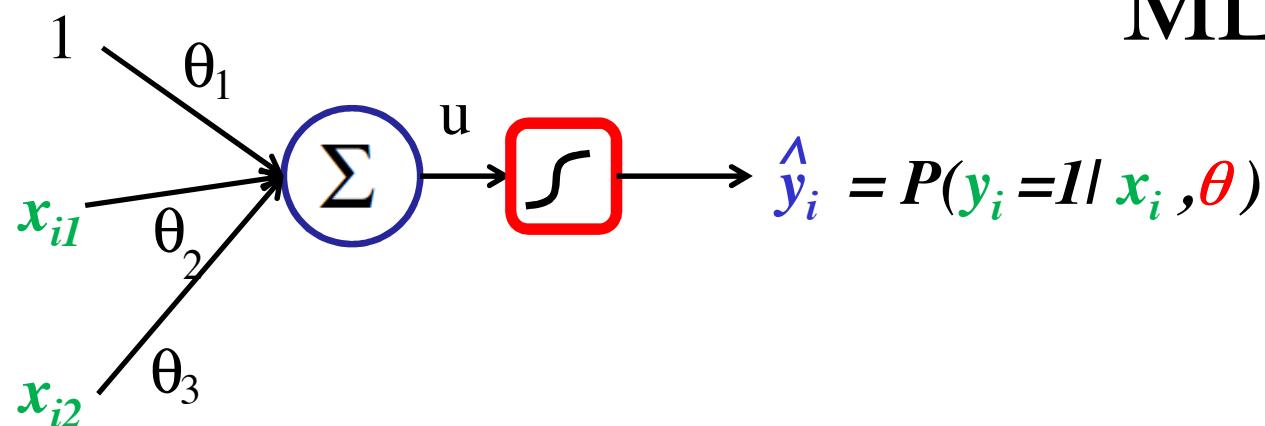


Outline of the lecture

This lecture introduces you to the fascinating subject of classification and regression with artificial neural networks. In particular, it

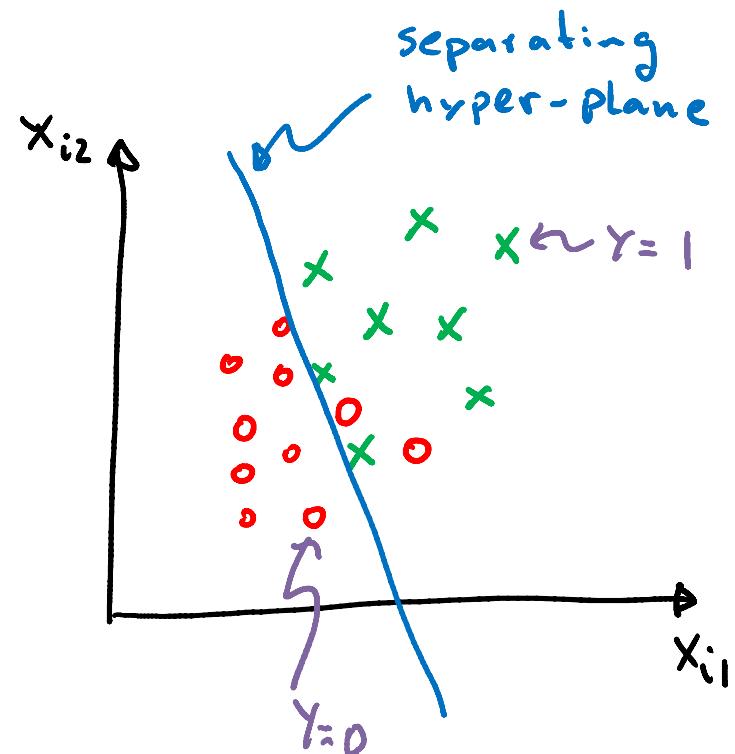
- Introduces multi-layer perceptrons (MLPs)
- Teaches you how to combine probability with neural networks so that the nets can be applied to regression, binary classification and multivariate classification.
- Describes the relation between energy functions (cost/loss functions) and probabilistic models.

MLP – 1 neuron

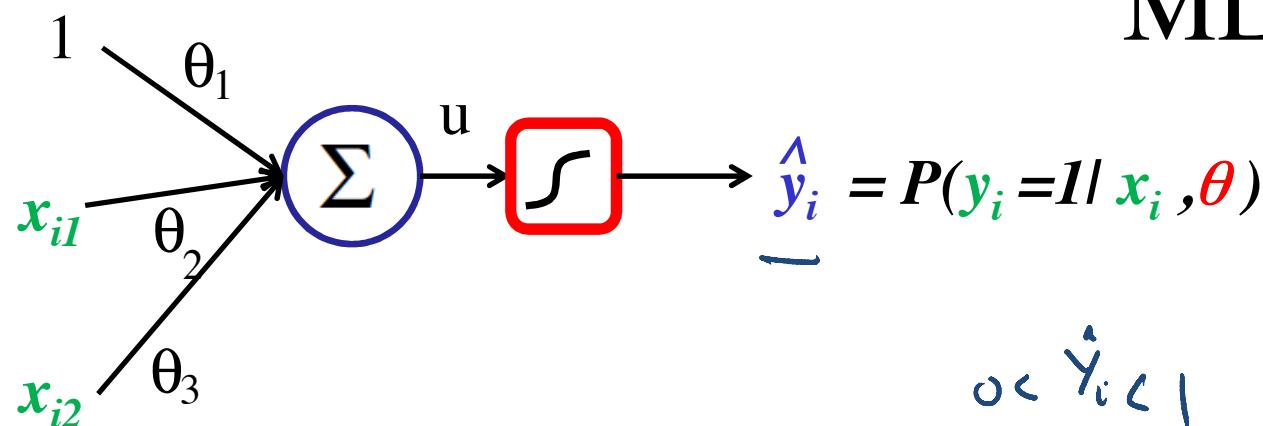


We are given the data $\{x_i, y_i\}_{i=1}^n$
e.g.

	x_{i1}	x_{i2}	y_i
$i=1$	0.2	6	0
$i=2$	0.3	22	1
$i=3$	0.6	-0.6	1
$i=4$	-0.4	58	0
:			



