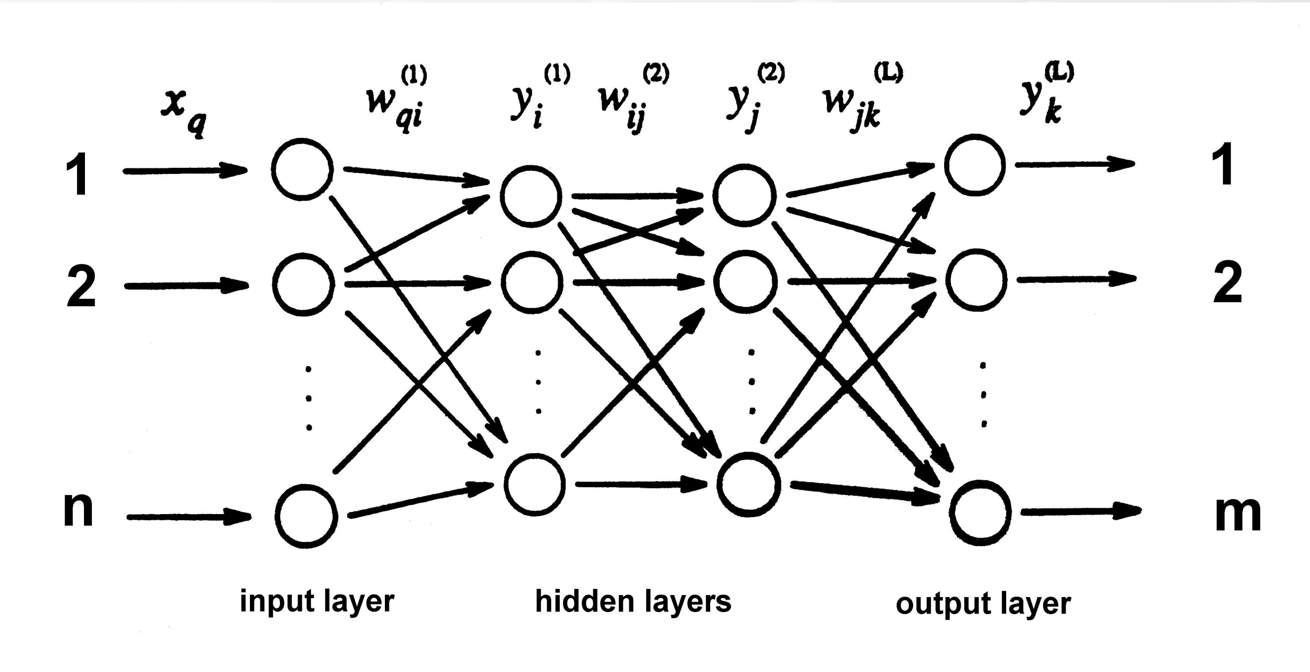
Backpropagation Learning Algorithm

- An algorithm for learning the weights in the network, given a training set of input-output pairs $\{ x^{\mu}, o^{\mu} \}$
- The algorithm is based on gradient descent method.
- Architecture



$w_{ij}^{(l)}$: Weight on connection between the i^{th} unit in layer (l-1) to j^{th} unit in layer l

Supervised Learning

Supervised learning algorithms require the presence of a "teacher" who provides the right answers to the input questions.

Technically, this means that we need a *training set* of the form

$$L = \left\{ \left(\overline{x}^1, \overline{y}^1 \right), \; \ldots\ldots \; \left(\overline{x}^p, \overline{y}^p \right) \right\}$$

where :

- $\overline{x}^h \quad (h = 1 \ldots p)$ is the network input vector
- $\overline{y}^h \quad (h = 1 \ldots p)$ is the network output vector

Supervised Learning

The learning (or training) phase consists of determining a configuration of weights in such a way that the network output be as close as possible to the desired output, for all examples in the training set.

