Assignment2 Solution

Artificial neurons

1. y = w^Tx = 0.6

2. on one side the value is larger than 0, on the other side the value is smaller than 0, and on the decision boundary the value is equal to 0.

3. θ_0 is the negative distance from decision boundary to origin; θ_1 and θ_2 is the normal vector of the decision boundary.

4. Normal is (2, 3), distance is -1/$\sqrt{13}$.

5. Bias coefficient is the first θ marked as θ_0.
   Feature vector $(x_1, x_2 \ldots x_n)$ is rewritten to (1, $x_1, x_2 \ldots x_n$).

6. Step function: h(z) = $\begin{cases} 1 \; if \; z > 0 \\ 0 \; if \; z \leq 0 \end{cases}$ Sigmoid function: h(z) = $\frac{1}{1+\mathrm{e}^{-z}}$
   We can compute derivative for sigmoid activation, but derivative is always 0 for step activation. When θ is small, h(a) is closest to 0.5 which means it's hard to decide class.

7. P(y=1|x) = exp($\theta^T x$) P(y=0|x) = exp($\theta^T x$) (1 - P(y=1|x)) =$\frac{\exp(\theta^T x)}{1+\exp(\theta^T x)} = \frac{1}{1+\exp(-\theta^T x)}$

8. $\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))$ $\frac{\partial \log(\sigma(z))}{\partial z} = 1 - \sigma(z)$, where σ(z) is the sigmoid.

9. Compute the direction by computing the derivative of the loss function w.r.t the parameters. Control the size of the update by learning rate.

10. Stop when the loss change is smaller than a threshold. The condition use loss change because small change in parameters would make big change in loss.

11. If learning rate is too small, it will take too long to converge or get stuck in an undesirable local minimum. If learning rate is too high, it will make the learning jump over minima.

12. The empirical error loss is the number of misclassified examples.
    The problem is that it does not have the derivative for the gradient descent.

13. $$\begin{aligned} l(\theta) &= \log\left(\prod_{x^{(i)} \in c1} P(y = 1|x^{(i)}) \prod_{x^{(i)} \in c0} P(y = 0|x^{(i)})\right) \\ &= \log \prod_{i=1}^{m} P(y = 1|x^{(i)})^{y(i)} P(y = 0|x^{(i)})^{1-y(i)} \\ &= \sum_{i=1}^{m} \log(\mathrm{P}(\mathrm{y} = 1|x^{(i)}) + (1 - y^{(i)}) \log(1 - \mathrm{P}(\mathrm{y} = 1|x^{(i)}) \\ &= \sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)})) \end{aligned}$$

14. $\frac{dl(\theta)}{d\theta} = \frac{d}{d\theta} \sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$ = $\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))\, x^{(i)}$
    $\frac{d(-l(\theta))}{d\theta} = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x^{(i)}$
    $\theta \leftarrow \theta - \eta \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x^{(i)}$

15. One against all others: k discriminant functions, each class has one discriminant function.
    One against each other: k(k-1)/2 discriminant functions, a pair of classes have one discriminant functions.
    One against each other is easier to discriminate the data.

16. The rows of θ^T are templates and k rows of θ^T means k templates (one template per class). $\theta^T x$ measures how well x matches each of templates; high similarities to a template of a particular class indicates high membership in this class.
    Using multiple linear discriminant functions means multiple templates per class.

17. Each component will be in the interval (0, 1), and the components will add up to 1, so that they can be interpreted as probabilities. Furthermore, the larger components will correspond to larger probabilities.

18. $\frac{\partial \sigma(z_j)}{\partial z_i} = \sigma(z_j)\left(\delta_{i,j} - \sigma(z_i)\right)$, where σ(z) is the softmax.
    $\frac{\partial \log(\sigma(z_j))}{\partial z_i} = \frac{1}{\sigma(z_j)}\sigma(z_j)\left(\delta_{i,j} - \sigma(z_i)\right) = \delta_{i,j} - \sigma(z_i)$, where σ(z) is the softmax.

