Deep Learning

Homework 2

Q 1) ~~J(w,b)~~ Equation 6.1 in the Deep Learning book is as follows

$$J(\theta) = \frac{1}{4} \sum_{x \in X} (f^*(x) - f(x;\theta))^2$$

When we choose a linear model with θ consisting of w, b the form of model $f(x;\theta)$ is defined to be

$$f(x; w, b) = x^T w + b$$

$$\therefore J(\theta) = \frac{1}{4} \sum_{x \in X} (f^*(x) - (x^T w + b))^2$$

Differentiating $J(\theta)$ w.r.t w

$$\nabla_w J(\theta) = \frac{\partial}{\partial w}\left[\frac{1}{4} \sum_{x \in X} \nabla_w (f^*(x) - (x^T w + b))^2\right]$$

$$= -\frac{1}{2} \sum_{x \in X} x (f^*(x) - x^T w - b).$$

$\hookrightarrow$ (1).

for XOR function

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \qquad X^T = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

$$f^*(x) = y = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

replacing X and y in eqn (1)

$$= -\frac{1}{2} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \left(\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \begin{bmatrix} b \\ b \\ b \\ b \end{bmatrix} \right)$$

Let us set gradient to 0 to get optimal solution

$$0 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \left(\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ w_2 \\ w_1 \\ w_1+w_2 \end{bmatrix} - \begin{bmatrix} b \\ b \\ b \\ b \end{bmatrix} \right)$$

$$= \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \left(\begin{bmatrix} -b \\ 1-\omega_2-b \\ 1-\omega_1-b \\ -\omega_1-\omega_2-b \end{bmatrix} \right)$$

$$0 = \begin{bmatrix} 1-\omega_1-b-\omega_1-\omega_2-b \\ 1-\omega_2-b-\omega_1-\omega_2-b \end{bmatrix}$$

$$1-2\omega_1-\omega_2-2b = 0 \rightarrow (2)$$
$$1-\omega_1-2\omega_2-2b = 0 \rightarrow (3)$$

Differentiating $J(\theta)$ w.r.t b.

$$\nabla_b J(\theta) = \frac{1}{4} \sum_{X \in x} \nabla_b \left(f^*(x) - (x^T\omega + b) \right)^2$$

$$\nabla_b J(\theta) = -\frac{1}{2} \sum_{X \in x} \left[f^*(x) - (x^T\omega + b) \right]$$

$\hookrightarrow (4)$

Similar to $\nabla \omega$ we take the values of XOR function and setting gradient to 0

Replacing the values in eqn (4).

$$0 = \left(\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \begin{bmatrix} b \\ b \\ b \\ b \end{bmatrix} \right)$$

$$0 = \left(\begin{bmatrix} -b \\ 1 - w_2 - b \\ 1 - w_1 - b \\ -w_1 - w_2 - b \end{bmatrix} \right)$$

$$-b + 1 - w_2 - b + 1 - w_1 - b - w_1 - w_2 - b = 0$$

$$-2w_1 - 2w_2 - 2 = 4b$$

$$2(-w_1 - w_2 - 1) = 4b$$

$$-w_1 - w_2 - 1 - 2b = 0 \rightarrow (5).$$

Subtracting eqn (5) from eqn (2).

$$\cancel{1} - 2w_1 - \cancel{w_2} - 2b = 0$$

$$- \quad \cancel{1} - w_1 - \cancel{w_2} - 2b = 0$$

$$0 - w_1 - 0 - 0 = 0$$

$$\boxed{w_1 = 0}$$

Subtracting eqn 5 from eqn 3

$$1 - w_1 - 2w_2 - \cancel{2b} = 0$$
$$-\cancel{1} - \cancel{w_1} - w_2 - \cancel{2b} = 0$$

$$-w_2 = 0$$

$$\boxed{w_2 = 0}$$

putting w_1 & w_2 values in eqn 5

$$1 - 0 - 0 = 2b$$

$$b = 1/2$$

$$\boxed{b = 0.5}$$

$$\therefore \boxed{\begin{array}{c} w_1 = w_2 = 0 \\ b = 0.5 \end{array}}$$

Problem 2 – L2-regularized Linear Regression via Stochastic Gradient Descent

Test Set unregularized MSE: 723.6215674228685

Problem 3 – Logistic Sigmoid Identities

a) Prove that $\sigma(-x) = 1 - \sigma(x)\ \forall x$.

$$\sigma(-x) = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-(-x)}} = \frac{1}{1+e^{x}} * (e^{x}) = \frac{e^{x}}{e^{x}+1}$$

