# Unit 6 Softmax Regression
# Part 01 Softmax Regression

TFIP-AI Artificial Neural Networks and Deep Learning

# Softmax Regression Definition

Softmax regression (or multinomial logistic regression) is a generalization of logistic regression to the case where we want to handle multiple classes. In logistic regression we assumed that the labels were binary: $y(i) \in \{0,1\}$. For example, we can use such a classifier to distinguish between two kinds of hand-written digits. Softmax regression allows us to handle $y(i) \in \{1,\ldots,K\}$ where K is the number of classes.

# Softmax Regression Definition cont...

Recall that in logistic regression, we had a training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ of $m$ labeled examples, where the input features are $x^{(i)} \in \Re^n$. With logistic regression, we were in the binary classification setting, so the labels were $y^{(i)} \in \{0, 1\}$. Our hypothesis took the form:

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^\top x)},$$

and the model parameters $\theta$ were trained to minimize the cost function

$$J(\theta) = -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]$$

# Softmax Regression Definition cont...

In the softmax regression setting, we are interested in multi-class classification (as opposed to only binary classification), and so the label y can take on K different values, rather than only two. Thus, in our training set $\{(x(1),y(1)),\ldots,(x(m),y(m))\}$, we now have that $y(i)\in\{1,2,\ldots,K\}$. (Note that our convention will be to index the classes starting from 1, rather than from 0.) For example, in the MNIST digit recognition task, we would have K=10 different classes.

# Softmax Regression Definition cont...

Given a test input $x$, we want our hypothesis to estimate the probability that $P(y = k|x)$ for each value of $k = 1, \ldots, K$. I.e., we want to estimate the probability of the class label taking on each of the $K$ different possible values. Thus, our hypothesis will output a $K$-dimensional vector (whose elements sum to 1) giving us our $K$ estimated probabilities. Concretely, our hypothesis $h_\theta(x)$ takes the form:

$$h_\theta(x) = \begin{bmatrix} P(y = 1|x;\theta) \\ P(y = 2|x;\theta) \\ \vdots \\ P(y = K|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x)} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \\ \vdots \\ \exp(\theta^{(K)\top} x) \end{bmatrix}$$

Here $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)} \in \Re^n$ are the parameters of our model. Notice that the term $\frac{1}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x)}$ normalizes the distribution, so that it sums to one.

# Softmax Regression Definition cont...

For convenience, we will also write $\theta$ to denote all the parameters of our model. When you implement softmax regression, it is usually convenient to represent $\theta$ as a $n$-by-$K$ matrix obtained by concatenating $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}$ into columns, so that

$$\theta = \begin{bmatrix} | & | & | & | \\ \theta^{(1)} & \theta^{(2)} & \cdots & \theta^{(K)} \\ | & | & | & | \end{bmatrix}.$$

# Softmax Regression Definition cont...

In the softmax regression setting, we are interested in multi-class classification (as opposed to only binary classification), and so the label y can take on K different values, rather than only two. Thus, in our training set $\{(x(1),y(1)),\ldots,(x(m),y(m))\}$, we now have that $y(i)\in\{1,2,\ldots,K\}$. (Note that our convention will be to index the classes starting from 1, rather than from 0.) For example, in the MNIST digit recognition task, we would have K=10 different classes.

If C=2, softmax reduces to logistic regression.

