathcal{L}}{\partial w_j}$ & $\frac{\partial \mathcal{L}}{\partial b}$ directly — $\mathcal{L}$ is not defined (directly) using w_j and b.