Contents

Problem 1: XOR Problem

① XOR →

given → $\begin{array}{ccccc} & x^{(1)} & x^{(2)} & x^{(3)} & x^{(4)} \\ \mathbf{X} = & \left[\begin{matrix} 0 \\ 0 \end{matrix}\right. & \begin{matrix} 0 \\ 1 \end{matrix} & \begin{matrix} 1 \\ 0 \end{matrix} & \left.\begin{matrix} 1 \\ 1 \end{matrix}\right] \end{array}$

give → $y = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^T$ $(f^*(x))$
(ground truth)

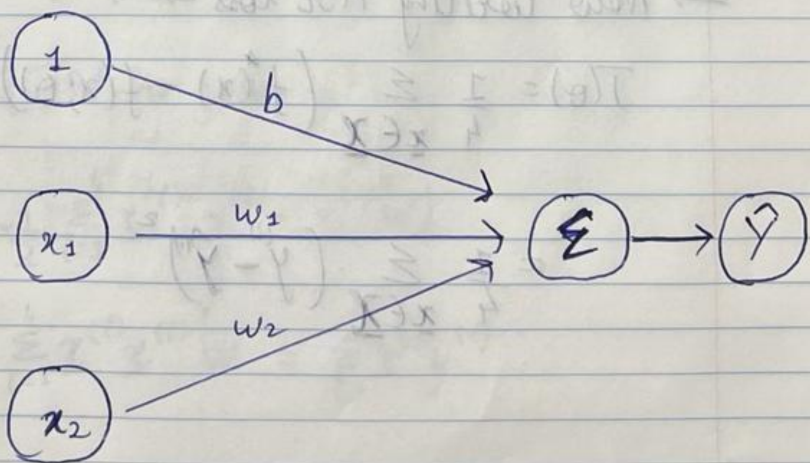
MSE ⇒ $J(\theta) = \frac{1}{4} \sum_{x \in X} \left(f^*(x) - f(x; \theta) \right)^2$

$f(x; w, b) = x^T \mathbf{w} + b$

Here, $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ (vector) & $x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$b = b$ (scalar)

→ 2 layer NN representation →



→ After understanding the NN model, modifying the given data to make the matrix equations more consistent.

$$W = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix} \quad \begin{pmatrix} \text{weight} \\ \text{matrix} \end{pmatrix}$$

$$\underline{X} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\underline{y} = [\underset{y^{(1)}}{0} \quad \underset{y^{(2)}}{1} \quad \underset{y^{(3)}}{1} \quad \underset{y^{(4)}}{0}]^T$$

$$\underline{x}^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 = 1 \end{bmatrix}$$

→ Now writing MSE loss →

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

$$\hat{J} = \frac{1}{4} \sum_{\underline{x} \in \underline{X}} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

$$\hat{y} = \underline{x}^{(i)^T} W$$ (Since this includes the bias no need for extra (+b) term)

$$\therefore \; J(\theta) = \frac{1}{4} \sum_{i=1}^{n=4} \left(y^{(i)} - \underline{x}^{(i)^T} W \right)^2$$

→ differentiating with respect to ∇w (i.e whole θ both $W(w_1 \,\&\, w_2) \,\&\, b$)

$$\nabla J(\theta) = \nabla w \frac{1}{4} \sum_{i=1}^{4} \left(y^{(i)} - \underline{x}^{(i)^T} W \right)^2$$

$$= \frac{1}{4} \sum_{i=1}^{4} \nabla w \left(y^{(i)} - \underline{x}^{(i)^T} W \right)^2$$

$$= \frac{1}{4} \sum_{i=1}^{4} \underline{x}^{(i)} \cdot 2 \left(y^{(i)} - \underline{x}^{(i)^T} W \right)$$

$$= -\frac{1}{2} \sum_{i=1}^{4} \underline{x}^{(i)} \left(y^{(i)} - \underline{x}^{(i)^T} W \right)$$

equating $\nabla J(\theta)$ to zero

$$0 = -\frac{1}{2} \sum_{i=1}^{4} \underline{x}^{(i)} \left(y^{(i)} - \underline{x}^{(i)^T} W \right)$$

$$\sum_{i=1}^{4} \underline{x}^{(i)} \underline{x}^{(i)^T} W = \sum_{i=1}^{4} \underline{x}^{(i)} y^{(i)}$$

$$\therefore \; w = \left(\sum_{i=1}^{4} \underline{x}^{(i)} \underline{x}^{(i)T} \right)^{-1} \sum_{i} \underline{x}^{(i)} y^{(i)}$$

