CS 7650 : Problem Set 2

1).

$$L(\theta) = \prod_{j=1}^{m} p(y_j, x_j) = \prod_{j=1}^{m} p(y_j) \cdot \left[\prod_{i=1}^{n} p(x_{j,i} | y_j) \right]$$

naive assumption : conditional independence

$$= \theta_y^{c(y=1)} \cdot (1-\theta_y)^{c(y=0)} \cdot \prod_{i=1}^{n} \prod_{y \in \{0,1\}} \theta_{x_i=1|y}^{c(x_i=1, y)} \cdot (1-\theta_{x_i=1|y})^{c(x_i=0, y)}$$

models parameters

$\theta_y = p(y=1)$

$\theta_{x|y} = p(x|y)$

- Obtain the log-likelihood function.

$$l(\theta) = \log L(\theta) = c(y=1) \cdot \log \theta_y + c(y=0) \log 1-\theta_y + \sum_{i=1}^{n} \sum_{y \in \{0,1\}} c(x=1,y) \cdot \log \theta_{x_i=1|y} + c(x_i=0, y) \log 1-\theta_{x=1|y}$$

- Take the derivative w.r.t θ_y, set to 0.

$$\frac{\partial l(\theta)}{\partial \theta_y} = \frac{c(y=1)}{\theta_y} + \frac{c(y=0)}{1-\theta_y} (-1) = 0$$

$$\rightarrow \frac{c(y=1)}{\theta_y} = \frac{c(y=0)}{1-\theta_y} \qquad c(y=0) = m - c(y=1)$$

$$c(y=1) - \theta_y \cdot c(y=1) = m \cdot \theta_y - \theta_y \, c(y=1)$$

$$\boxed{\theta_y = \frac{c(y=1)}{m}}$$

2) $P_{logistic}(y=1|x) = \frac{e^{w^T x}}{1+e^{w^T x}}$

$P_{softmax}(y|x) = \frac{e^{w_y^T x}}{\sum_{y' \in y} e^{w_{y'}^T x}}$

Given: $y \in \{0,1\}$ — Binary classification

Find w for which:

$$P_{logistic}(y=1|x) = P_{softmax}(y=1|x)$$

$$\frac{e^{w^T x}}{1+e^{w^T x}} = \frac{e^{w_1^T x}}{e^{w_0^T x} + e^{w_1^T x}}$$

$$e^{(w^T + w_0^T)^T x} + e^{(w+w_1)^T x} = e^{w_1^T x} + e^{(w+w_1)^T x}$$

divide by $e^{w_1^T x}$

$$\therefore e^{(w+w_0^T)^T x - w_1^T x} = 1$$

$$e^{(w+w_0-w_1)^T x} = 1$$

Take ln on both sides

$$(w+w_0-w_1)^T x \cdot \ln e^1 = 0$$

$$\Rightarrow (w+w_0-w_1)^T x = 0$$

This can be true if $x=0$ or if $w+w_0-w_1=0$ provided that $x \neq 0$

or $(w+w_0-w_1)$ is orthogonal to x

$$\boxed{w = w_1 - w_0}$$

$\sum_j w^1_{ji} \cdot x_j$

3.) $z_i = \text{ReLU}(w_i^1 \cdot x + b_i)$
$y = w^2 \cdot z \longrightarrow \sum_i w^2_{i1} z_i$

3.1) Node z_i is "dead" when the linear combination of features ~~to~~ that is the input to the node is less than or equal to zero. For such inputs, the output of ReLU is 0 and the gradients will be 0 as well. This results in a "dead" node

> when $w_i^1 \cdot x + b_i \leq 0$, the neuron z_i is "dead"

3.2) $\frac{\partial L}{\partial y} = 1$.