MLP – 1 neuron



$$0 < \hat{y}_i < 1$$

$$u = \Theta_0 + \Theta_1 x_{i1} + \Theta_2 x_{i2}$$

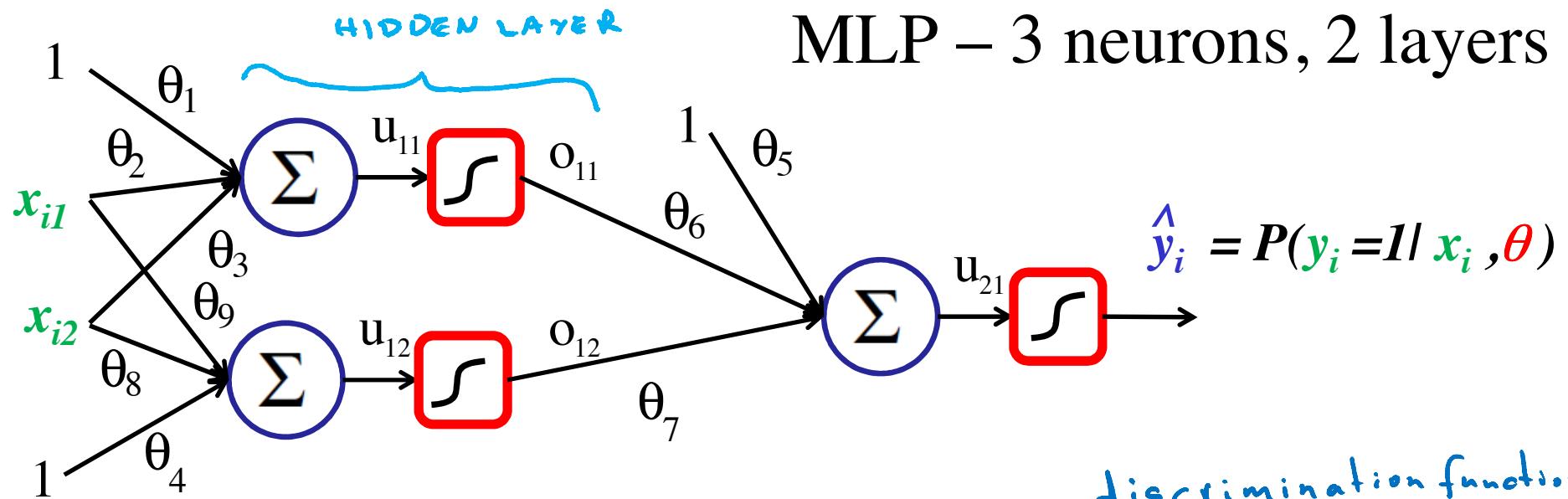
$$\hat{y}_i = \frac{1}{1+e^{-u}} = \frac{1}{1+e^{-\Theta_0 - \Theta_1 x_{i1} - \Theta_2 x_{i2}}} = P(y_i=1 | x_i, \theta)$$

$$P(y_i=1 | x_i, \theta) = \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} = \begin{cases} \hat{y}_i & \text{When } y_i = 1 \\ 1 - \hat{y}_i & \text{Otherwise} \end{cases}$$

For n independent observations (Bernoulli)

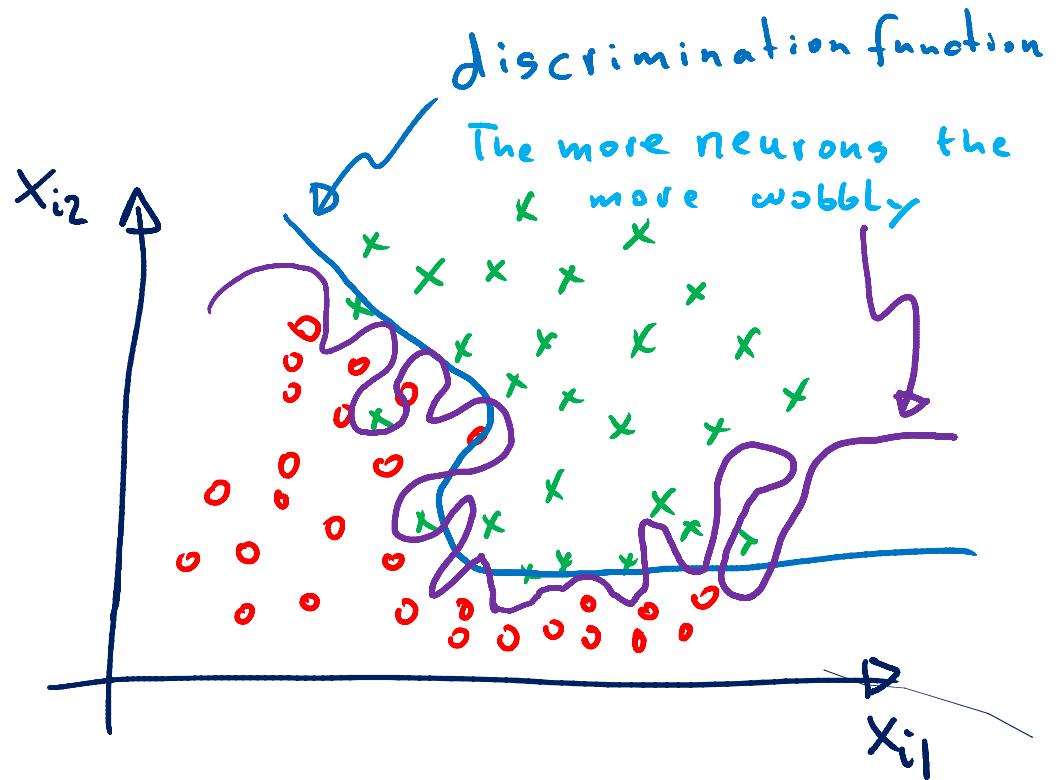
$$P(Y | X, \theta) = \prod_{i=1}^n P(y_i | x_i, \theta)$$

