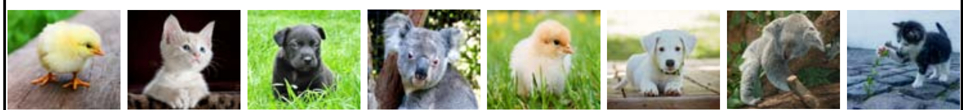
8. Multi-class Classification

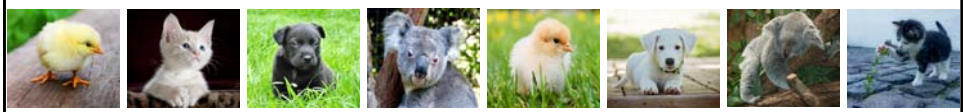
Most of this material is from Prof. Andrew Ng'and Chang's slides.

Recognizing cats, dogs, baby chicks, and others



3 1 2 0 3 2 0 1

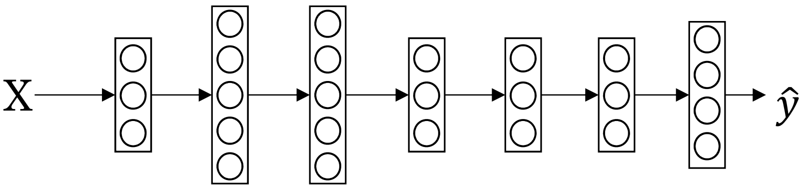
Recognizing cats, dogs, baby chicks, and others



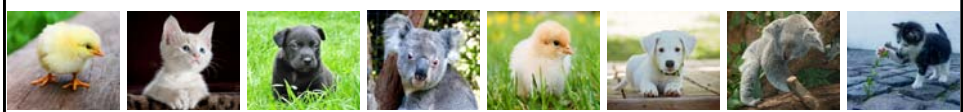
3 1 2 0 3 2 0 1

C = #classes = 4

$n^{[L]} = 4 = C$



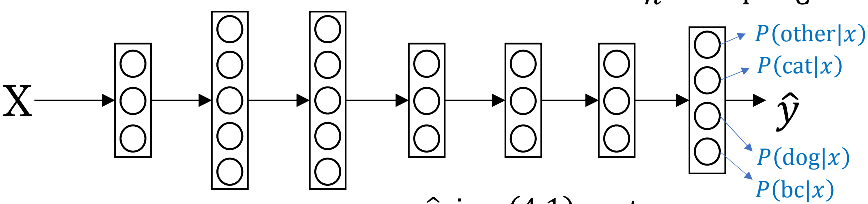
Recognizing cats, dogs, baby chicks, and others



3 1 2 0 3 2 0 1

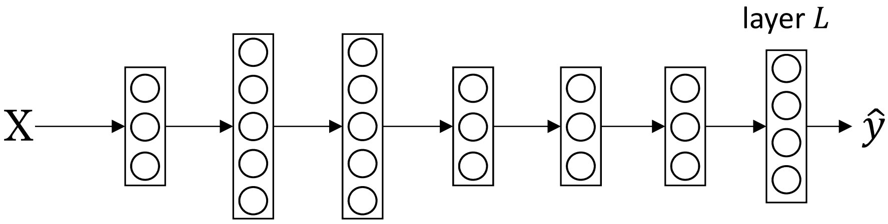
C = #classes = 4

$n^{[L]} = 4 = C$



$\hat{y}$ is a (4,1) vector

Recognizing cats, dogs, baby chicks, and others



(4,1)

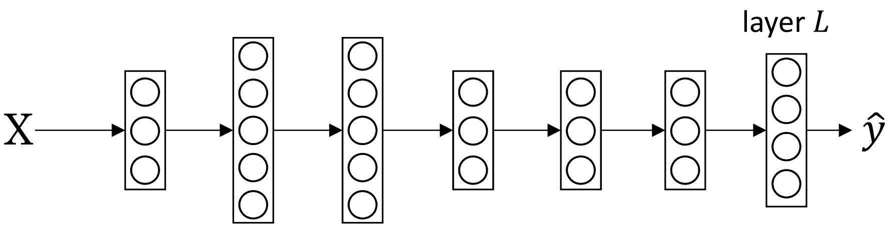
$$Z^{[L]} = W^{[L]} a^{[L-1]} + b^{[L]}$$