Formally, this amounts to minimizing the following *error function* :

$$E = \frac{1}{2} \sum_{h=1}^{p} \left\| \overline{out}_h - \overline{y}_h \right\|_2^2$$

$$= \frac{1}{2} \sum_{h} \sum_{k} \left(out_k^h - y_k^h \right)^2$$

where $\overline{out}_h$ is the output provided by the network when given $\overline{x}^h$ as input.

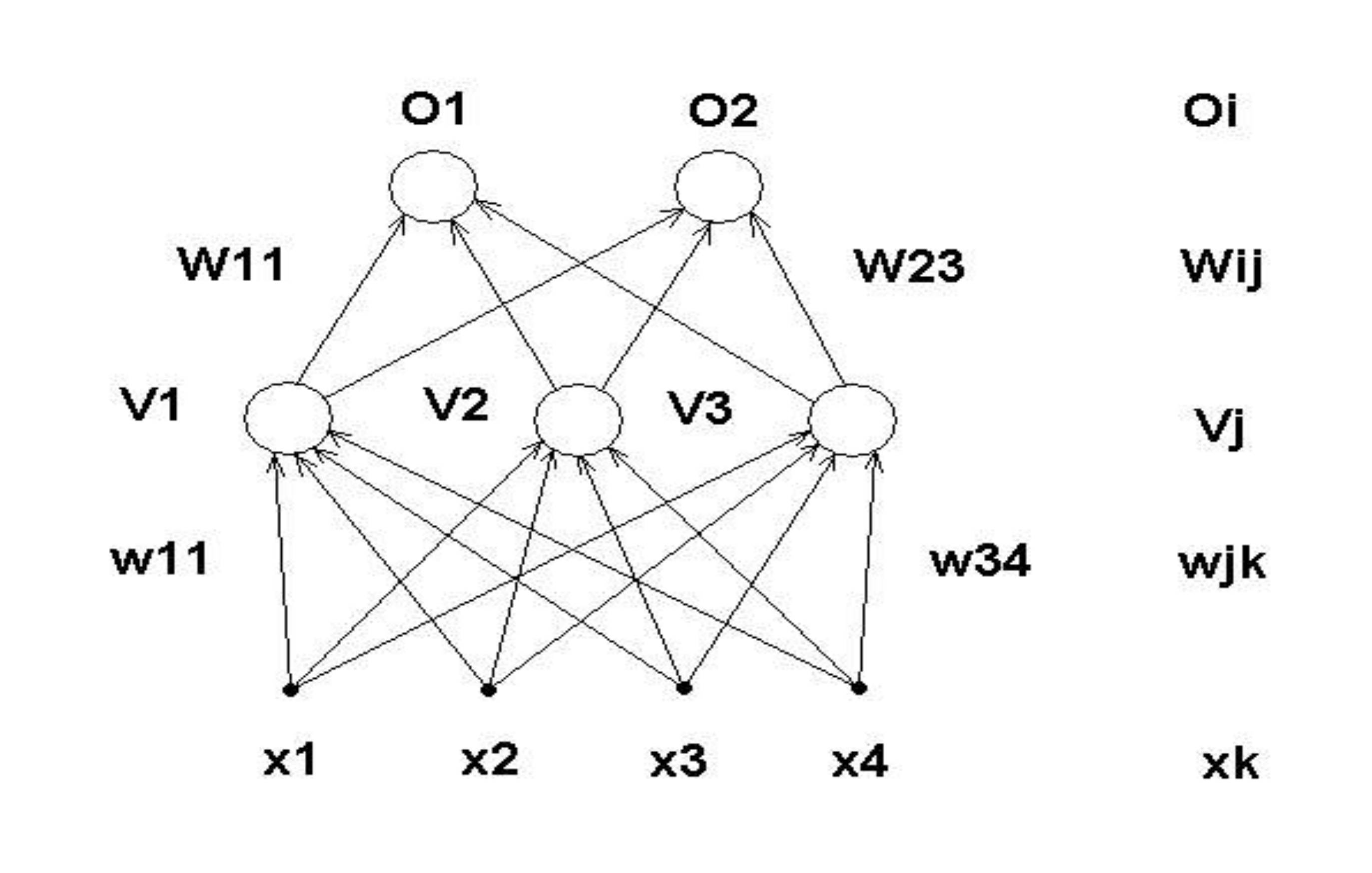
Back-Propagation

To minimize the error function *E* we can use the classic gradient – descent algorithm.

