Brian Bertness
1478201

1. We are given the following:

$$E(w, v|X) = -\sum_{t=1}^{N}\left[r^t \log y^t + (1-r^t)\log(1-y^t)\right] + \sum_{h=1}^{H} \|w_h\|_2^2$$

$$y^t = \text{Sigmoid}\left(\sum_{h=1}^{H} v_h z_h^t + v_0\right)$$

$$z_h^t = \text{reLU}(w_h^T + w_{h0})$$

H1 Hint: Given reLU(f(x)) = max(0, f(x)) the derivative of reLU(f(x)) is f'(x) if f(x) > 0 and 0 otherwise.

H2 Given $y = \text{Sigmoid}(\alpha) = 1/(1+e^{-\alpha})$, $\frac{\partial y}{\partial \alpha} = y(1-y)$

$v_h = v_h + \Delta v_h$ where $\Delta v_h = -\eta \frac{\partial E}{\partial v_h}$

We find $\frac{\partial E}{\partial v_h}$

$$\frac{\partial E}{\partial v_h} = \frac{\partial E}{\partial y^t}\frac{\partial y^t}{\partial \alpha^t}\frac{\partial \alpha^t}{\partial v_h}$$ We find each partial derivative in turn.

① $$\frac{\partial E}{\partial y^t} = -\sum_{t=1}^{N} \frac{r^t}{y^t} - \frac{1-r^t}{1-y^t}$$ (Note: the last term $\sum_{h=1}^{H}\|W_h\|_2^2$ has $\partial E/\partial y^t = 0$)

② $$\frac{\partial y^t}{\partial \alpha^t} = y^t(1-y^t)$$ by H_2 above

Since $y^t = \text{sigmoid}\left(\sum_{h=1}^{H} v_h Z_h^t + v_0\right)$ we know

$$\alpha = \sum_{h=1}^{H} v_h Z_h^t + v_0$$

So,

(3) $$\frac{\partial \alpha^t}{\partial v_h} = Z_h^t$$

Thus $$\Delta v_h = -\eta \frac{\partial E}{\partial y^t} \frac{\partial y^t}{\partial \alpha^t} \frac{\partial \alpha^t}{\partial v_h}$$

$$= \eta \sum_{t=1}^{N} \frac{r^t}{y^t} - \frac{1-r^t}{1-y^t} \cdot y^t(1-y^t) Z_h^t$$ (by 1, 2 and 3)

$$= \eta \sum_{t=1}^{N} \frac{r^t(1-y^t) - y^t(1-r^t)}{\cancel{y^t(1-y^t)}} \cancel{y^t(1-y^t)} Z_h^t$$

$$= \eta \sum_{t=1}^{N} (r^t - \cancel{r^t y^t} - y^t + \cancel{r^t y^t}) Z_h^t$$

$$= \eta \sum_{t=1}^{N} (r^t - y^t) Z_h^t$$

$$\boxed{\begin{aligned} \text{Thus, } v_h &= v_h + \Delta v_h = v_h + -\eta \frac{\partial E}{\partial v_h} \\ &= v_h + \eta \sum_{t=1}^{N} (r^t - y^t) Z_h^t \end{aligned}}$$

$\underline{v_0 = v_0 + \Delta v_0}$ where $\Delta v_0 = -\eta \frac{\partial E}{\partial v_0}$

We find $\frac{\partial E}{\partial v_0}$

$$\frac{\partial E}{\partial v_0} = \frac{\partial E}{\partial y^t} \frac{\partial y^t}{\partial \alpha^t} \frac{\partial \alpha^t}{\partial v_0}$$

We note that $\frac{\partial E}{\partial y^t}$ and $\frac{\partial y^t}{\partial \alpha^t}$ have been derived in 1 and 2 above.

4 $\frac{\partial \alpha^t}{\partial v_0} = 1$ since $\alpha = \sum_{h=1}^{H} v_h z_h^t + v_0$

Thus, $\Delta v_0 = -\eta \frac{\partial E}{\partial v_0} \frac{\partial y^t}{\partial \alpha^t} \frac{\partial \alpha^t}{\partial v_h}$

$$= \eta \sum_{t=1}^{N} (r^t - y^t) \quad (\text{by } 1, 2, \text{ and } 4)$$

$$\boxed{\text{Thus, } v_0 = v_0 + \Delta v_0 = v_0 + -\eta \frac{\partial E}{\partial v_0} = v_0 + \eta \sum_{t=1}^{N} (r^t - y^t)}$$

$\underline{w_h = w_h + \Delta w_h}$ where $\Delta w_h = -\eta \dfrac{\partial E}{\partial w_h}$

We have $z^t_h = \text{reLU}(\sigma)$ where $\sigma = w_h^T x^t + w_0$
and $y^t = \text{Sigmoid}(\alpha)$ where $\alpha = v_h z^t_h + v_0$

We want to find $\dfrac{\partial E}{\partial w_h}$ for both terms in $E(w, v \mid X)$.