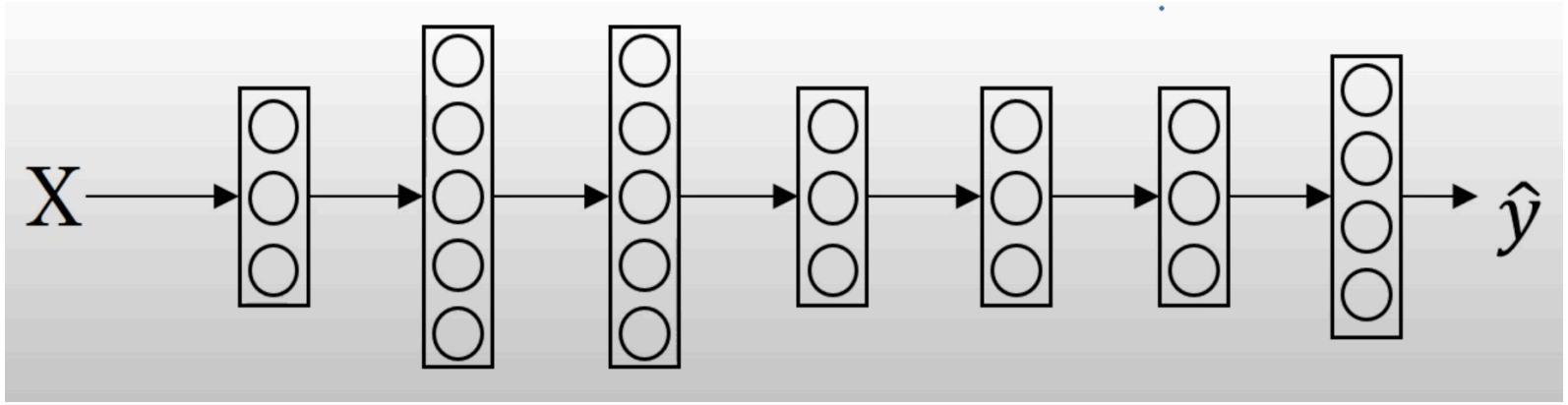
# Softmax Regression

## Recognizing cats, dogs, and baby chicks

3 1 2 0 3 2 0 1

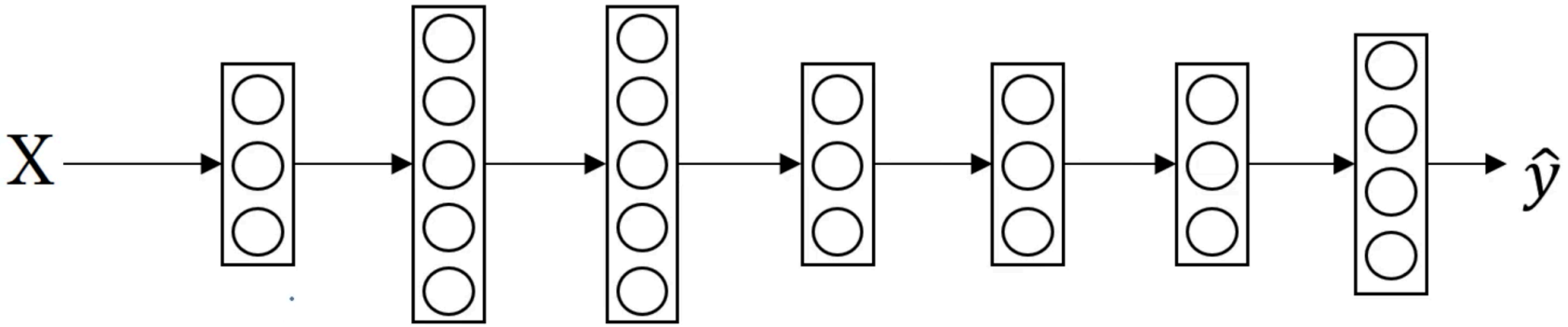
C = #classes = 4 (0, ..., 3)

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



# Softmax Regression cont...

Softmax layer




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

# Softmax Regression cont...

## Softmax layer

Activation Function:

$$t = e^{(Z^{[l]})}$$

$$a^{[L]} = \frac{e^{Z^{[L]}}}{\sum_{j=1}^{n} t_i}$$

$$a_i^{[L]} = \frac{t_i}{\sum_{j=1}^{n}}$$

# Softmax Regression cont...

## Softmax layer

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

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

$$\sum_{j=1}^{4} t_j = 176.3$$

$$a^{[L]} = \frac{t}{176.3} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}$$

$$a^{[L]} = g^{[L]}(Z^{[L]})$$

# Softmax Regression cont...

Softmax examples