19. $$l(\theta) = \log \prod_{i=1}^{m} \prod_{j=1}^{k} P\left(y^{(i)} = j \middle| x^{(i)}\right)^{1(y^{(i)}=j)}$$
    $$= \sum_{i=1}^{m} \sum_{j=1}^{k} 1(y^{(i)} = j) log P\left(y^{(i)} = j \middle| x^{(i)}\right)$$
    $$= \sum_{i=1}^{m} \sum_{j=1}^{k} 1(y^{(i)} = j) \log (h_{\theta_j}(x^{(i)}))$$

20. $\frac{d(-l(\theta))}{d\theta} = \sum_{i=1}^{m}(h_{\theta_j}\left(x^{(i)}\right) - 1(y^{(i)} = j))\, x^{(i)}$
    $\theta_j \leftarrow \theta_j - \eta \sum_{i=1}^{m}(h_{\theta_j}\left(x^{(i)}\right) - 1(y^{(i)} = j))\, x^{(i)}$

Neural networks

1. Mapping the input to higher dimensional spaces.
   Benefit: it can find the relations in high dimensional spaces.

2. Condense the input to lower dimensional spaces.
   Benefit: extract the features with useful information.

3. Because they have different gradient values:
   for hidden layer, the gradient is $\frac{\partial E}{\partial w_j} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_j} \frac{\partial z_j}{\partial w_j}$;
   for output layer, the gradient is $\frac{\partial E}{\partial v} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v}$

4. Hidden layer: $\frac{\partial E}{\partial w_j} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_j} \frac{\partial z_j}{\partial w_j} = \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) v_j {z_j}^{(i)} (1 - {z_j}^{(i)}) x^{(i)}$
   Output layer: $\frac{\partial E}{\partial v} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v} = \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) z^{(i)}$

5. Hidden layer: $\frac{\partial E}{\partial w_j} = \sum_{l=1}^{k} \frac{\partial E}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial z_j} \frac{\partial z_j}{\partial w_j} = \sum_{i=1}^{m} \sum_{l=1}^{k} ({\hat{y}_l}^{(i)} - {y_l}^{(i)}) v_{lj} {z_j}^{(i)} (1 - {z_j}^{(i)}) x^{(i)}$
   Output layer: $\frac{\partial E}{\partial v_j} = \frac{\partial E}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial v_j} = \sum_{i=1}^{m} ({\hat{y}_j}^{(i)} - {y_j}^{(i)}) z^{(i)}$

6. Hidden layer: $\frac{\partial E}{\partial w_j} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_j} \frac{\partial z_j}{\partial w_j} = \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) v_j {z_j}^{(i)} (1 - {z_j}^{(i)}) x^{(i)}$
   Output layer: $\frac{\partial l}{\partial v} = \frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v} = \sum_{i=1}^{m} (y^{(i)} \frac{1}{\hat{y}^{(i)}} - (1 - y^{(i)}) \frac{1}{1 - \hat{y}^{(i)}}) \hat{y}^{(i)} 1 - \hat{y}^{(i)}) z^{(i)}$
   $= \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) z^{(i)}$

7. Hidden layer: $\frac{\partial E}{\partial w_j} = \sum_{l=1}^{k} \frac{\partial E}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial z_j} \frac{\partial z_j}{\partial w_j} = \sum_{i=1}^{m} \sum_{l=1}^{k} ({\hat{y}_l}^{(i)} - {y_l}^{(i)}) v_{lj} {z_j}^{(i)} (1 - {z_j}^{(i)}) x^{(i)}$
   Output layer: $\frac{\partial E}{\partial v_j} = \frac{\partial E}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial v_j} = \sum_{i=1}^{m} ({\hat{y}_j}^{(i)} - {y_j}^{(i)}) z^{(i)}$

8. The weights should be initialized randomly and close to 0.
   The reason is that breaking symmetry between different units.