Activation function:

(4,1) $t = e^{(z^{[L]})}$

(4,1) $a^{[L]} = \frac{e^{(z^{[L]})}}{\sum_{j=1}^{4} t_j} \rightarrow a_i^{[L]} = \frac{e^{\left(z_i^{[L]}\right)}}{\sum_{j=1}^{4} e^{\left(z_j^{[L]}\right)}}$

Recognizing cats, dogs, baby chicks, and others



(4,1)

$$Z^{[L]} = W^{[L]} a^{[L-1]} + b^{[L]}$$

Activation function:

(4,1) $t = e^{(z^{[L]})}$

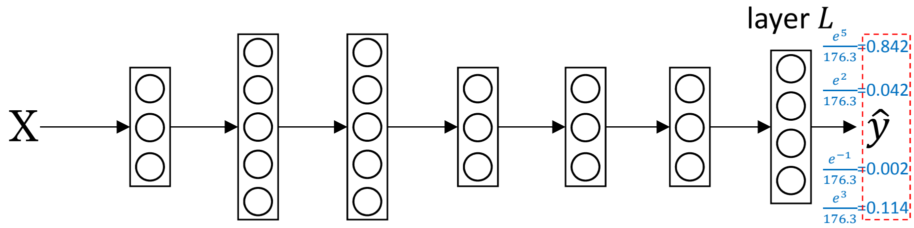
(4,1) $a^{[L]} = \frac{e^{(z^{[L]})}}{\sum_{j=1}^{4} t_j} \rightarrow a_i^{[L]} = \frac{e^{\left(z_i^{[L]}\right)}}{\sum_{j=1}^{4} e^{\left(z_j^{[L]}\right)}}$

$$z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

$$t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix} \qquad \sum_{j=1}^{4} t_j = 176.3$$

$$a^{[L]} = \frac{t}{176.3}$$

Recognizing cats, dogs, baby chicks, and others



(4,1)

$$Z^{[L]} = W^{[L]} a^{[L-1]} + b^{[L]}$$

Activation function:

(4,1) $t = e^{(Z^{[L]})}$

(4,1) $a^{[L]} = \frac{e^{(Z^{[L]})}}{\sum_{j=1}^{4} t_j} \rightarrow a_i^{[L]} = \frac{e^{\left(Z_i^{[L]}\right)}}{\sum_{j=1}^{4} e^{\left(Z_j^{[L]}\right)}}$

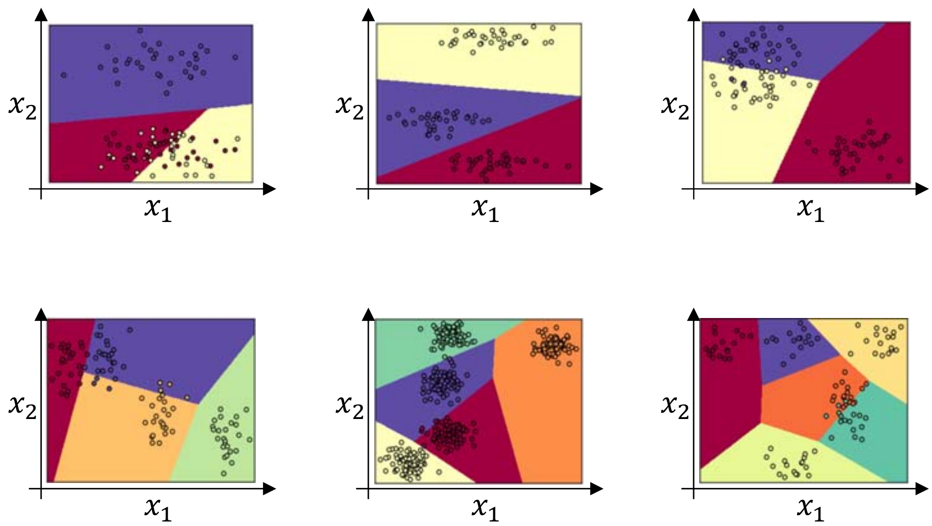
$= \hat{y}$

$$Z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

$$t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix} \quad \sum_{j=1}^{4} t_j = 176.3$$

