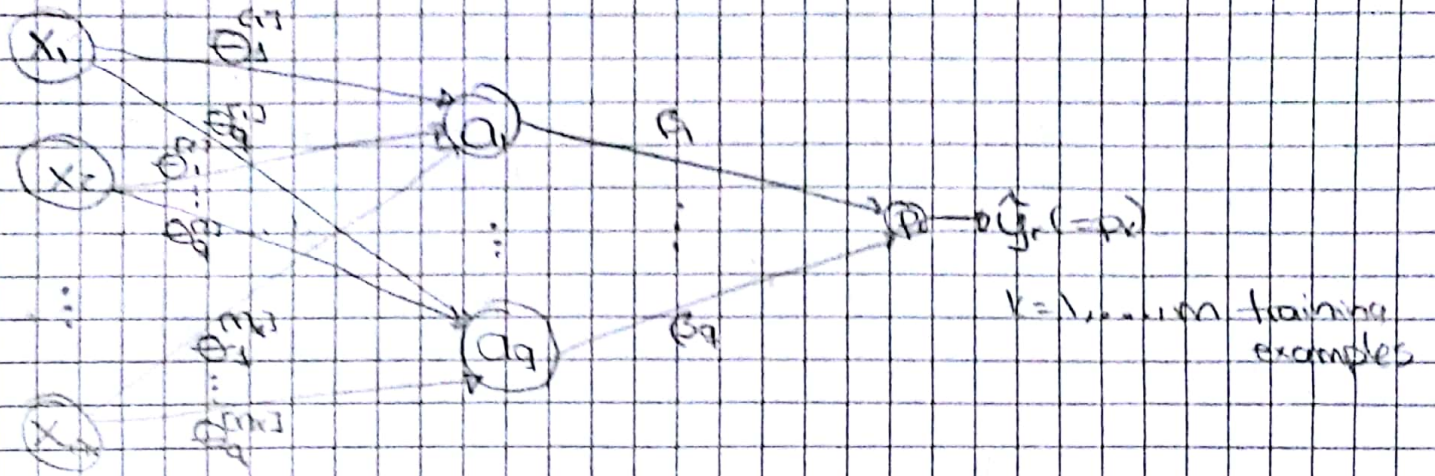
Backprop

Shallow neural network (1 hidden layer) for classification



$a_j = \sigma_1\left(\sum_{i=0}^{m} \theta_j^{[1]} x_i\right) = \sigma_1(z_j)$; $j = 1,\dots,q$; $\sigma_1(z) = \tanh(z)$ or $\sigma_1(z) = \frac{1}{1+e^{-z}}$; $x_i \in \mathbb{R}^m$

$a_0 = x_0 = 1$

$p = \sigma_2\left(\sum_{j=0}^{q} a_j \beta_j\right) = \sigma_2\left(\sum_{j=0}^{q} \beta_j \sigma_1\left(\sum_{i=0}^{m} \theta_j^{[1]} x_i\right)\right)$

with $\sigma_2(z) = \frac{1}{1+e^{-z}} = (1+e^{-z})^{-1} = \frac{e^z}{1+e^z}$

$p_k = \sigma_2(w_k) = \sigma_2\left(\sum_j a_{jk} \beta_j\right)$ $\quad * = (y-1)\log(1-\sigma_2(w)) - y\log(\sigma_2(w))$

Distance loss: $l_k(\hat{y}, y) = -\left(y_k \log(\hat{y}_k) + (1-y_k)\log(1-\hat{y}_k)\right) = *$

$J(\theta, \beta) = \frac{1}{m}\sum_{k=1}^{m} l_k(\hat{y}_k, y_k)$

We want $\frac{\partial J}{\partial \theta_j^{[1]}}$ and $\frac{\partial J}{\partial \beta_j}$ - $l_k(\hat{y}, y) = (y-1)\log(1-p_k) - y\log(p_k)$

$\frac{\partial l_k}{\partial \beta} = \frac{\partial l_k}{\partial p_k} \cdot \frac{\partial p_k}{\partial w} \cdot \frac{\partial w}{\partial \beta} = \left[\frac{y-1}{1-p_k}(-1) - \frac{y}{p_k}\right]\left[\sigma_2'(w)\right]\left[a_k\right]$

but $\sigma_2'(w) = -(1+e^{-w})^{-2}(-e^{-w}) = \frac{e^{-w}}{(1+e^{-w})^2} = \sigma_2(w)\frac{e^{-w}}{1+e^{-w}}$

$= \sigma_2(w)(1-\sigma_2(w)) = \sigma_2\left(\sum_{j=0}^{q} a_j \beta_j\right)\left(1-\sigma_2\left(\sum_{j=0}^{q} a_j \beta_j\right)\right)$

$= p_k(1-p_k)$

Then $\frac{\partial l_k}{\partial \beta_j} = \left[\frac{1-y_k}{1-p_k} - \frac{y_k}{p_k}\right]\left[p_k(1-p_k)\right]a_{jk} = \left[\frac{p_k - y_k p_k - y + y p_k}{(1-p_k)p_k}\right]\left(p_k(1-p_k)\right)a_{jk}$