MLP – 3 neurons, 2 layers

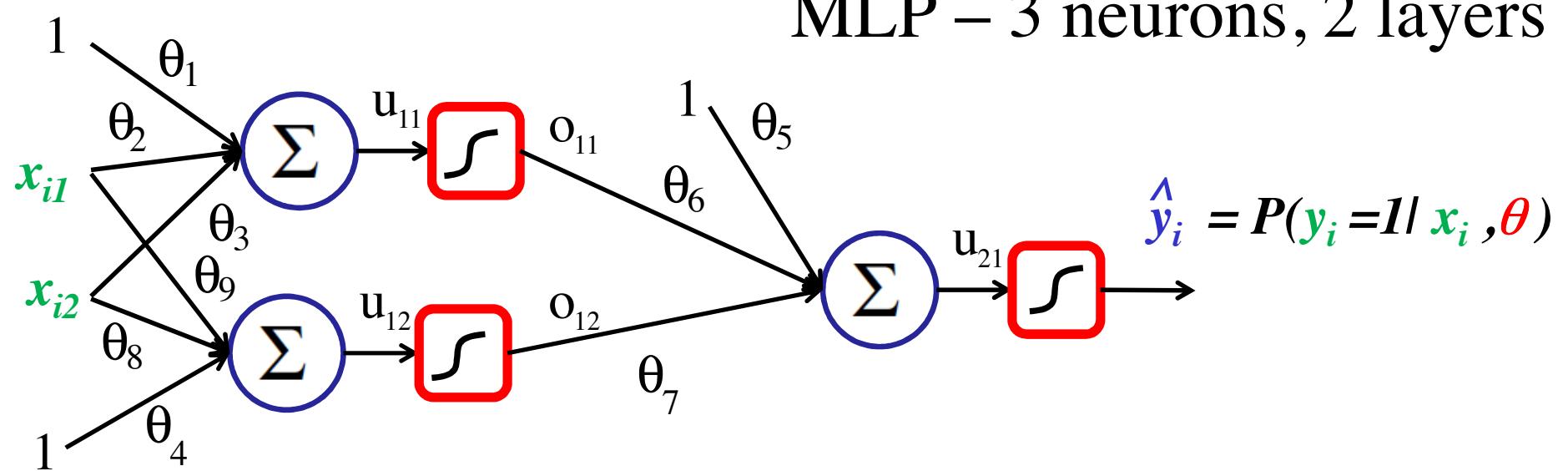


Data:

	x_{i1}	x_{i2}	y_i
i=1	6	2	1
i=2	0.2	-5	0
	-100	3.1	1
	6	9	0
	5	8	0



MLP – 3 neurons, 2 layers



$$u_{11} = \theta_1 + \theta_2 x_{i1} + \theta_3 x_{i2}$$

$$u_{12} = \theta_4 + \theta_5 x_{i1} + \theta_6 x_{i2}$$

$$o_{11} = \frac{1}{1+e^{-u_{11}}}$$

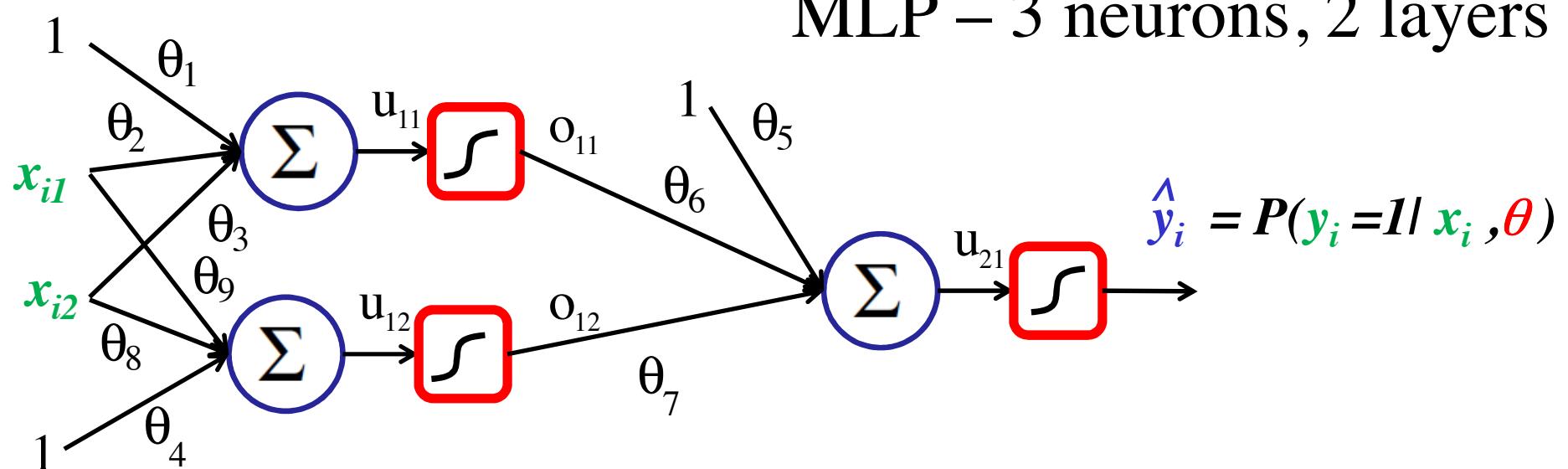
$$o_{12} = \frac{1}{1+e^{-u_{12}}}$$

$$\hat{y}_i = \frac{1}{1+e^{-u_{21}}}$$

$$u_{21} = \theta_5 + \theta_6 o_{11} + \theta_7 o_{12}$$

$$P(y_i | x_i, \theta) = \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i}$$

MLP – 3 neurons, 2 layers



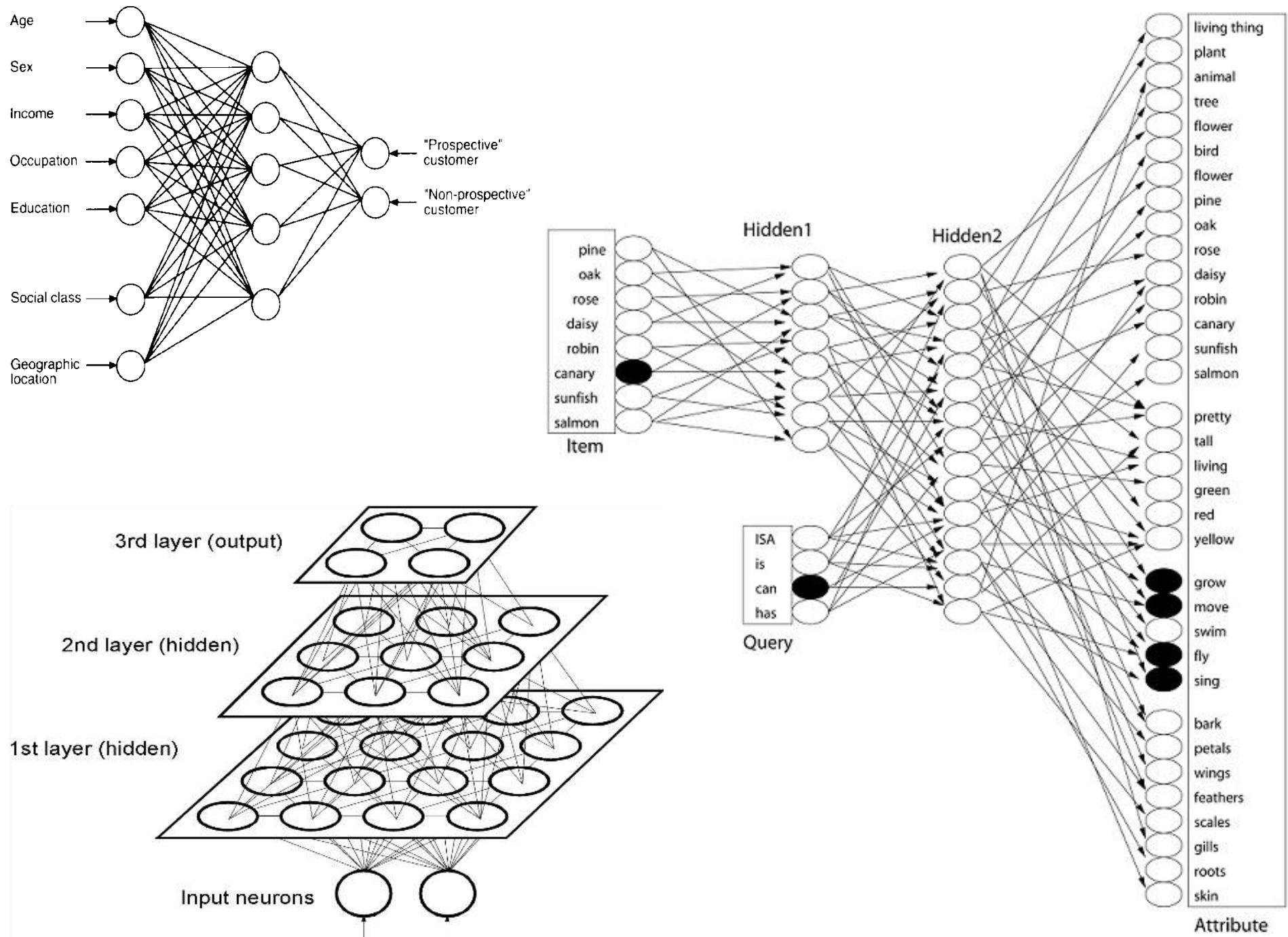
For n independent observations.