Computation graphs:

Note: The L2 loss we define as: $\frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$, so for some of you who define the L2 loss as: $\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$, all the solution results using L2 should be multiplied by 2.

1. It is a method of expressing and evaluating the mathematical expression. It helps better and easier understanding the way it is organized to perform forward pass followed by a backward pass or backward propagation step.

   In the forward pass, it computes the outputs for each node. In the backward pass, it computes gradients/derivatives with respect to each input.

   Each node is able to compute and store the output based on its corresponding mathematical operation during forward pass, and also the gradients with respect to its input variables during backward pass.

   The reason is that it can be used for realizing chain rule, which is used to effectively train a neural network, when calculating the derivatives for a nested function.

2. $\hat{y} = f_2(W_2, f_1(W_1, x))$

   L2:

   $L = \frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - f_2(W_2, f_1(W_1, x^{(i)})))^2$

   $\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1} = -\sum_{i=1}^{m}(y^{(i)} - f_2(W_2, f_1(W_1, x^{(i)})))\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1}$

   $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial W_2} = -\sum_{i=1}^{m}(y^{(i)} - f_2\left(W_2, f_1\left(W_1, x^{(i)}\right)\right))\frac{\partial f_2}{\partial W_2}$

   Cross entropy:

   $L = \sum_{i=1}^{m} y^{(i)} \log f_2\left(W_2, f_1\left(W_1, x^{(i)}\right)\right) + (1 - y^{(i)})\log\,(1 - f_2\left(W_2, f_1\left(W_1, x^{(i)}\right)\right))$

   $\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1} = \sum_{i=1}^{m}(\frac{y^{(i)}}{f_2\left(W_2, f_1\left(W_1, x^{(i)}\right)\right)} - \frac{1 - y^{(i)}}{1 - f_2\left(W_2, f_1\left(W_1, x^{(i)}\right)\right)})\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1}$

   $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial W_2} = \sum_{i=1}^{m}\frac{y^{(i)}}{f_2(W_2, f_1(W_1, x^{(i)}))}\frac{\partial f_2}{\partial W_2}$

3. $\hat{y} = f_2\left(W_2, f_1(W_1, x)\right)$

   L2:

   $L = \frac{1}{2}\sum_{i=1}^{m}\sum_{c=1}^{C}(y_c^{(i)} - f_2(W_2, f_1(W_1, x^{(i)})))^2$

   $\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1} = -\sum_{i=1}^{m}\sum_{c=1}^{C}(y_c^{(i)} - f_2(W_2, f_1(W_1, x^{(i)})))\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1}$

   $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial W_2} = -\sum_{i=1}^{m}\sum_{c=1}^{C}(y_c^{(i)} - f_2(W_2, f_1(W_1, x^{(i)})))\frac{\partial f_2}{\partial W_1}$

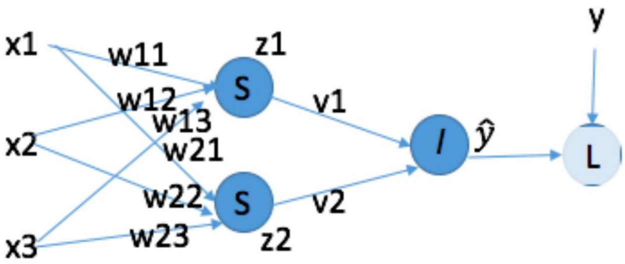
Cross entropy:

$$L = \sum_{i=1}^{m} \sum_{c=1}^{C} y_c^{(i)} \log f_2(W_2, f_1(W_1, x^{(i)}))$$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial f_1}\frac{\partial f_1}{\partial W_1} = \sum_{i=1}^{m} \sum_{c=1}^{C} \frac{y^{(i)}}{f_2(W_2,f_1(W_1,x^{(i)}))}\frac{\partial f_2}{\partial f_1}\ \frac{\partial f_1}{\partial W_1}$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial f_2}\frac{\partial f_2}{\partial W_2}=\sum_{i=1}^{m} \sum_{c=1}^{C} \frac{y^{(i)}}{f_2(W_2,f_1(W_1,x^{(i)}))}\frac{\partial f_2}{\partial W_2}$$

