# Assignment# 6

# Group Members


| | | |
|---|---|---|
| **Muhammad Hamza jamil** | **2572890** | **s8mujami@stud.uni-saarland.de** |
| **Hacane Hechehouche** | **2571617** | **S8hahech@stud.uni-Saarland.de** |

Exersise 6.1

a) Maximum Likelihood estimator for $\theta$.

$$P_{model}(x;\theta) = \prod_{i=1}^{n} P_{model}(x^{(i)};\theta)$$

$$L(\theta) = \prod_{i=1}^{n} P_{model}(x^{(i)};\theta)$$

$$L(\theta) = P_m(x^1;\theta).P_m(x^2;\theta).P_m(x^3;\theta)\ldots P_m(x^N;\theta)$$

Taking Log on both sides

$$LL(\theta) = \sum_{i=1}^{N} \log P_{model}(x^i;\theta)$$

Taking derivate w.r.t $\theta$

$$\frac{LL(\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\left(\sum_{i=1}^{N} \log P_{model}(x^{(i)};\theta)\right)$$

(b) data-generated: it is the actual probability of what we should get as a result of estimation.

whereas the empirical distribution shows what we "did" estimate (what we got rather than What we should get).

we note that ~~also~~ the empirical is not smooth compared to the data generated one.

(c) $$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p_{model}(x^{(i)}; \theta)$$

because the argmax does not change when we rescale the cost function we can divide by m to obtain a version that is expressed as Expectation with respect to the empirical distribution $\hat{p}_{data}$.

to clarify this let $f(x, \theta) = \log p_{model}(x, \theta)$

then dividing by m the maximized expression gives:

$$\frac{1}{m} \cdot \sum_{i=1}^{m} f(x^{i}; \theta)$$

let's take a new random variable Y which follows the empirical distribution of the sample. that's a descrete random variable with $\frac{1}{m}$ probabi

then

$$\sum_{i=1}^{m} \frac{1}{m} \cdot f(x^{i}) = \sum_{i=1}^{m} P(Y = x^{i})\, f(x^{i}) = E_{Y \sim \hat{p}_{data}} f(Y)$$

So back to our maximization we get

$$\boxed{\theta_{ML} = \arg\max_{\theta} E_{x \sim \hat{p}_{data}} \log p_{model}(x; \theta)}$$

D here we need to minimize the dissimilarity between the actual data and the predicted ones. means to minimize the KL divergence. that is

$$D_{KL}(\hat{P}_{data} \| P_{model}) = E_{x \sim \hat{P}_{data}}\left[\log \hat{P}_{data}(x) - \log P_{model}(x)\right]$$

to minimize the KL divergence we only need to minimize

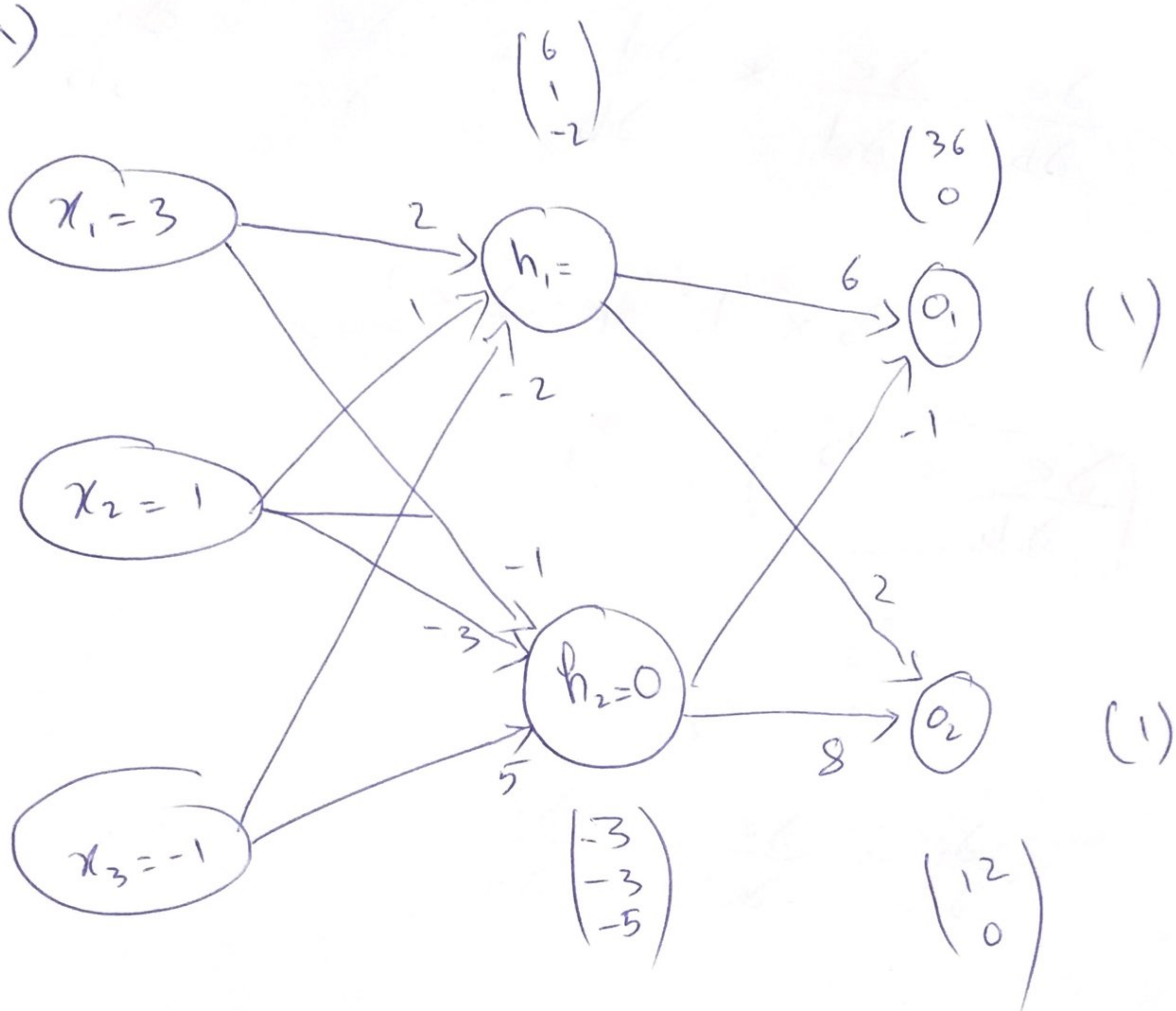
$$-E_{x \sim \hat{P}_{data}}\left[\log P_{model}(x)\right]$$