$$P(Y|X, \Theta) = \prod_{i=1}^n \hat{y}_i^{y_i} (1-\hat{y}_i)^{1-y_i} = \prod_{i=1}^n P(y_i | x_i, \Theta)$$

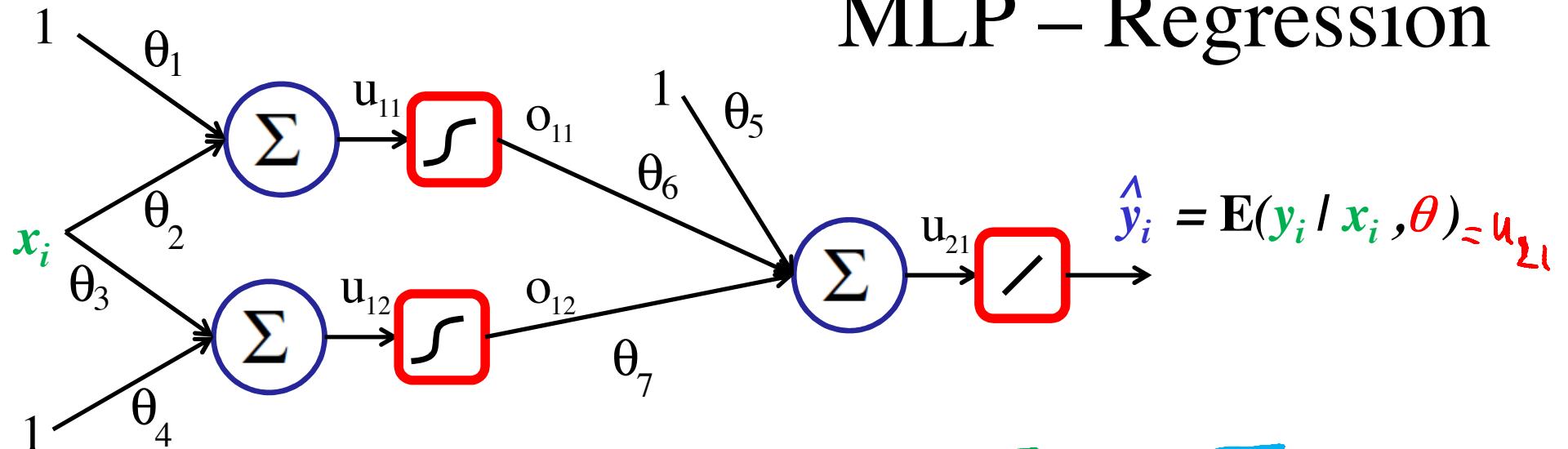
Cost:

$$C(\Theta) = -\log P(Y|X, \Theta) = -\sum_{i=1}^n y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i)$$

i.e minimize the cross-entropy error.

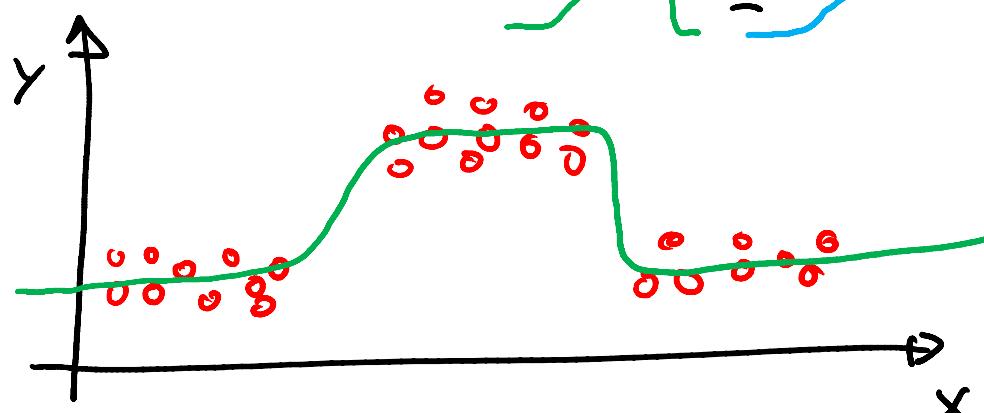


MLP – Regression



Data :