4.



$$\frac{\partial L}{\partial v0} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v0} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} z_0^{(i)}$$

$$\frac{\partial L}{\partial v1} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v1} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} z_1^{(i)}$$

$$\frac{\partial L}{\partial v2} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v2} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} z_2^{(i)}$$

$$\frac{\partial L}{\partial w10} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w10} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v1 \cdot z_1^{(i)}(1- z_1^{(i)}) \cdot 1$$

$$\frac{\partial L}{\partial w11} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w11} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v1 \cdot z_1^{(i)}\left(1- z_1^{(i)}\right) \cdot x_1^{(i)}$$

$$\frac{\partial L}{\partial w12} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w12} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v1 \cdot z_1^{(i)}(1- z_1^{(i)}) \cdot x_2^{(i)}$$

$$\frac{\partial L}{\partial w13} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w13} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v1 \cdot z_1^{(i)}(1- z_1^{(i)}) \cdot x_3^{(i)}$$

$$\frac{\partial L}{\partial w20} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w20} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v2 \cdot z_2^{(i)}\left(1- z_2^{(i)}\right) \cdot 1$$

$$\frac{\partial L}{\partial w21} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w21} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v2 \cdot z_2^{(i)}(1- z_2^{(i)}) \cdot x_1^{(i)}$$

$$\frac{\partial L}{\partial w22} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w22} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v2 \cdot z_2^{(i)}(1- z_2^{(i)}) \cdot x_2^{(i)}$$

$$\frac{\partial L}{\partial w23} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w23} = \sum_{i=1}^{m} \frac{\partial L}{\partial \hat{y}} \cdot v2 \cdot z_2^{(i)}(1- z_2^{(i)}) \cdot x_3^{(i)}$$

5. Addition of a constant: incoming gradient is multiplied by 1.

$$\frac{\partial}{\partial x}(x + c) = 1$$

Addition of two inputs: incoming gradient is multiplied by 1 with respect to each input.

$$\frac{\partial}{\partial x}(x + y) = 1 \qquad \frac{\partial}{\partial y}(x + y) = 1$$

Multiplication of a constant: incoming gradient is multiplied by the constant.

$$\frac{\partial}{\partial x}cx = c$$

Multiplication of two inputs: incoming gradient is multiplied by the other input.

$$\frac{\partial}{\partial x}(xy) = y \qquad \frac{\partial}{\partial y}(xy) = x$$

Max of two inputs: incoming gradient is multiplied by 1 only for max input.

$$\frac{\partial}{\partial x}(\max(x, y)) = 1 \;\; if\ x \geq y \qquad \frac{\partial}{\partial y}(\max(x, y)) = 1 \;\; if\ y \geq x$$

$$\frac{\partial}{\partial x}(\max(x, y)) = 0 \;\; otherwise \qquad \frac{\partial}{\partial y}(\max(x, y)) = 0 \;\; otherwise$$

Min of two inputs: incoming gradient is multiplied by 1 only for min input.

$$\frac{\partial}{\partial x}(\min(x, y)) = 1 \;\; if\ x \leq y \qquad \frac{\partial}{\partial y}(\min(x, y)) = 1 \;\; if\ y \leq x$$

$$\frac{\partial}{\partial x}(\min(x, y)) = 0 \;\; otherwise \qquad \frac{\partial}{\partial y}(\min(x, y)) = 0 \;\; otherwise$$

6. $\frac{\partial f}{\partial x} = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}]$