To compute the partial derivates $\partial E / \partial w_{ij}$, we use the *error back propagation* algorithm.

It consists of two stages:

- *Forward pass* : the input to the network is propagated layer after layer in forward direction
- *Backward pass* : the " error " made by the network is propagated backward, and weights are updated properly



Dato il pattern *μ*, l'unità nascosta *j* riceve un input netto dato da

$$h_j^\mu = \sum_k w_{jk} x_k^\mu$$

e produce come output :

$$V_j^\mu = g\left(h_j^\mu\right) = g\left(\sum_k w_{jk} x_k^\mu\right)$$

Back-Prop : Updating Hidden-to-Output Weights

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}}$$

$$= -\eta \frac{\partial}{\partial W_{ij}} \left[\frac{1}{2} \sum_{\mu} \sum_{k} \left(y_k^{\mu} - O_k^{\mu} \right)^2 \right]$$

$$= \eta \sum_{\mu} \sum_{k} \left(y_k^{\mu} - O_k^{\mu} \right) \frac{\partial O_k^{\mu}}{\partial W_{ij}}$$

$$= \eta \sum_{\mu} \left(y_i^{\mu} - O_i^{\mu} \right) \frac{\partial O_i^{\mu}}{\partial W_{ij}}$$

$$= \eta \sum_{\mu} \left(y_i^{\mu} - O_i^{\mu} \right) g'\left(h_i^{\mu} \right) V_j^{\mu}$$

$$= \eta \sum_{\mu} \delta_i^{\mu} V_j^{\mu}$$

where: $\delta_i^{\mu} = \left(y_i^{\mu} - O_i^{\mu} \right) g'\left(h_i^{\mu} \right)$

Back-Prop : Updating Input-to-Hidden Weights （I）

$$\Delta w_{jk} = -\eta \frac{\partial E}{\partial w_{jk}}$$

$$= \eta \sum_{\mu} \sum_{i} \left(y_i^{\mu} - O_i^{\mu} \right) \frac{\partial O_i^{\mu}}{\partial w_{jk}}$$

$$= \eta \sum_{\mu} \sum_{i} \left(y_i^{\mu} - O_i^{\mu} \right) g'\left(h_i^{\mu} \right) \frac{\partial h_i^{\mu}}{\partial w_{jk}}$$

$$\frac{\partial h_i^{\mu}}{\partial w_{jk}} = \sum_{l} W_{il} \frac{\partial V_l^{\mu}}{\partial w_{jk}}$$

$$= W_{ij} \frac{\partial V_j^{\mu}}{\partial w_{jk}}$$

$$= W_{ij} \frac{\partial g\left(h_j^{\mu} \right)}{\partial w_{jk}}$$

$$= W_{ij} \, g'\left(h_j^{\mu} \right) \frac{\partial h_j^{\mu}}{\partial w_{jk}}$$