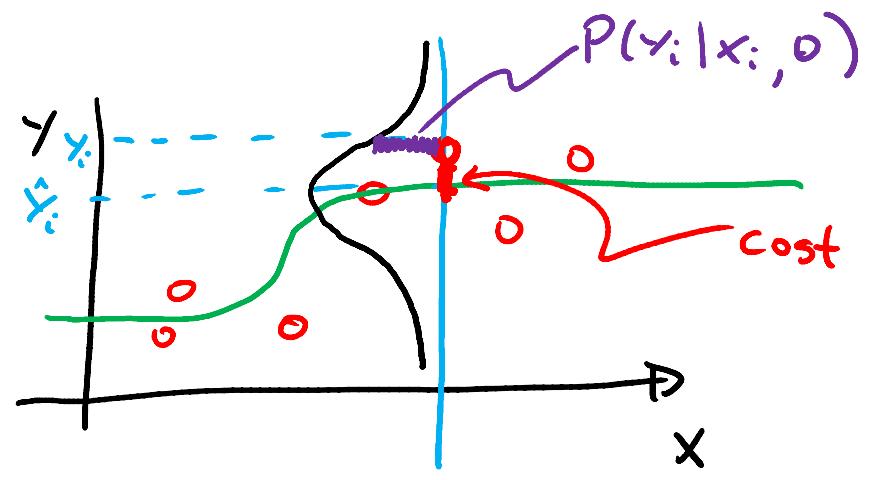
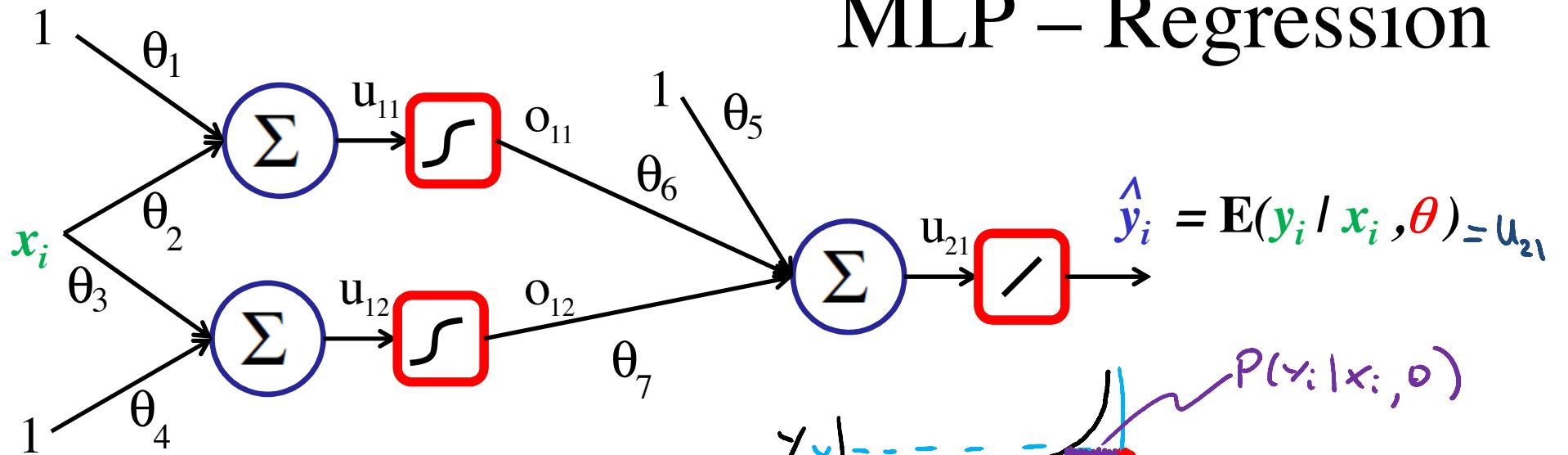
x_i	y_i
0.2	0.6
0.9	0.4
-0.6	5
-0.3	6.2



$$\hat{y}_i = \theta_5 + \frac{\theta_6}{1 + e^{-\theta_1 - \theta_2 x_i}} + \frac{\theta_7}{1 + e^{-\theta_4 - \theta_3 x_i}}$$



MLP – Regression

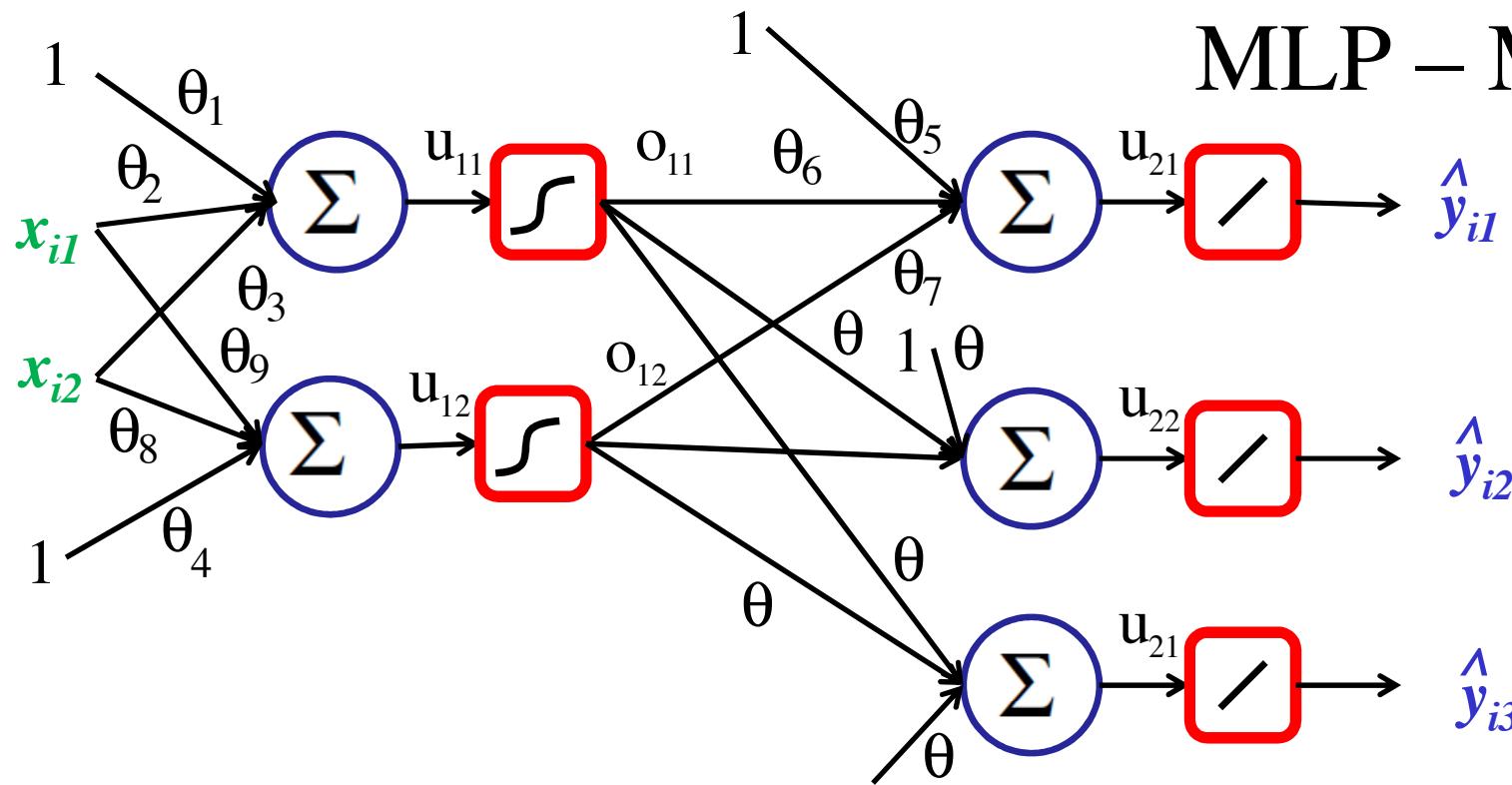


$$P(y_i | x_i, \theta) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2} (\hat{y}_i - y_i)^2}$$

$$C(\theta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = f(\theta, x_i)$$