7. $\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$

8. $\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_{11}} & \cdots & \frac{\partial f_1}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_{m1}} & \cdots & \frac{\partial f_1}{\partial x_{mn}} \end{bmatrix}$

9. $D(\mathrm{F} \circ \mathrm{G}) = D(F(G))D(G)$ or $D(\mathrm{G} \circ \mathrm{F}) = D(G(F))D(F)$

10. $(\mathrm{F} \circ \mathrm{G})(\mathrm{x}, \mathrm{y}) = \ [3x^2y - 15xy^2 \quad xy - x + y]^T$

directly:

$$D(\mathrm{F} \circ \mathrm{G})(\mathrm{x}, \mathrm{y}) = \begin{bmatrix} 6xy - 15y^2 & 3x^2 - 30xy \\ y - 1 & x + 1 \end{bmatrix}$$

chain rule: $D\big(F(x, y, z)\big) = \begin{bmatrix} 3y & 3x & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad D\big(\mathrm{G}(\mathrm{x}, \mathrm{y})\big) = \begin{bmatrix} 1 & -5 \\ y & x \\ 1 & -1 \end{bmatrix}$

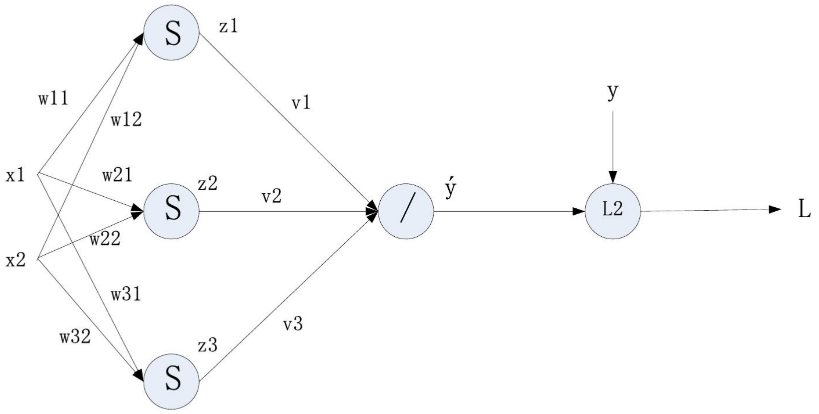
$$D(\mathrm{F} \circ \mathrm{G})(\mathrm{x}, \mathrm{y}) = D\left(F\big(G(x, y)\big)\right) D\big(G(x, y)\big)$$

$$= \begin{bmatrix} 3xy & 3x - 15y & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -5 \\ y & x \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 6xy - 15y^2 & 3x^2 - 30xy \\ y - 1 & x + 1 \end{bmatrix}$$

11. It is necessary for realizing chain rule when performing back propagation, meaning that the outer function node should be computed after inner function node for forward pass, while in the opposite direction for backward pass.

12. We define w10, w20, w30, v0 as bias, and w11, w12, w21, w22, w31, w32, v1, v2, v3 as weight.

(a):



(b): (1) x1, y1 = [(1,2), 8]:

$z_1^{(1)}$ = 1 / (1 + $e^{-(w_{10}+w_{11}x_{11}+w_{12}x_{12})}$) = 0.52248

$z_2^{(1)}$ = 1 / (1 + $e^{-(w_{10}+w_{11}x_{21}+w_{12}x_{22})}$) = 0.51999

$z_3^{(1)}$ = 1 / (1 + $e^{-(w_{10}+w_{11}x_{31}+w_{12}x_{32})}$) = 0.51749

$\hat{y}$ = 0.01+0.02*$z_1^{(1)}$+0.03*$z_2^{(1)}$+0.04*$z_3^{(1)}$ = 0.05672

(2) x2, y2 = [(1,3), 11]:

$\hat{y}$ = 0.01+0.02*$z_1^{(2)}$+0.03*$z_2^{(2)}$+0.04*$z_3^{(2)}$ = 0.05715

(3) x1, y1 = [(2,2), 10]:

$\hat{y}$ = 0.01+0.02*$z_1^{(3)}$+0.03*$z_2^{(3)}$+0.04*$z_3^{(3)}$ = 0.0572 (c): $\frac{\partial L}{\partial v0} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v0} = \sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})z_0^{(i)}$ = -28.829

$\frac{\partial L}{\partial v1} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v1} = \sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})z_1^{(i)}$ = -15.186

$\frac{\partial L}{\partial v2} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v2} = \sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})z_2^{(i)}$ = -15.066

$\frac{\partial L}{\partial v3} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v3} = \sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})z_3^{(i)}$ = -15.017

$\frac{\partial L}{\partial w10} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w10} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v1 \cdot z_1^{(i)}(1 - z_1^{(i)}) \cdot 1$ = -0.1437

$\frac{\partial L}{\partial w11} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w11} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v1 \cdot z_1^{(i)}(1 - z_1^{(i)}) \cdot x_1^{(i)}$ = -0.1933

$\frac{\partial L}{\partial w12} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w12} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v1 \cdot z_1^{(i)}(1 - z_1^{(i)}) \cdot x_2^{(i)}$ = -0.34219

$\frac{\partial L}{\partial w20} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w20} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v2 \cdot z_2^{(i)}\left(1 - z_2^{(i)}\right) \cdot 1$ = -0.2157

$\frac{\partial L}{\partial w21} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w21} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v2 \cdot z_2^{(i)}(1 - z_2^{(i)}) \cdot x_1^{(i)}$ = -0.2897

$\frac{\partial L}{\partial w22} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w22} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v2 \cdot z_2^{(i)}(1 - z_2^{(i)}) \cdot x_2^{(i)}$ = -0.5133

$\frac{\partial L}{\partial w30} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z3}\frac{\partial z3}{\partial w30} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v3 \cdot z_3^{(i)}\left(1 - z_3^{(i)}\right) \cdot 1$ = -0.2877

$\frac{\partial L}{\partial w31} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z3}\frac{\partial z3}{\partial w31} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v3 \cdot z_3^{(i)}\left(1 - z_3^{(i)}\right) \cdot x_1^{(i)}$ = -0.3866

$\frac{\partial L}{\partial w32} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z3}\frac{\partial z3}{\partial w32} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v3 \cdot z_3^{(i)}(1 - z_3^{(i)}) \cdot x_2^{(i)}$ = -0.6847

(d): $\frac{\partial L}{\partial v} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial v} = \sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})z^{(i)}$ = (-28.83, -15.186, -15.066, -15.017)

$\frac{\partial L}{\partial w1} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z1}\frac{\partial z1}{\partial w1} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v1 \cdot z_1^{(i)}(1 - z_1^{(i)})x^{(i)}$ = (-0.1437, -0.1933, -0.3419)

$\frac{\partial L}{\partial w2} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z2}\frac{\partial z2}{\partial w2} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v2 \cdot z_2^{(i)}(1 - z_2^{(i)})x^{(i)}$ = (-0.2157, -0.2897, -0.5133)

$\frac{\partial L}{\partial w3} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z3}\frac{\partial z3}{\partial w3} = \sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v3 \cdot z_3^{(i)}(1 - z_3^{(i)})x^{(i)}$ = (-0.2877, -0.3866, -0.6847)

13. (a): $\begin{bmatrix} 8x + 12y \\ 12x + 18y \end{bmatrix}$

(b): $\begin{bmatrix} 2x & 2 \\ 3 & 8y \end{bmatrix}$ = $\begin{bmatrix} 2 & 2 \\ 3 & 16 \end{bmatrix}$

(c): $\begin{bmatrix} 6x \\ 16x^3 + 3 \end{bmatrix}$ = $\begin{bmatrix} 12 \\ 131 \end{bmatrix}$