$= (p_k - y_k)a_{jk}$

And $\frac{\partial J}{\partial \beta_j} = \frac{1}{m}\sum_{k=1}^{m}(p_k - y_k)a_{jk}$ with $a_{jk} = \sigma_1\left(\sum_{i=0}^{m} \theta_j^{[1]} x_{ik}\right)$, $x_k \in \mathbb{R}^m$

$$\frac{\partial \mathcal{L}_{ik}}{\partial \theta_l^{[n]}} = \frac{\partial \mathcal{L}_k}{\partial p_k} \cdot \frac{\partial p_k}{\partial w_k} \cdot \frac{\partial w_k}{\partial a_{lk}} \cdot \frac{\partial a_{lk}}{\partial z_{lk}} \cdot \frac{\partial z_{lk}}{\partial \theta_l^{[n]}}$$

$$= \left[\frac{p_k - y_k}{(1-p_k)p_k} \right] \left[p_k(1-p_k) \right] \left[\beta_l \right] \left[\sigma_1'(z_{lk}) \right] \left[x_{nk} \right]$$

with $\sigma_1'(z_{lk}) = \sigma_1(z_{lk})(1-\sigma_1(z_{lk}))$ if logistic

or $\sigma_1'(z_{lk}) =$ if tanh

$$= (p_k - y_k)(\beta_l)(\sigma_1(z_{lk}))(1-\sigma_1(z_{lk}))x_{nk}$$

$$\therefore \frac{\partial \mathcal{L}}{\partial \theta_l^{[n]}} = \frac{1}{m} \sum_{k=1}^{m} \left[(p_k - y_k)(\beta_l)(\sigma_1(z_{lk}))(1-\sigma_1(z_{lk}))x_{nk} \right]$$

σ_{lk} $\quad$ $-\alpha_{lk}$

$n = 1, 2, \ldots, n_x \quad n=4$

$l = 1, 2, \ldots, q \quad q=3$

$k = 1, 2, \ldots, m \quad m=100$

$$z_{lk} = \sum_{i=0}^{n_x} \theta_l^{[n]} x_{ik} \Rightarrow \frac{\partial z_{lk}}{\partial \theta_l^{[n]}} = x_{nk}$$

$A \in \mathbb{R}^{100 \times 3}$

```
for l = 1, 2, 3 {
    for n = 1, 2, 3, 4 {
        dLk_dtheta = (p - y)(beta[l])(A[:, l])(1 - A[:, l])X_n
```

(p − y): 100×1 · beta[l]: 1×1 · A[:, l]: 100×1 · (1 − A[:, l]): 100×1 · X_n: 100×1

m = # training examples
n_x = # of variables

$$A = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_{n_x}^{(1)} \\ x_1^{(2)} & x_2^{(2)} & & \\ \vdots & & & \\ x_1^{(m)} & x_2^{(m)} & \dots & x_{n_x}^{(m)} \end{bmatrix} \in \mathbb{R}^{m \times n_x}$$

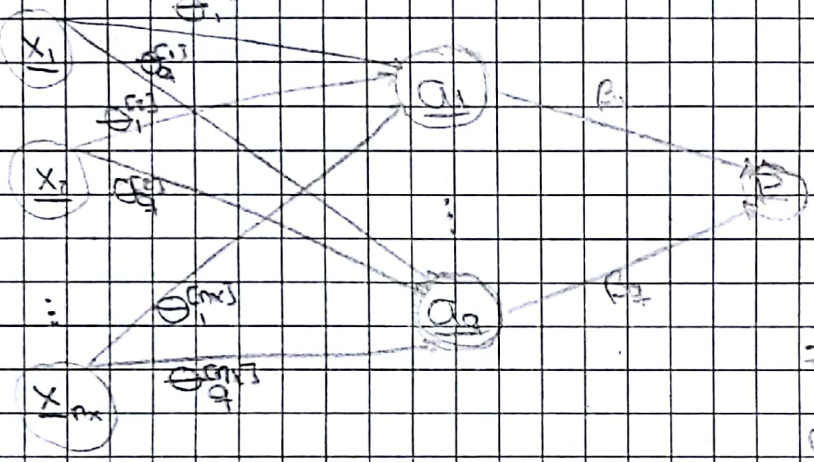
x_1 = vector of obs

$$\begin{bmatrix} 1 & 1 & & 1 \\ x_1 & x_2 & \dots & x_{n_x} \\ 1 & 1 & & 1 \end{bmatrix} = \begin{bmatrix} - & x^{(1)} & - \\ - & x^{(2)} & - \\ & \vdots & \\ - & x^{(m)} & - \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \{0,1\}^m$$