For the first term:

$$\frac{\partial E}{\partial w_h} = \frac{\partial E}{\partial y^t} \frac{\partial y^t}{\partial \alpha^t} \frac{\partial \alpha^t}{\partial z_h} \frac{\partial z^t_h}{\partial \sigma} \frac{\partial \sigma}{\partial w_h}$$

We have already calculated $\dfrac{\partial E}{\partial y^t}$ and $\dfrac{\partial y^t}{\partial \alpha^t}$ in 1 and 2

5 We have $\dfrac{\partial \alpha^t}{\partial z^t_h} = v_h$

6 $\dfrac{\partial z^T_h}{\partial \sigma}$ represents the conditional on the piecewise reLU function. That is if $w_h^T x^t + w_0 > 0$ then reLU will return $f'(x)$ otherwise it will return 0.

7 Also, $\dfrac{\partial \sigma}{\partial w^h} = x^t$

For the second term we want to find $\dfrac{\partial E}{\partial w_h}$.

We expand the sum:

$$\sum_{h=1}^{H} \|W_h\|_2^2 = \|W_1\|_2^2 + \|W_2\|_2^2 + \dots + \|W_H\|_2^2$$

$$= W_1^2 + W_2^2 + \dots + W_H^2$$

8 Taking the derivative $\frac{\partial E}{\partial w_h}$ we get

$$= 2w_1 + 2w_2 + \dots + 2w_H$$

$$= \sum_{h=1}^{H} 2w_h$$

Thus, by 1 2 5 6 7 8 we get

$$\Delta w_h = -\eta \frac{\partial E}{\partial w_h} = \eta \sum_{t=1}^{N} (r^t - y^t) v_h x^t + \sum 2w_h \quad \text{if } w_h^T x^t + w_0 > 0$$

or

0 otherwise

with complete update equation being

$w_h = w_h + \Delta w_h$ where Δw_h is as defined above.

$w_0 = w_0 + \Delta w_0$ is the same as above except $\frac{\partial \sigma}{\partial w_0} = 1$ and $\frac{\partial E}{\partial w_h}$ of the second term is 0.

Thus,

$$w_0 = w_0 + \Delta w_0 = w_0 + -\eta \frac{\partial E}{\partial w_0}$$

$$= w_0 + \eta \sum_{t=1}^{N} (r^t - y^t) v_h x^t \quad \text{if } w_h^T x^t + w_0 > 0$$

otherwise it is just

$$= w_0 + 0 = w_0$$

2a. We determine the vector of coefficients for Figure 2.

The equation of the line which separates the two classes is

$$x_1 = 2 \Rightarrow x_1 - 2 = 0 \Rightarrow 0x_2 + x_1 - 2 = 0$$

Hence the options for the activation functions are

(1) $$y = S(0x_2 + x_1 - 2)$$

or

(2) $$y = S(0x_2 - x_1 + 2)$$

By the figure we want the activation function to return 0 for $(\underset{x_1}{0}, \underset{x_2}{0})$. Thus

$$S(0(0) + (0) - 2) = S(-2) = 0 \quad \text{Test (1)}$$
$$S(0(0) - (0) + 2) = S(2) = 1 \quad \text{Test (2)}$$

We see that (1) is the correct activation function. So, $y = S(0x_2 + x_1 - 2)$

$\vec{w}$ $$\Rightarrow \vec{w} = [-2, 1, 0]^T \text{ where } \vec{w} = [w_0, w_1, w_2]^T$$

2a. (cont'd) We determine the vector of coefficients for Figure 3.

The equation of the line which separates the two classes is

$$x_1 - 4x_2 = 0 \Rightarrow -4x_2 + x_1 + 0 = 0$$

Hence, the options for the activation functions are

(3) $y = S(-4x_2 + x_1 + 0)$

or

(4) $y = S(4x_2 - x_1 + 0)$

By the figure we want the activation function to return a 1 for $(\underset{x_1}{1}, \underset{x_2}{0})$. Thus,

$S(-4(0) + 1 + 0) = S(1) = 1$ Test (3)

$S(4(0) - 1 + 0) = S(-1) = 0$ Test (4)

We see that (3) is the correct activation function. So, $y = S(-4x_2 + x_1 + 0)$

$\vec{w}$ $\Rightarrow \vec{w} = [0, 1, -4]^T$ where $\vec{w} = [w_0, w_1, w_2]^T$

2b. We determine coefficients W and $\vec{v}$ in the two-layer perceptron in Figure 4 to recognize the shaded region in Figure 5.