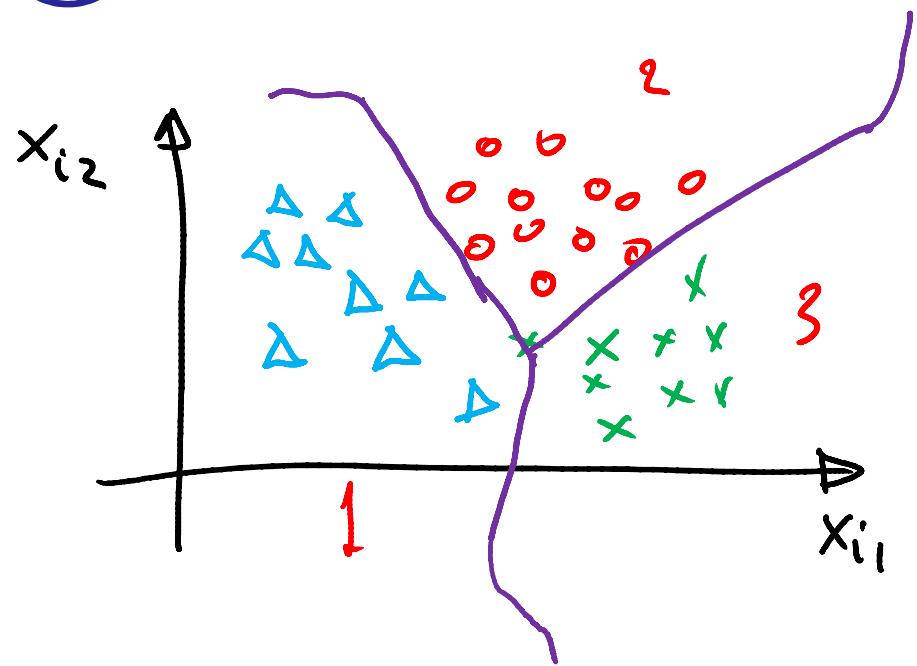
MLP – Multiclass

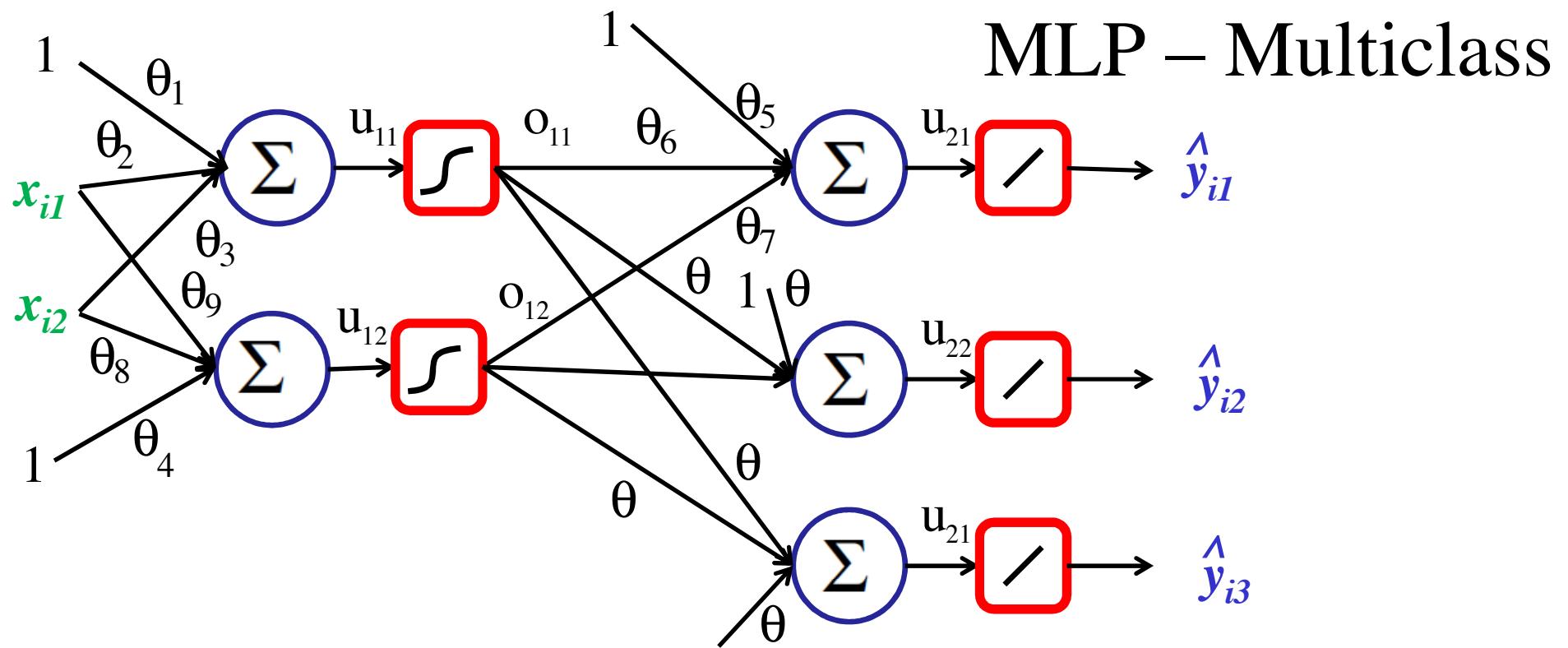


Data :

x_{i1}	x_{i2}	y_{i1}	y_{i2}	y_{i3}
0.2	0.3	0	1	0
-5	-6	1	0	0
-20	4	1	0	0
42	6.8	0	0	1

class 2
 class 1
 // 1
 class 3





To get a probabilistic model, define: SOFTMAX

$$P(y_i = (0|1|0) | x_i, \theta) = P(y_i = 2 | x_i, \theta) = \frac{e^{\hat{y}_2}}{e^{\hat{y}_1} + e^{\hat{y}_2} + e^{\hat{y}_3}}$$