→ This equation can also be represented in terms of the design matrix $\underline{X}$ & $\underline{Y}$

$$\therefore \; w = (\underline{X}\underline{X}^T)^{-1} \underline{X} \underline{Y}$$

$\therefore$ w =

$$(\underline{X}\underline{X}^T)^{-1} = \left(\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \right)^{-1}$$

$$= \left(\begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 2 \\ 2 & 2 & 4 \end{bmatrix} \right)^{-1}$$

$$= \begin{bmatrix} 1 & 0 & -0.5 \\ 0 & 1 & -0.5 \\ -0.5 & -0.5 & 0.75 \end{bmatrix}$$ (calculated using MATLAB)

$$\underline{X}\underline{Y} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$\therefore \; w = (\underline{X}\underline{X}^T)^{-1} \underline{X}\underline{Y}$$

$$= \begin{bmatrix} 1 & 0 & -0.5 \\ 0 & 1 & -0.5 \\ -0.5 & -0.5 & 0.75 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 + (-(0.5) \times 2) \\ 1 + (-(0.5) \times 2) \\ (-0.5) + (-0.5) + (2 \times 0.75) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.5 \end{bmatrix}$$

$$\therefore \; \begin{bmatrix} w1 \\ w2 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.5 \end{bmatrix}$$

$$\boxed{\therefore \; w_1 = 0 \qquad w_2 = 0 \qquad b = 0.5}$$

Problem 2: L_2 Regularized Linear Regression Via Stochastic Gradient Descent

- We developed a 2-layer neural network using the data given in homework 1. The file is saved as ***homework2_590146168_249812226.py.***
- The model is trained using Linear regression via stochastic gradient descent method.
- There are 4 Hyperparameters that can be altered:
    - Mini-Batch Size
    - Learning Rate
    - Number of Epochs
    - L_2 Regularization Strength
- The training data is split into Training (80%) and Validation (20%) dataset.
- We have selected a set of 10 values for each hyperparameter i.e., 10^4 combinations.
- The weights are initialized randomly and this same weight is used for training the model for each set of hyperparameters to keep the results fair.
- Once training is done, we calculate the cost on validation set. If the cost is minimum, the best hyperparameter is updated. This process is done till all hyperparameters are checked.
- It took approximately 22 hours to check all 10^4 combinations.
- At the end we obtain the best set of hyperparameters. Using these hyperparameters we train the model on validation + training dataset.
- Now this trained model is used to calculate the cost (unregularized MSE) for the test dataset and the result is reported below:

```
C:\Users\Fenil\anaconda3\python.exe "C:/THIS PC/WPI/SPRING 2022/CS 541 Deep Learning/H
----------------------------------------------------------------------
Performing Grid Search for 10e4 combinations of Hyperparameters:
----------------------------------------------------------------------
 Grid Search Completed
 Results after performing Grid Search:
 Best Hyperparameters:
   Epochs=  750
   Alpha=  1
   Learning Rate=  0.00075
   Mini Batch Size=  32
 Cost on Validation Set with Best Hyperparameters=  80.66432176996726

----------------------------------------------------------------------
Training on Training + Validation Dataset:
----------------------------------------------------------------------
 Training Completed

----------------------------------------------------------------------
Performance Evaluation
----------------------------------------------------------------------
 Cost on Test Dataset=  85.14249556730823
```

Problem 3: Logistic Sigmoid Identity

a. Prove $\boldsymbol{\sigma(-x) = 1 - \sigma(x)}\ \forall \boldsymbol{x}$

$$L.H.S = \frac{1}{1+e^{-(-x)}} = \frac{1}{1+e^{x}}$$

$$R.H.S = 1 - \frac{1}{1+e^{-x}} = \frac{1+e^{-x}-1}{1+e^{-x}}$$

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

$$= \boxed{\frac{1}{1+e^{x}}}$$ (dividing numerator & denominator by e^{-x})

$$\boxed{L.H.S = R.H.S}$$

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

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

→ $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\sigma'(x) = \frac{\partial}{\partial x}\left(\frac{1}{1+e^{-x}}\right) =$$

$$= \frac{-1}{(1+e^{-x})^2}\cdot\left(-e^{-x}\right) = \boxed{\frac{+e^{-x}}{(1+e^{-x})^2}}$$

$$= \left(\frac{1}{1+e^{-x}}\right)\cdot\left(\frac{e^{-x}}{1+e^{-x}}\right)$$

$$= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{+x}}\right) = \text{L.H.S}$$

$$= \sigma(x)\left(1-\sigma(x)\right)$$

$\left\{\left(\frac{1}{1+e^{+x}} = (1-\sigma(x))\right.\right.$ from last proof$\left.\right)\left.\right\}$