We see that the shaded region in Figure 5 is equal to the intersection of the regions of figures 2 and 3.

Thus, the equations of the lines which represent the shaded areas are

$$z_1 = 0x_2 + x_1 - 2$$

and

$$z_2 = -4x_2 + x_1 + 0$$

So $y = 1$ when

$$z_1 = 1$$

and

$$z_2 = 1$$

And so we choose a middle point between 1 and 2. Since 1.5 splits the difference we have

$$z_2 + z_1 = 1.5 \Rightarrow z_2 + z_1 - 1.5 = 0$$

2b. (cont'd) Thus, the options for the activation functions are

(5) $S(z_2 + z_1 - 1.5)$

or

(6) $S(-z_2 - z_1 + 1.5)$

We want the activation function to return a 1 for (1, 1) and a 0 for (0, 0). We see that (5) does both of these while (6) does not!

$S(1+1-1.5) = S(.5) = 1$ Tests for 5

$S(0+0-1.5) = S(-1.5) = 0$

$S(-1-1+1.5) = S(-.5) = 0$ Tests for 6

$S(-0-0+1.5) = S(1.5) = 1$

Thus: $y = S(z_2 + z_1 - 1.5)$

$\vec{v}$ $\Rightarrow \vec{v} = [-1.5, 1, 1]^T$ where $\vec{v} = [v_0, v_1, v_2]^T$

and

W $W = \begin{bmatrix} -2 & 1 & 0 \\ 0 & 1 & -4 \end{bmatrix}^T$

3.a) Programming Piece

b)

Output from hw3:

```
Validation accuracy for 4 hidden units is 0.826
Validation accuracy for 16 hidden units is 0.915
Validation accuracy for 20 hidden units is 0.912
Validation accuracy for 24 hidden units is 0.919
Validation accuracy for 32 hidden units is 0.923
Validation accuracy for 48 hidden units is 0.900
Test accuracy with 32 hidden units is 0.912
```

Output from my code run which reports the validation accuracy by the number of hidden units.

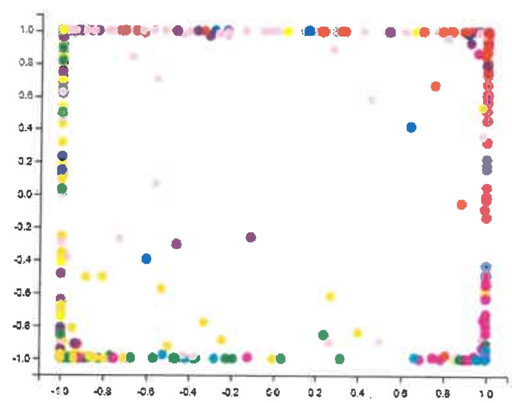
We should use 32 hidden units since this had the highest validation accuracy.

The accuracy on the test set with 32 hidden units is 0.912.

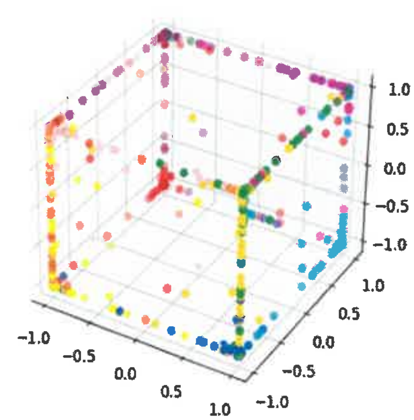
We see that as the number of hidden units increases the validation accuracy increases, up to a certain point. This is due to hidden units adding more dimensions which allows for better classification. At a certain point, though, adding more hidden units will result in an overtrained model. Also, more hidden units implies a more complex model so gradient descent could find a less than optimal local minimum.

3. c)

Visualization with 2 hidden units:



Visualization with 3 hidden units:



With two hidden units we only have two dimensions to classify the data. This results in poor separation of the classes. With 3 hidden units there is another dimension which the classifier can use for separation. Thus, the separation of classes is better in the 3d plot, corresponding to 3 hidden units, than in the 2d plot, corresponding to two hidden units.