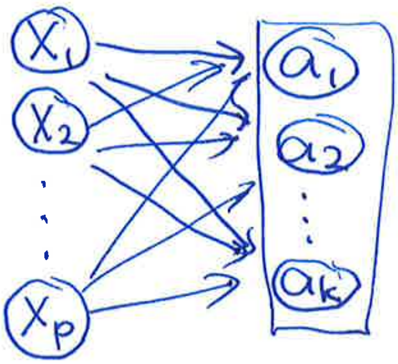
Previously:

Multinomial Logistic Regression: K classes



For m observations, in columns,

$$\begin{bmatrix} z_1^{(1)} & z_1^{(2)} & \cdots & z_1^{(m)} \\ z_2^{(1)} & z_2^{(2)} & \cdots & z_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ z_K^{(1)} & z_K^{(2)} & \cdots & z_K^{(m)} \end{bmatrix} = \begin{bmatrix} b_1+w_1^T x^{(1)} & b_1+w_1^T x^{(2)} & \cdots & b_1+w_1^T x^{(m)} \\ b_2+w_2^T x^{(1)} & b_2+w_2^T x^{(2)} & \cdots & b_2+w_2^T x^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ b_K+w_K^T x^{(1)} & b_K+w_K^T x^{(2)} & \cdots & b_K+w_K^T x^{(m)} \end{bmatrix}$$

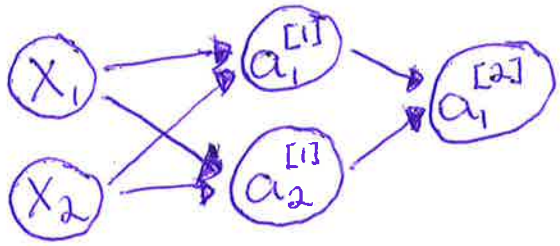
$$= \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} + \begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_K^T \end{bmatrix} \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & & | \end{bmatrix}$$

(using broadcasting for b vector)

For a sigmoid activation apply to each column of the z matrix
(each column sums to 1, each column corresponds to one observation's probability of being in each class)

Neural Network example model from day 1:

- 2 inputs
- 1 hidden layer with 2 units and tanh activation
- Output layer with 1 unit and sigmoid activation



square bracket notation says which layer; subscript is which unit in that layer

$a_2^{[1]}$ is the second unit in the first layer

convention: $a_1^{[0]} = x_1, \ldots, a_p^{[0]} = x_p$

Each circle means:

- calculate z as linear combination of outputs from previous layer
- calculate $a = g(z)$

layer 1

$$z_1^{[1]} = b_1^{[1]} + (w_1^{[1]})^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \quad a_1^{[1]} = \tanh(z_1^{[1]})$$

$$z_2^{[1]} = b_2^{[1]} + (w_2^{[1]})^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad a_2^{[1]} = \tanh(z_2^{[1]})$$

$$\boxed{z^{[1]} = b^{[1]} + W^{[1]T} a^{[0]} \text{ where } a^{[0]} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} \quad a^{[1]} = \tanh(z^{[1]})$$

layer 2

$$z_1^{[2]} = b_1^{[2]} + (w_1^{[2]})^T \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \end{bmatrix} \rightarrow z^{[2]} = b^{[2]} + w^{[2]T} a^{[1]}$$

$$a_1^{[2]} = \sigma(z_1^{[2]}) \rightarrow a^{[2]} = \sigma(z^{[2]})$$

In general, $z^{[l]} = b + W^T a^{[l-1]}$

- $z^{[l]}$: column vector of z's for layer l, length = # of units in that layer, n_l
- b: column vector of bias, length n_l
- $a^{[l-1]}$: column vector of activations from previous layer
- W^T is n_l by n_{l-1}

$$a^{[l]} = g^{[l]}(z^{[l]})$$

First Day:

Multiple Layers with non-linear transformations are helpful (the whole idea)

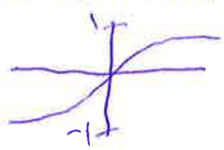
General notation:

- $a_j^{(i)[l]}$ is the activation value for unit j in layer l for observation i
- $b_j^{[l]}$ is the bias for unit j in layer l
- $w_j^{[l]}$ is the <u>vector</u> of weights for unit j in layer l
- $g^{[l]}$ is the activation function for layer l
- n_l is the number of units in layer l

---

2 common choices for activation functions in hidden layers

- $\tanh(z) = \dfrac{e^{2z} - 1}{e^{2z} + 1}$



- $\text{relu}(z) = \max(0, z)$

↑ rectified linear unit