(d): For each data,

$$\frac{\partial L}{\partial v^{(i)}} = D(L \circ Y)(v) = \mathrm{DL}\big(\mathrm{Y(v)}\big) \cdot DY(v) = [\hat{y}^{(i)} - y^{(i)}] \cdot \begin{bmatrix} z_0^{(i)} & z_1^{(i)} & z_2^{(i)} & z_3^{(i)} \end{bmatrix}$$

$$\frac{\partial L}{\partial w^{(i)}} = D(L \circ Y)(w) = \mathrm{DL}\big(\mathrm{Y(w)}\big) \cdot DY(w)$$

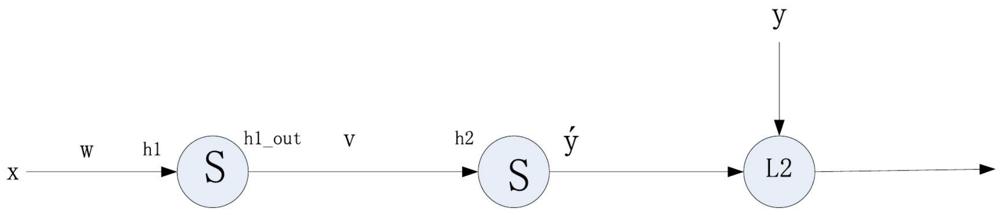
$$= [\hat{y}^{(i)} - y^{(i)}] \cdot \begin{bmatrix} v1 \cdot z_1^{(i)}(1 - z_1^{(i)}) \cdot x_0^{(i)} & v1 \cdot z_1^{(i)}(1 - z_1^{(i)}) \cdot x_1^{(i)} & v1 \cdot z_1^{(i)}(1 - z_1^{(i)}) \cdot x_2^{(i)} \\ v2 \cdot z_2^{(i)}(1 - z_2^{(i)}) \cdot x_0^{(i)} & v2 \cdot z_2^{(i)}(1 - z_2^{(i)}) \cdot x_1^{(i)} & v2 \cdot z_2^{(i)}(1 - z_2^{(i)}) \cdot x_2^{(i)} \\ v3 \cdot z_3^{(i)}(1 - z_3^{(i)}) \cdot x_0^{(i)} & v3 \cdot z_3^{(i)}(1 - z_3^{(i)}) \cdot x_1^{(i)} & v3 \cdot z_3^{(i)}(1 - z_3^{(i)}) \cdot x_2^{(i)} \end{bmatrix}$$

We sum the three data's $\frac{\partial L}{\partial v^{(i)}}$ and $\frac{\partial L}{\partial w^{(i)}}$, and then we get:

$$\frac{\partial L}{\partial v} = [-28.829, -15.186, -15.066, -15.017]$$

$$\frac{\partial L}{\partial w} = \begin{bmatrix} -0.1437 & -0.19332 & -0.3419 \\ -0.2157 & -0.2897 & -0.5133 \\ -0.2874 & -0.3866 & -0.6847 \end{bmatrix}$$

14. graph:



$$\frac{\partial L}{\partial v} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h2} \frac{\partial h2}{\partial v} = \sum_{i=1}^{m} \big(\hat{y}^{(i)} - y^{(i)}\big) \cdot \hat{y}^{(i)}(1 - \hat{y}^{(i)}) \cdot h_{1\_out}^{(i)}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h2} \frac{\partial h2}{\partial h_{1\_out}} \frac{\partial h_{1\_out}}{\partial h_1} \frac{\partial h1}{\partial w}$$

$$= \sum_{i=1}^{m} \big(\hat{y}^{(i)} - y^{(i)}\big) \cdot \hat{y}^{(i)}(1 - \hat{y}^{(i)}) \cdot v \cdot h_{1\_out}^{(i)}(1 - h_{1\_out}^{(i)}) \cdot x^{(i)}$$