Back-Prop : Updating Input-to-Hidden Weights （II）

$$\frac{\partial h_j^\mu}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \sum_m w_{jm} x_m^\mu$$

$$= x_k^\mu$$

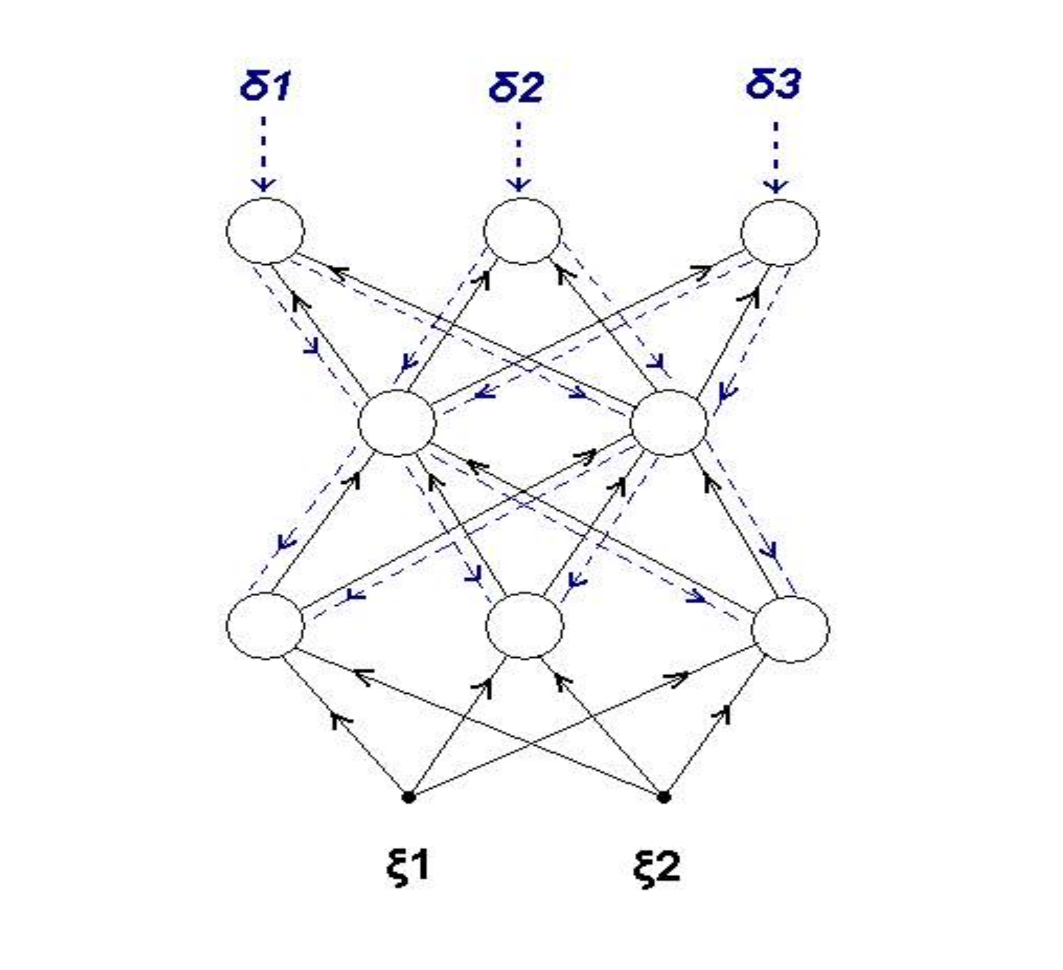
Hence, we get :

$$\Delta w_{jk} = \eta \sum_{\mu,i} \left(y_i^\mu - O_i^\mu\right) g'\left(h_i^\mu\right) W_{ij}\ g'\left(h_j^\mu\right) x_k^\mu$$

$$= \eta \sum_{\mu,i} \delta_i^\mu\ W_{ij}\ g'\left(h_j^\mu\right) x_k^\mu$$

$$= \eta \sum_{\mu} \hat{\delta}_j^\mu\ x_k^\mu$$

where : $\hat{\delta}_j^\mu = g'\left(h_j^\mu\right) \sum_i \delta_i^\mu\ W_{ij}$



Retropropagazione dell’errore :

- le linee nere indicano il segnale propagato in avanti
- Le linee blu indicano l’errore (i *δ*) propagato all’indietro