$$a^{[L]} = \frac{t}{176.3}$$

(1-layer) Softmax classifier examples



Understanding softmax

$$z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \qquad t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix}$$

$$g^{[L]}(z^{[L]}) = \begin{bmatrix} e^5/(e^5+e^2+e^{-1}+e^3) \\ e^2/(e^5+e^2+e^{-1}+e^3) \\ e^{-1}/(e^5+e^2+e^{-1}+e^3) \\ e^3/(e^5+e^2+e^{-1}+e^3) \end{bmatrix} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}$$

Understanding softmax

$$z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \qquad t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix}$$

"soft max" "hard max"

$$g^{[L]}(z^{[L]}) = \begin{bmatrix} e^5/(e^5+e^2+e^{-1}+e^3) \\ e^2/(e^5+e^2+e^{-1}+e^3) \\ e^{-1}/(e^5+e^2+e^{-1}+e^3) \\ e^3/(e^5+e^2+e^{-1}+e^3) \end{bmatrix} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix} \quad \text{vs.} \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Understanding softmax

$$z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \qquad t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix}$$

$$g^{[L]}(z^{[L]}) = \begin{bmatrix} e^5/(e^5+e^2+e^{-1}+e^3) \\ e^2/(e^5+e^2+e^{-1}+e^3) \\ e^{-1}/(e^5+e^2+e^{-1}+e^3) \\ e^3/(e^5+e^2+e^{-1}+e^3) \end{bmatrix} = \underset{\text{"soft max"}}{\begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}} \quad \text{vs.} \quad \underset{\text{"hard max"}}{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}$$

Softmax regression generalizes logistic regression to C classes.

Understanding softmax

$$z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \qquad t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix}$$

$$g^{[L]}(z^{[L]}) = \begin{bmatrix} e^5/(e^5+e^2+e^{-1}+e^3) \\ e^2/(e^5+e^2+e^{-1}+e^3) \\ e^{-1}/(e^5+e^2+e^{-1}+e^3) \\ e^3/(e^5+e^2+e^{-1}+e^3) \end{bmatrix} = \underset{\text{"soft max"}}{\begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}} \quad \text{vs.} \quad \underset{\text{"hard max"}}{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}$$

Softmax regression generalizes logistic regression to C classes.

If $\mathrm{C} = 2$, softmax essentially reduces to logistic regression.

Loss function

assume

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad a^{[L]} = \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

Loss function

assume

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad a^{[L]} = \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{4} y_j \log \hat{y}_j$$

Loss function

assume

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad a^{[L]} = \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{4} y_j \log \hat{y}_j$$

$$= -y_2 \log \hat{y}_2 = -\log \hat{y}_2$$

to make $\mathcal{L}(\hat{y}, y)$ small,
we should make $\hat{y}_2$ big.

minimizing loss function is justified by maximum likelihood principle
→ find $\text{argmin}_{W,b}(-\log \hat{y}_i)$ = $\text{argmax}_{W,b}(\hat{y}_i)$ = $\text{argmax}_{W,b}(p(y_i|W,b))$ such that $y_i = 1$

Loss function

assume

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad a^{[L]} = \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{4} y_j \log \hat{y}_j$$

$$= -y_2 \log \hat{y}_2 = -\log \hat{y}_2$$

to make $\mathcal{L}(\hat{y}, y)$ small,
we should make $\hat{y}_2$ big.

$$J(W^{[1]}, b^{[1]}, \dots) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

Loss function

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad a^{[L]} = \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{4} y_j \log \hat{y}_j$$

$$= -y_2 \log \hat{y}_2 = -\log \hat{y}_2$$

to make $\mathcal{L}(\hat{y}, y)$ small,
we should make $\hat{y}_2$ big.

$$J(W^{[1]}, b^{[1]}, \dots) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