$x_i \in \mathbb{R}^m$

$x_i = 1$

$$\Theta = \begin{bmatrix} \theta_1^{(1)} & \theta_2^{(1)} & \dots & \theta_{n_x}^{(1)} \\ \theta_1^{(2)} & \theta_2^{(2)} & \dots & \theta_{n_x}^{(2)} \end{bmatrix}$$



$$z_1^{(i)} = \sum_{i=1}^{n_x} \theta_1^{(i)} x_i^{(i)}$$

$$a_1^{(i)} = \sigma\left(\sum_{i=1}^{n_x} \theta_1^{(i)} x_i^{(i)}\right)$$

$$\mathbb{R}^m \to a_1 = \sigma\left(\sum_{i=1}^{n_x} \theta_1^{[i]} x_i\right) = \sigma(z_1) \qquad \text{with } z_1 = \sum_{i=1}^{n_x} \theta_1^{[i]} x_i$$

$k = 1, \dots, q$; $\theta_1^{[i]} \in \mathbb{R}^{n_x}$; $x_i \in \mathbb{R}^{n_x}$

and $\sigma(u) = \dfrac{1}{1+e^{-u}}$

$$\mathbb{R}^m \to \hat{P} = \sigma\left(\sum_{l=1}^{q} \beta_l a_l\right) = \sigma(u) \qquad \text{with } u = \sum_{l=1}^{q} \beta_l a_l$$

$\beta_l \in \mathbb{R}$; $a_l \in \mathbb{R}^m$

and $u_k = \sum_{l=1}^{q} \beta_l a_l^{(k)}$

$\hat{P}_k = \sigma(u_k)$

Loss function: $\ell(\Theta, \beta) = -\left(y \log(\hat{P}) + (1-y)\log(1-\hat{P})\right)$

$$L(\Theta, \beta) = \text{mean}(\ell(\Theta, \beta))$$

$$\ell_k(\Theta, \beta) = -\left(y^{(k)} \log(\hat{P}_k) + (1-y^{(k)}) \log(1-\hat{P}_k)\right)$$

$$\frac{\partial \ell_k}{\partial \beta_l} = \frac{\partial \ell_k}{\partial \hat{P}_k} \cdot \frac{\partial \hat{P}_k}{\partial u_k} \cdot \frac{\partial u_k}{\partial \beta_l} = \left[\frac{1-y^{(k)}}{1-\hat{P}_k} - \frac{y^{(k)}}{\hat{P}_k}\right]\left[\frac{\sigma(u_k)(1-\sigma(u_k))}{\hat{P}_k}\right]\left[a_l^{(k)}\right]$$

$$= \left[\frac{\hat{P}_k - \hat{P}_k y^{(k)} - y^{(k)} + \hat{P}_k y^{(k)}}{(1-\hat{P}_k)\hat{P}_k}\right]\left[\hat{P}_k(1-\hat{P}_k)\right] a_l^{(k)} = (\hat{P}_k - y^{(k)}) a_l^{(k)}$$

$$\frac{\partial L}{\partial \beta_l} = \underbrace{(\hat{P} - y)}_{\mathbb{R}^m} \underbrace{a_l}_{\mathbb{R}^m} \in \mathbb{R}^m$$

$$\frac{\partial l_k}{\partial \theta_h^{(m)}} = \frac{\partial l_k}{\partial p_k} \cdot \frac{\partial p_k}{\partial u_k} \cdot \frac{\partial u_k}{\partial a_h^{(k)}} \cdot \frac{\partial a_h^{(k)}}{\partial z_h^{(k)}} \cdot \frac{\partial z_h^{(k)}}{\partial \theta_h^{(m)}}$$

$$= [\hat{p}_k - y^{(k)}] [\beta_h] [\underbrace{\sigma(z_h^{(k)})}_{a_h^{(k)}} (1 - \underbrace{\sigma(z_h^{(k)})}_{a_h^{(k)}})] x_n^{(k)}$$

$h = 1, \ldots, q$
$n = 1, \ldots, n_k$
$k = 1, \ldots, m$

$$\frac{\partial l}{\partial \theta_h^{(m)}} = \underbrace{(\hat{p} - y)}_{\mathbb{R}^m} \underbrace{\beta_h}_{\mathbb{R}} \underbrace{[a_h (1 - a_h)]}_{\mathbb{R}^m} \underbrace{x_n}_{\mathbb{R}^m}$$

$n_k \times q \times m$ elements

Let $\text{temp}_h = (\hat{p} - y) \beta_h [a_h (1 - a_h)] \in \mathbb{R}^m \Rightarrow \frac{\partial l}{\partial \theta_h^{(m)}} = \underbrace{\text{temp}_h}_{\mathbb{R}^m} \cdot \underbrace{x_n}_{\mathbb{R}^m} \in \mathbb{R}^m$

Then $T_h = \begin{bmatrix} \text{temp}_1 & \text{temp}_2 & \cdots & \text{temp}_{n_k} \end{bmatrix} \in \mathbb{R}^{m \times n_k}$

for h in $1, \ldots, q$:

T_h $*$ X

$n \times n_k$ ↑ element-wise $m \times n_k$