$$\mathbb{I}_2(y_i) = \begin{cases} 1 & y_i = 2 \\ 0 & \text{o.w.} \end{cases}$$

MLP – Multiclass

Then,

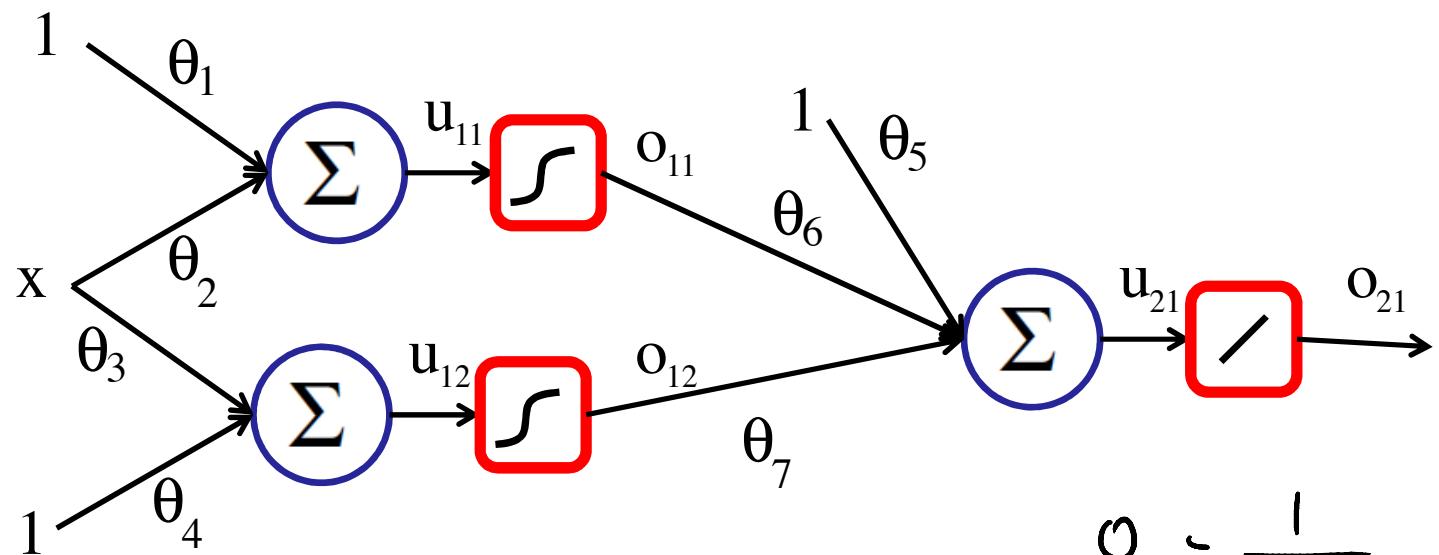
$$P(y_i | x_i, \theta) = \left[\frac{e^{\hat{y}_{i1}}}{e^{\hat{y}_{i1}} + e^{\hat{y}_{i2}} + e^{\hat{y}_{i3}}} \right] \mathbb{I}_1(y_i) \left[\frac{e^{\hat{y}_{i2}}}{e^{\hat{y}_{i1}} + e^{\hat{y}_{i2}} + e^{\hat{y}_{i3}}} \right]^2 \mathbb{I}_2(y_i) \left[\frac{e^{\hat{y}_{i3}}}{e^{\hat{y}_{i1}} + e^{\hat{y}_{i2}} + e^{\hat{y}_{i3}}} \right] \mathbb{I}_3(y_i)$$

$$= \begin{cases} e^{\hat{y}_{i1}}/\text{sum} & y_i = 1 \\ e^{\hat{y}_{i2}}/\text{sum} & y_i = 2 \\ e^{\hat{y}_{i3}}/\text{sum} & y_i = 3 \end{cases}$$

Cost:

$$C(\theta) = -\log P(y_i | x_i, \theta) = -\sum_{i=1}^n \sum_{j=1}^3 \mathbb{I}_j(y_i) \log \frac{e^{\hat{y}_{ij}}}{\text{sum}}$$