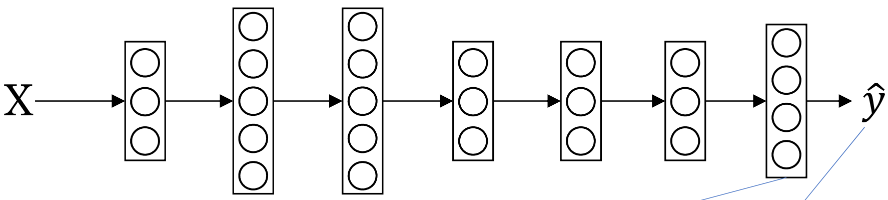
$Y = [y^{(1)} \quad y^{(2)} \quad \dots \quad y^{(m)}]$

$(4, m)$

$\hat{Y} = [\hat{y}^{(1)} \quad \hat{y}^{(2)} \quad \dots \quad \hat{y}^{(m)}]$

$(4, m)$

Gradient descent with softmax



$z^{[L]} \longrightarrow a^{[L]} = \hat{y} \longrightarrow \mathcal{L}(\hat{y}, y)$

softmax loss function

Forward propagation is trivial.

Backpropagation: $dz^{[L]} = \dfrac{\partial J}{\partial z^{[L]}} = \hat{y} - y$

$(4,1)$

Proof of $dz^{[L]} = \frac{\partial J}{\partial z^{[L]}} = \hat{y} - y$

$$da = \frac{\partial J}{\partial a} = p - y$$

For simplicity, set $p = \hat{y}$ and $p_i = \hat{y}_i$

$a = z^{[L]}$ and $a_i = z_i^{[L]}$

$$\therefore p_i = \frac{e^{(a_i)}}{\sum_k e^{(a_k)}}$$

$$\begin{aligned}
\frac{\partial p_i}{\partial a_i} &= \frac{\partial \frac{exp(a_i)}{\sum_k exp(a_k)}}{\partial a_i} \\
&= \frac{exp(a_i)\sum_k exp(a_k) - exp(a_i)\,exp(a_i)}{\left(\sum_k exp(a_k)\right)^2} \\
&= \frac{exp(a_i)\left[\sum_k \{exp(a_k)\} - exp(a_i)\right]}{\left(\sum_k exp(a_k)\right)^2} \\
&= \frac{exp(a_i)}{\sum_k exp(a_k)} \frac{\sum_k \{exp(a_k)\} - exp(a_i)}{\sum_k exp(a_k)} \\
&= \frac{exp(a_i)}{\sum_k exp(a_k)} \left(1 - \frac{exp(a_i)}{\sum_k exp(a_k)}\right) \\
&\boxed{= p_i(1 - p_i)}
\end{aligned}$$

for $i \neq j$,

$$\begin{aligned}
\frac{\partial p_i}{\partial a_j} &= \frac{0 - exp(a_i)\,exp(a_j)}{\left(\sum_k exp(a_k)\right)^2} \\
&= -\frac{exp(a_i)}{\sum_k exp(a_k)} \frac{exp(a_j)}{\sum_k exp(a_k)} \\
&\boxed{= -p_i p_j}
\end{aligned}$$

Proof of $da = \frac{\partial J}{\partial a} = p - y$

$$\begin{aligned}
\frac{\partial J}{\partial a_i} &= \frac{\partial\left(-\sum_j y_j \log p_j\right)}{\partial a_i} \\
&= -\sum_j y_j \frac{\partial \log p_j}{\partial a_i} \\
&= -\sum_j y_j \frac{1}{p_j}\frac{\partial p_j}{\partial a_i} = -\frac{y_i}{p_i} p_i(1 - p_i) - \sum_{i \neq j} \frac{y_j}{p_j}(-p_i p_j) \\
&= -y_i + y_i p_i + \sum_{i \neq j} y_j p_i \\
&= -y_i + \sum_j y_j p_i \\
&= -y_i + p_i \sum_j y_j \\
&= p_i - y_i
\end{aligned}$$

$$\therefore \frac{\partial J}{\partial a} = [p_i - y_i] = p - y \Rightarrow dz^{[L]} = \frac{\partial J}{\partial z^{[L]}} = \hat{y} - y$$