Note: a_i denotes the input to neuron i

$- \frac{\partial L}{\partial b_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z_i} \cdot \frac{\partial z_i}{\partial b_i}$

$\frac{\partial y}{\partial z_i} \rightarrow w^2_{i1}$ (only 1 node in o/p layer)

$\frac{\partial z_i}{\partial b_i} \rightarrow \frac{\partial z_i}{\partial a_i} \cdot \frac{\partial a_i}{\partial b_i}$, $\frac{\partial a_i}{\partial b_i} \rightarrow 1$

$\frac{\partial L}{\partial y} \rightarrow 1$

$\frac{\partial L}{\partial b_i} = w^2_{i1} \cdot \text{drelu}(a_i)$ where $\text{drelu}(a_i) = \begin{cases} 1 \text{ if } a_i > 0 \\ 0 \text{ if } a_i \leq 0 \end{cases}$

$- \frac{\partial L}{\partial w_{ji}^1} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z_i} \cdot \frac{\partial z_i}{\partial a_i} \cdot \frac{\partial a_i}{\partial w_{ji}^1} \times x_j$

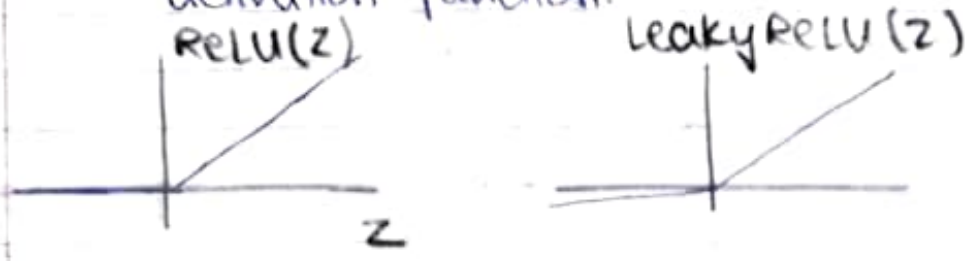
$\frac{\partial y}{\partial z_i} \rightarrow w^2_{i1}$, $\frac{\partial z_i}{\partial a_i} \rightarrow \text{drelu}(a_i)$

$\frac{\partial L}{\partial w_{ji}^1} = w^2_{i1} \cdot \text{drelu}(a_i) \cdot x_j$ where $\text{drelu}(a_i) = \begin{cases} 1 \text{ if } a_i > 0 \\ 0 \text{ if } a_i \leq 0 \end{cases}$

3.3) During gradient based learning, a dead neuron can never be brought back to life. This is because when the input to a node (a_i) is less than or equal to zero, the output (z_i) will be zero (Forward propagation). During backward propagation, the gradient term $\frac{\partial z_i}{\partial a_i} = d\text{relu}(a_i) = 0$.

As a result, the weight that are connected to this term can never be updated since $w^{t+1} = w^t - \alpha \nabla J(\theta)$ will not change the weights. As a result, the dead neuron can't ever be brought to life.

3.4) To overcome this problem, I would suggest using the leaky RELU activation function.



The gradient of leakyRELU takes the form

$$d\text{leakyRELU}(a_i) \begin{cases} 1 & \text{if } a_i > 0 \\ 0.3 & \text{if } a_i \leq 0 \end{cases}$$

- small value

Now, the gradient wont be zero when $a_i \leq 0$, and therefore the neurons won't "die".

4.) $P(y|x) = \text{softmax}(W\, q(\underbrace{V f(x)}_{z}))$

$L(x, i^*) = wz \cdot e_{i^*} - \log \sum_j \exp(wz) \cdot e_j$

$\text{error}(\text{root}) \overset{\text{def}}{=} e_{i^*} - p(y|x)$.

Show that $\text{error}(z) \overset{\text{def}}{=} \frac{\partial L(x, i^*)}{\partial z} = w^T [e_{i^*} - p(y|x)]$

$\overset{\text{def}}{=} w^T \text{error}(\text{root})$

$\frac{\partial L(x, i^*)}{\partial z} = \frac{\partial}{\partial z} \left(wz \cdot e_{i^*} - \log \sum_j \exp(wz) \cdot e_j \right)$

$= w^T e_{i^*} - \frac{1 \cdot \exp(wz) \cdot w^T}{\sum_j \exp(wz) \cdot e_j}$

$\Rightarrow$ conditional probability $p(y|x)$

$\text{softmax}(wz)$ $= w^T e_{i^*} - w^T p(y|x)$

$= w^T [e_{i^*} - p(y|x)]$

$\boxed{\text{error}(z) = w^T \cdot \text{error}(\text{root})}$