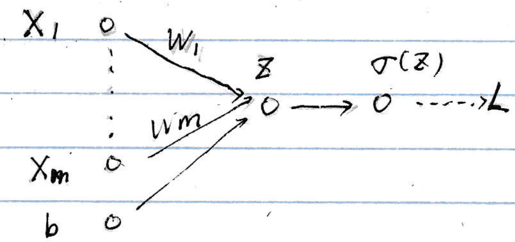
Intro to Deep Learning
HW 1

Yifan Li
yl506

Problem 1:

logistic regression (binary outcome)

a)



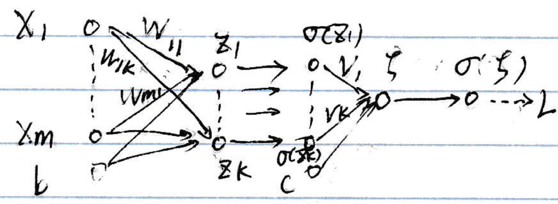
$z = w_1x_1 + w_2x_2 + \dots + w_m x_m + b$

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

$L(y, \sigma(z)) = -y \log\sigma(z) - (1-y)\log(1-\sigma(z))$

- $\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \sigma(z)} \frac{\partial \sigma(z)}{\partial z} \frac{\partial z}{\partial w_i} = \frac{\sigma(z)-y}{(1-\sigma(z))\sigma(z)} \cdot \sigma(z)(1-\sigma(z)) \cdot x_i = x_i(\sigma(z)-y)$
- $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \sigma(z)} \frac{\partial \sigma(z)}{\partial z} \frac{\partial z}{\partial b} = \frac{\sigma(z)-y}{\sigma(z)(1-\sigma(z))} \cdot \sigma(z)(1-\sigma(z)) \cdot 1 = \sigma(z) - y$

b) MLP w/ a single hidden layer (binary outcome)



$z_i = w_{1i}x_1 + w_{2i}x_2 + \dots + w_{mi}x_m + b$

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

$\zeta = v_1\sigma(z_1) + v_2\sigma(z_2) + \dots + v_k\sigma(z_k) + c$

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

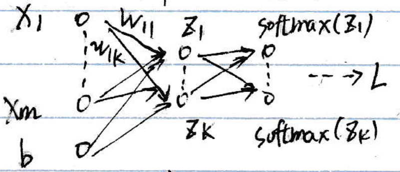
$L(y, \sigma(\zeta)) = -y\log\sigma(\zeta) - (1-y)\log(1-\sigma(\zeta))$

- $\frac{\partial L}{\partial v_i} = \frac{\partial L}{\partial \sigma(\zeta)} \frac{\partial \sigma(\zeta)}{\partial \zeta} \frac{\partial \zeta}{\partial v_i} = \frac{\sigma(\zeta)-y}{\sigma(\zeta)(1-\sigma(\zeta))} \cdot \sigma(\zeta)(1-\sigma(\zeta)) \cdot \sigma(z_i) = \sigma(z_i)(\sigma(\zeta)-y)$
- $\frac{\partial L}{\partial c} = \frac{\partial L}{\partial \sigma(\zeta)} \frac{\partial \sigma(\zeta)}{\partial \zeta} \frac{\partial \zeta}{\partial c} = \sigma(\zeta) - y$
- $\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial \sigma(\zeta)} \frac{\partial \sigma(\zeta)}{\partial \zeta} \frac{\partial \zeta}{\partial \sigma(z_j)} \frac{\partial \sigma(z_j)}{\partial z_j} \frac{\partial z_j}{\partial w_{ij}} = \frac{\sigma(\zeta)-y}{\sigma(\zeta)(1-\sigma(\zeta))} \cdot \sigma(\zeta)(1-\sigma(\zeta)) \cdot v_j$