Use equation: $\frac{a}{b} = \frac{a+1}{b} - \frac{1}{b}, a = e^{x}, b = e^{x}+1$

$$= \frac{e^{x}+1}{e^{x}+1} - \frac{1}{e^{x}+1} = 1 - \sigma(x)$$

b) Prove $\sigma'(x) = \frac{\partial \sigma}{\partial x}(x) = \sigma(x)(1-\sigma(x)) \quad \forall x$

$$\sigma'(x) = \frac{1}{1+e^{-x}} = \frac{d}{dx}\,\sigma(x) = \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = \frac{d}{dx}(1+e^{-x})^{-1}$$

$$= -(1+e^{-x})^{-2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^{2}} = \frac{1}{1+e^{-x}} * \frac{e^{-x}}{1+e^{x}}$$

$$= \sigma(x) * \sigma(-x)$$

Problem 4 – Regularization to Encourage Symmetry

W = [w1, w2] and represent the weights of the 1 x 2 “image” pixels, respectively. Optimally we want (w1 – w2)^2 to be small so the weights should have similar values. Mathematically:

$$\frac{\alpha}{2}\left(w^{T}\begin{bmatrix}1\\-1\end{bmatrix}\right)([1\ -1]w) = \frac{\alpha}{2}w^{T}\begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix}w$$

Thus, the matrix S that best encourages the weights to be symmetric is $S = \begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix}$

~~Q5) P(y+~~

Q5) $P(y|x) = \mathcal{N}(\mu = x^T\omega, \sigma^2)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - x^T\omega)^2}{2\sigma^2}\right)$$

Training dataset $D = \{x^{(i)}, y^{(i)}\}_{i=1}^{n}$

$$\mu = x^T\omega$$

$$P(D|\omega,\sigma^2) = \prod_{i=1}^{n} P(y^{(i)}/x^{(i)}, \omega, \sigma^2)$$

$$\ln P(D|\omega,\sigma^2) = \ln \prod_{i=1}^{n} P(y^{(i)}/x^{(i)}, \omega, \sigma^2)$$

$$= \sum_{i=1}^{n} \ln P(y^{(i)}|x^{(i)}, \omega, \sigma^2)$$

$$= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(y - x^T\omega)^2}{2\sigma^2}\right)}\right)$$

$$= \sum_{i=1}^{n} \left[\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \ln e^{\left(-\frac{(y - x^T\omega)^2}{2\sigma^2} \right)} \right]$$

$\hookrightarrow$ 1

Differentiating with ω.

$$\nabla_\omega \ln P(D, \omega, \sigma^2) = \nabla_\omega \sum_{i=1}^{n} \left(\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \right)$$

$$+ \nabla_\omega \sum_{i=1}^{n} \left(\frac{-(y - x^T\omega)^2}{2\sigma^2} \right)$$

$$= 0 + \sum_{i=1}^{n} \left(\frac{-1}{2\sigma^2} \right) - 2x(y - x^T\omega)$$

$$\nabla_\omega \ln P(D | \omega, \sigma^2) = -\frac{1}{\sigma^2} \left[\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x x^T\omega \right]$$

To get optimized value set the above equ to 0.

$$0 = -\frac{1}{\sigma^2}\left[\sum_{i=1}^{n} xy - \sum_{i=1}^{n} xx^T\omega\right]$$

$$\omega \sum_{i=1}^{n} xx^T = \sum_{i=1}^{n} xy$$

$$\boxed{\omega = \left(\sum_{i=1}^{n} x^{(i)} x^{T(i)}\right)^{-1}\left(\sum_{i=1}^{n} x^{(i)} y^{(i)}\right)}$$

from equ (1).

$$\ln P(D|\omega,\sigma^2) = \sum_{i=1}^{n}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln e^{\left(-\frac{(y-x^T\omega)^2}{2\sigma^2}\right)}\right]$$

Differentiating with σ

$$\nabla_\sigma \ln P(D \mid \omega, \sigma^2) = \sum_{i=1}^{n} \nabla_\sigma \left(\ln(1) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) \right)$$

$$+ \sum_{i=1}^{n} \nabla_\sigma \left(\frac{-(y - x^T\omega)^2}{2\sigma^2} \right)$$

$$= \sum_{i=1}^{n} \left(0 - 0 - \frac{1}{2}\frac{1}{\sigma^2} 2\sigma \right) + \sum_{i=1}^{n} \frac{(y - x^T\omega)^2}{\sigma^3}$$

$$\nabla_\sigma \ln P(D \mid \omega, \sigma^2) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y - x^T\omega)^2$$

Set the above to 0 for optimal solution

$$\frac{n}{\sigma} = \frac{1}{\sigma^3} \sum_{i=1}^{n} (y - x^T\omega)^2$$

$$\boxed{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x^T\omega - y)^2}$$