$\therefore$ L.H.S = R.H.S

Problem 4: Regularization to encourage symmetry

(4) L_2 regularization →

$$f_{MSE}(w) = \frac{1}{2n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 + \underbrace{\frac{\alpha}{2n} w^T w}_{\text{regularization term}}$$

→ our goal - discourage the weights from becoming too asymmetric.

let $W = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

→ according to our goal we should add $(w_2 - w_1)^2$ to the $f_{MSE}(w)$ for regularization. $(w_2 - w_1)$ ensures the weights are as similar to each-other as possible.
- The squared term ensures regularization is non-negative.

∴ Now, we are adviced to use $\frac{\alpha}{2n} w^T I w$ for regularization

∴ let us assume a square matrix $S_{2\times 2}$ such that

$$S = \begin{bmatrix} s_1 & s_2 \\ s_3 & s_4 \end{bmatrix}$$

S follows the following property,

$$(w_2 - w_1)^2 = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} s_1 & s_2 \\ s_3 & s_4 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$(w_2 - w_1)^2 = s_1 w_1^2 + w_1 w_2 (s_2 + s_3) + s_4 w_2^2$$

→ now expanding the L.H.S

$$w_2^2 - 2w_1 w_2 + w_1^2 = s_1 w_1^2 + w_1 w_2 (s_2 + s_3) + s_4 w_2^2$$

→ so from comparision we get to know that

$s_1 = 1$
$s_4 = 1$

$s_2 + s_3 = -2 \Rightarrow s_2 = -1$ & $s_3 = -1$

∴ $\boxed{S = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}}$

Problem 5: Linear-Gaussian Prediction Model

(5.) $P(y|x) = \mathcal{N}(\mu = x^T w, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - x^T w)^2}{2\sigma^2}\right)$

$\Rightarrow$ expected value of Y is $x^T w$

$\Rightarrow$ The variance of Y is constant (σ^2) for all possible x.

collection of dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n} \rightarrow$ MLE for w & σ

$\Rightarrow$ MLE - optimizing on w & σ^2

$P(D|w,\sigma^2) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}, w, \sigma^2)$ (considering the conditional independence)

$\rightarrow$ maximizing $\log P(D|w,\sigma^2)$

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

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

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y_{(i)} - x_{(i)}^T w)^2}{2\sigma^2}\right)\right)$$

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

$$= \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \left(\frac{(y^{(i)} - x^{(i)T} w)^2}{2\sigma^2}\right) \quad \text{— (1)}$$

→ Partially differentiating wrt to W.

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

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

— equating it to zero

$$\sum_{i=1}^{n} x^{(i)}y^{(i)} - \sum_{i=1}^{n} x^{(i)}x^{(i)T}W = 0$$

$$\sum_{i=1}^{n} x^{(i)}y^{(i)} = \sum_{i=1}^{n} x^{(i)}x^{(i)T}W$$

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

→ Partially differentiating eqn ① wrt to σ^2

rewriting eqn ① →

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

$$\sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2)\right) - \sum_{i=1}^{n} \left(\frac{(y^{(i)} - x^{(i)T}W)^2}{2\sigma^2}\right)$$

— now differentiating wrt σ^2

$$\sum_{i=1}^{n} \left(-\frac{1}{2} \times \frac{1}{2\pi\sigma^2} \times 2\pi\right) - \sum_{i=1}^{n} \frac{1}{2}(y^{(i)} - x^{(i)T}W)^2 \times \left(\frac{-1}{\sigma^4}\right) = 0$$

Now,

$$\sum_{i=1}^{n} \frac{1}{2}\frac{1}{\sigma^2} = \sum_{i=1}^{n} \frac{1}{2}\frac{(y^{(i)} - x^{(i)T}W)^2}{\sigma^4}$$

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

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

End Of Assignment