since the other term do not depend on the model.

- So minimizing this is the same maximization we obtained in C that is also the same maximization in a .

## Exercise 6.3

a)




because softmax function maps the values between $0 - 1$

$$\frac{e^{36}}{e^{36}+e^{0}} \approx 1$$

$$\frac{e^{12}}{e^{12}+e^{0}} \approx 1$$

b)

1) $$\frac{\partial e}{\partial b} = \frac{\partial e}{\partial d} \times \frac{\partial d}{\partial b} + \frac{\partial e}{\partial c} \times \frac{\partial c}{\partial b}$$

$$= 3 \times 1 + 2 \times 1$$

$$\boxed{\frac{\partial e}{\partial b} = 5}$$

2) $$\frac{\partial e}{\partial a} = \frac{\partial e}{\partial c} \times \frac{\partial c}{\partial a}$$

$$= 2 \times 1$$

$$\boxed{\frac{\partial e}{\partial a} = 2}$$

Exercise 6.2

a) Derivation of Sigmoid Function

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

$$\frac{\partial \sigma(x)}{\partial x} = -1\left(1+e^{-x}\right)^{-2} \times \left(-e^{-x}\right)$$

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

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

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

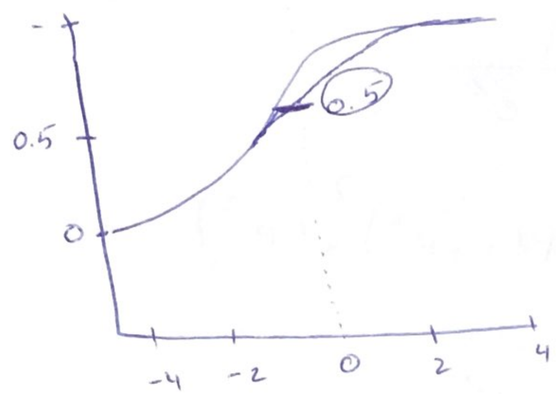
$$= \frac{1}{1+e^{-x}} \times \left(1 - \frac{1}{1+e^{-x}}\right)$$

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

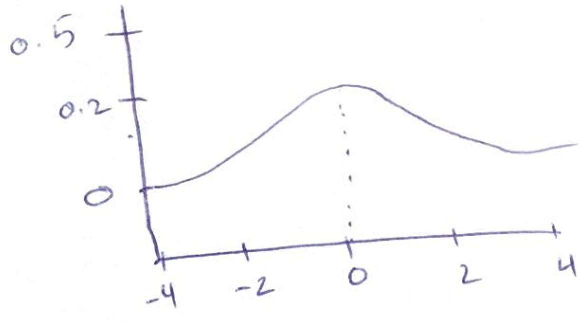
$$\boxed{= \sigma(x) - \sigma^2(x)}$$

b)

Sigmoid function is an activation function. it map's value between $0 \rightarrow 1$ here is the graph of simple sigmoid function




after taking derivative of sigmoid function it is normally distributed between 0 - 0.2




With this derivation logistic function for agiven layer can be evaluated using simple multiplication and subtraction rather than performing any re-evaluating the sigmoid function.

1) It transform linear input to nolear outputs

2) Sigmoid and the gradient of sigmoid function has Symmetric Propeties

c) Sigmoid function

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

A function is said to be symmetric
iff $g(-x) = g(x)$ or $g(-x) = -g(x)$

Sigmoid function

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

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

$$\cancel{\frac{\partial (x)}{\partial x} = \sigma(x) \times \sigma}$$

$$\frac{\partial(\sigma(x))}{\partial x} = \frac{e^{x}(1+e^{x}) - e^{x}.e^{x}}{(1+e^{x})^2}$$

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

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

The derivative of sigmoid function is even function

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

the sum of sigmoid function and its reflection about vertical axis $\sigma(-x)$ is

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

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

$$\boxed{\therefore \text{ symetri point } (0, 1/2)} \quad = 1$$

d)

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

From part a)

$$\sigma'(x) = \sigma(x) - \sigma^2(x)$$

$$\sigma'(0) = \frac{1}{2} - \frac{1}{4}$$

$$\sigma'(0) = \frac{1}{4} \Rightarrow 0.25$$

Second derivative

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

$$\sigma''(x) = \frac{-e^{-x}(1+e^{-x})^2 - e^{-x}\left(-2(1+e^{-x})e^{-x}\right)}{(1+e^{-x})^4}$$

$$= \frac{-e^{-x}(1+e^{-x})^2 + 2e^{-2x}(1+e^{-x})}{(1+e^{-x})^4}$$

$$\sigma''(x) = \frac{e^{-2x}(-e^{x}+1)}{(1+e^{-x})^3}$$

$$\sigma''(0) = \frac{1 \times (\cancel{-1} + \cancel{1})}{(1+\cancel{1})^3} = 0$$