Backpropagation



$$\hat{y} = o_{21} = u_{21} = \theta_5 + \theta_6 o_{11} + \theta_7 o_{12}$$

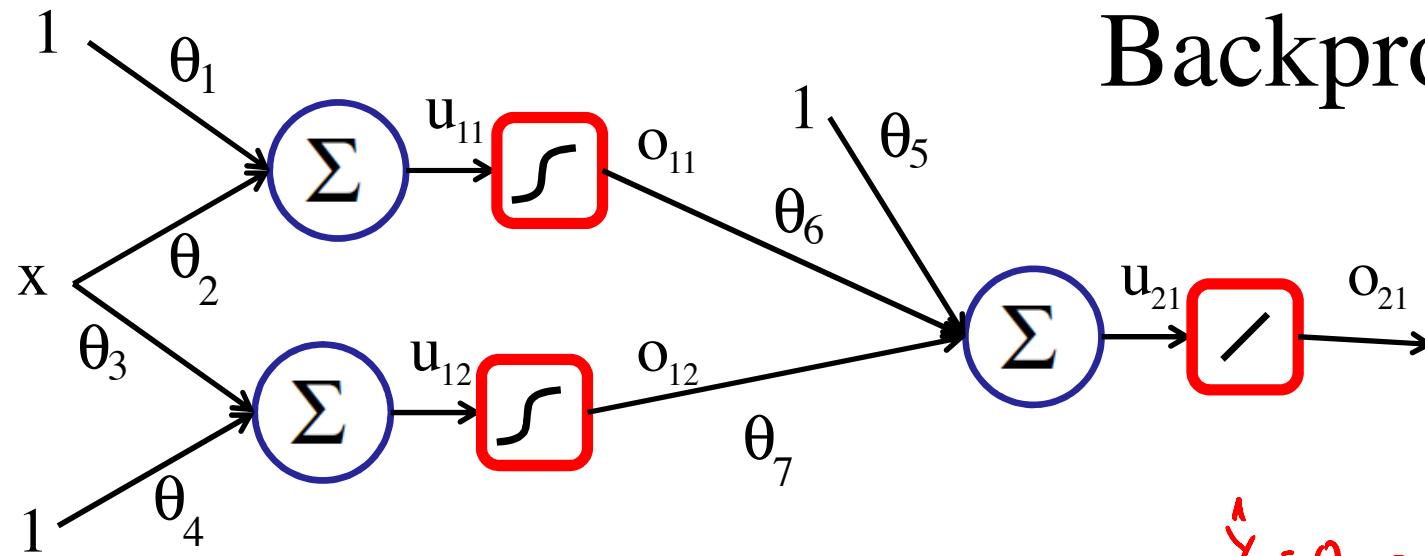
$$o_{11} = \frac{1}{1 + e^{-u_{11}}}$$

$$u_{11} = \theta_1 + \theta_2 x$$

$$o_{12} = \frac{1}{1 + e^{-u_{12}}}$$

$$u_{12} = \theta_4 + \theta_3 x$$

Backpropagation



$$E(\theta) = (y_i - \hat{y}_i(x_i, \theta))^2$$

$$\hat{Y} = o_{21} = \theta_5 + \theta_6 o_{11} + \theta_7 o_{12}$$

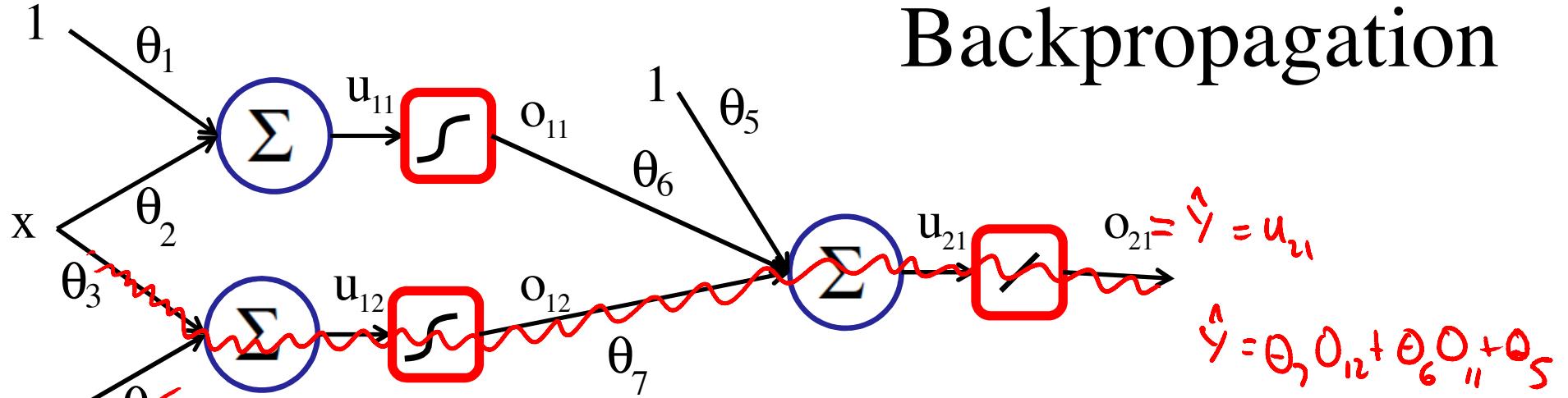
$$\frac{\partial E(\theta)}{\partial \theta_j} = -2 (y_i - \hat{y}_i(x_i, \theta)) \frac{\partial \hat{y}_i(x_i, \theta)}{\partial \theta_j}$$

$$\frac{\partial \hat{y}_i}{\partial \theta_5} = 1$$

$$\frac{\partial \hat{y}_i}{\partial \theta_6} = o_{11}$$

$$\frac{\partial \hat{y}_i}{\partial \theta_7} = o_{12}$$

Backpropagation



$$\frac{\partial \hat{y}}{\partial \theta_3} = \frac{\partial \hat{y}}{\partial u_{12}} \frac{\partial u_{12}}{\partial u_{11}} \frac{\partial u_{11}}{\partial \theta_3}$$

$$= \theta_1 O_{12} [1 - O_{12}] x$$

O_{12}

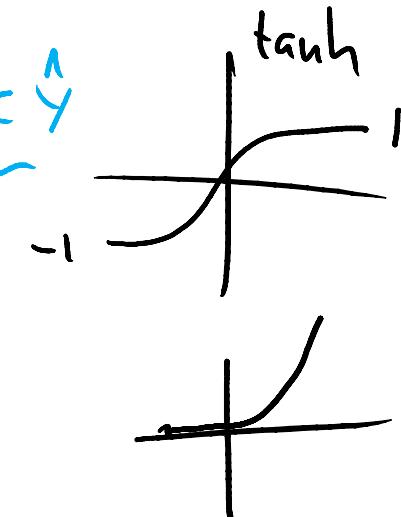
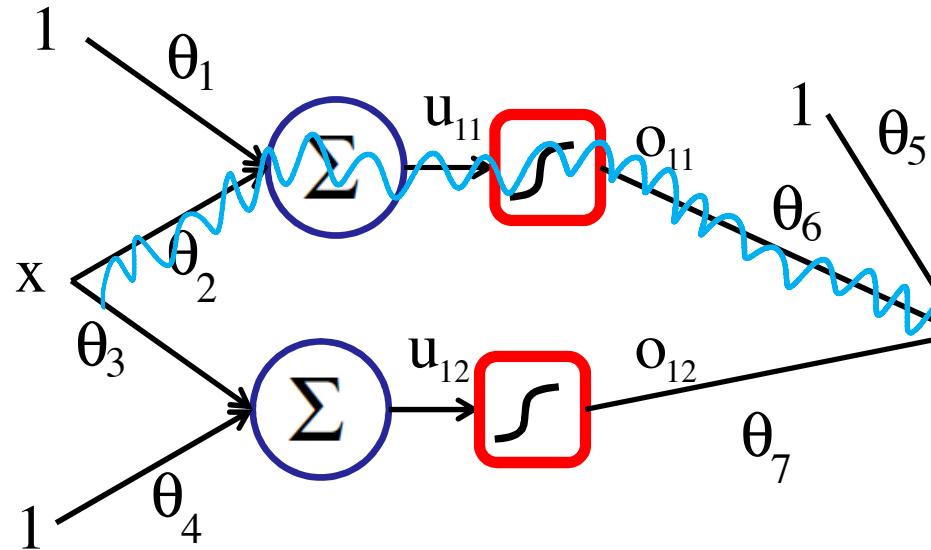
Thm

$$\frac{\partial O_{12}}{\partial u_{12}} = O_{12}(1 - O_{12})$$

i.e.

$$\frac{\partial}{\partial x} \frac{1}{1+e^{-x}} = \left(\frac{1}{1+e^{-x}} \right) \left(1 - \frac{1}{1+e^{-x}} \right)$$

Backpropagation



Batch:

$$\Theta_j^{(k+1)} = \Theta_j^{(k)} + \frac{\eta}{n} \sum_{i=1}^n (\gamma_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial \Theta_j}$$

iteration

Online:

$$\Theta_2^i = \Theta_2^{i-1} + \eta \left[(\gamma_i - \hat{y}_i) \Theta_2^{i-1} (o_2^i (1 - o_2^i)) x_i \right]$$