  $\cdot\, \sigma(z_j)(1-\sigma(z_j)) \cdot x_i = (\sigma(\zeta)-y)\, v_j\, \sigma(z_j)(1-\sigma(z_j))\, x_i$
- $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \sigma(\zeta)} \frac{\partial \sigma(\zeta)}{\partial \zeta} \frac{\partial \zeta}{\partial \sigma(z_j)} \frac{\partial \sigma(z_j)}{\partial z_j} \frac{\partial z_j}{\partial b} = (\sigma(\zeta)-y)\, v_j\, \sigma(z_j)(1-\sigma(z_j))$

HW 1 Yifan Li yl506

c) When using a softmax (multi-class) setup in part (a), the network now becomes



$z_i = W_{1i} x_1 + W_{2i} x_2 + \dots + W_{mi} x_m + b$

$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_K e^{z_i}}$

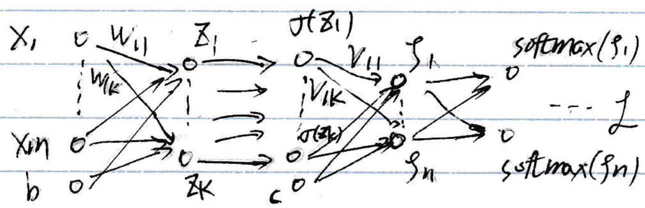
$\mathcal{L}(y, \text{softmax}(z_i)) = -\sum_{k=1}^{K} 1(y_i=1) \log \text{softmax}(z)_k = -\sum_K y_i \log P_j,\ P_j = \frac{e^{z_j}}{\sum_K e^{z_K}}$

let $O_j = \text{softmax}(z_j)$

- $\frac{\partial L}{\partial W_{ij}} = \frac{\partial L}{\partial O_j} \frac{\partial O_j}{\partial z_j} \frac{\partial z_j}{\partial W_{ij}} = (P_j - y_j) X_i$, where $P_j = \frac{e^{z_j}}{\sum_K e^{z_K}}$

- $\frac{\partial L}{\partial b} = P_j - y_j$. The loss function changes, so does $\frac{\partial L}{\partial W_{ij}}$ and $\frac{\partial L}{\partial b}$

in part (b), the network now becomes



$g_i = V_{1i}\sigma(z_1) + V_{2i}\sigma(z_2) + \dots + V_{Ki}\sigma(z_K) + c$

$\mathcal{L}(y, \text{softmax}(g_i)) = -\sum_n y \log P_j$, where $P_j = \frac{e^{g_j}}{\sum_n e^{g_n}}$

let $O_i = \text{softmax}(g_i)$

- $\frac{\partial L}{\partial V_{ij}} = \frac{\partial L}{\partial O_j} \frac{\partial O_j}{\partial g_j} \frac{\partial g_j}{\partial V_{ij}} = (P_j - y_j)\,\sigma(z_i)$. $\frac{\partial L}{\partial C} = \frac{\partial L}{\partial O_j} \frac{\partial O_j}{\partial g_j} \frac{\partial g_j}{\partial C} = P_j - y_j$

- $\frac{\partial L}{\partial W_{ij}} = \frac{\partial L}{\partial O_j} \frac{\partial O_j}{\partial g_j} \frac{\partial g_j}{\partial \sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial z_i} \frac{\partial z_i}{\partial W_{ij}} = (P_j - y_j) V_{ij}\, \sigma(z_i)(1-\sigma(z_i)) X_i$

- $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial O_j} \frac{\partial O_j}{\partial g_j} \frac{\partial g_j}{\partial \sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial z_i} \frac{\partial z_i}{\partial b} = (P_j - y_j) V_{ij}\, \sigma(z_i)(1-\sigma(z_i))$

The difference is that since the softmax is used for multi-class outcome, a few terms of the original gradient change as well.

Problem 5.

a) Yes.

b) About 12 hours.

c) I adhered to the Duke Community Standard in the completion of this assignment.