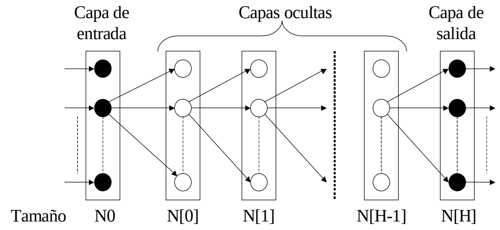
Perceptrón Multicapa (MLP)

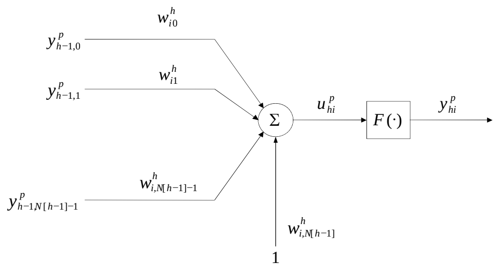
Teoría de Autómatas y Lenguajes Formales II. Curso 2011-12.
3º curso de Ingeniería Técnica en Informática de Sistemas
Escuela Técnica Superior de Ingeniería Informática
Universidad de Valladolid

Arquitectura y funcionamiento

- Organización por capas:
  - No hay conexión entre los diferentes integrantes de una capa, ni tan siquiera de uno consigo mismo
  - Las salidas de las neuronas sólo sirven de entrada a las de las capa(s) siguiente(s)



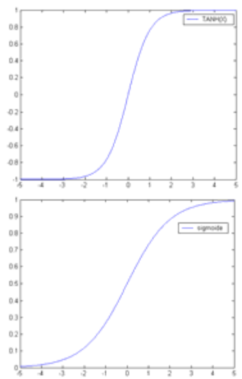
Funcionamiento: recuperación



$$y_{hi}^p = F(u_{hi}^p); \quad \text{donde} \quad u_{hi}^p = \left(\sum_{j=0}^{N[h-1]-1} w_{ij}^h y_{h-1,j}^p \right) + w_{i,N[h-1]}^h \quad \forall h = 1,2,\ldots H$$

$$y_{0i}^p = F(u_{0i}^p); \quad \text{donde} \quad u_{0i}^p = \left(\sum_{j=0}^{N0-1} w_{ij}^0 I_j^p \right) + w_{i,N0}^0$$

Funciones de activación (I)



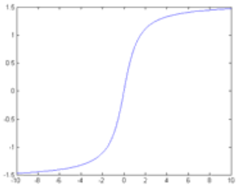
$$\tanh(x) = \frac{1-e^{-x}}{1+e^{-x}} = 2F(x) - 1$$

$$F(x) = \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx}F(x) = F(x)(1-F(x))$$

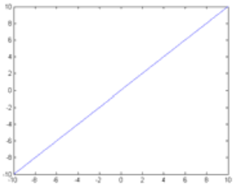
$$\frac{d}{dx}\tanh(x) = 2\frac{d}{dx}F(x)$$

Funciones de activación (II)



$$F(x) = arctg(x)$$

$$\frac{d}{dx}F(x) = \frac{1}{1+x^2}$$



$$F(x) = x$$

Sólo en la capa de salida
No se aplica el término momento

Aprendizaje (I)

- Error cuadrático medio de la muestra p-ésima:

$$E^p = \frac{1}{2}\sum_{i=0}^{N[H]-1}\left(d_i^p - y_{Hi}^p\right)^2$$

- Variación del peso j, de la neurona i, dentro de la capa h, provocado por la muestra p:

$$\Delta_p w_{ij}^h = -\gamma \frac{\partial E^p}{\partial w_{ij}^h}$$; que al aplicar la regla de la cadena:

$$\Delta_p w_{ij}^h = -\gamma \frac{\partial E^p}{\partial y_{hi}^p}\frac{\partial y_{hi}^p}{\partial w_{ij}^h} = -\gamma \frac{\partial E^p}{\partial y_{hi}^p}F'(u_{hi}^p)y_{h-1j}^p \quad \forall h = 1,2,\ldots H$$

$$\Delta_p w_{ij}^0 = -\gamma \frac{\partial E^p}{\partial y_{0i}^p}\frac{\partial y_{0i}^p}{\partial w_{ij}^0} = -\gamma \frac{\partial E^p}{\partial y_{0i}^p}F'(u_{0i}^p)I_j^p$$

Aprendizaje (II)

- Teniendo en cuenta la definición de E^p, para la capa de salida, se tiene que:

$$-\frac{\partial E^p}{\partial y_{Hi}^p} = d_i^p - y_{Hi}^p \quad \Rightarrow \quad \Delta_p w_{ij}^H = \gamma(d_i^p - y_{Hi}^p)F'(u_{Hi}^p)y_{H-1j}^p$$

- Dejando a un lado los índices *H y p*, el último término depende de la conexión (j) y, el resto, de la neurona en sí (i) agrupado bajo una variable "delta".

$$\Delta_p w_{ij}^H = \gamma \delta_{Hi}^p y_{H-1j}^p$$; donde: $$\delta_{Hi}^p = (d_i^p - y_{Hi}^p)F'(u_{Hi}^p)$$

Aprendizaje (III)

- Para la última capa oculta (h=H-1):

$$-\frac{\partial E^p}{\partial y_{H-1i}^p} = \sum_{k=0}^{N[H]-1}(d_k^p - y_{Hk}^p)\frac{\partial y_{Hk}^p}{\partial y_{H-1i}^p} = \sum_{k=0}^{N[H]-1}(d_k^p - y_{Hk}^p)F'(u_{Hk}^p)\frac{\partial u_{Hk}^p}{\partial y_{H-1i}^p}$$

- De acuerdo con la definición de los términos "delta" para la capa de salida:

$$-\frac{\partial E^p}{\partial y_{H-1i}^p} = \sum_{k=0}^{N[H]-1}\delta_{Hk}^p w_{ki}^H$$

Aprendizaje (IV)

- Los dos resultados anteriores llevados a la actualización de los pesos:

$$\Delta_p w_{ij}^{H-1} = \gamma \sum_{k=0}^{N[H]-1} \delta_{Hk}^p w_{ki}^H F'(u_{H-1i}^p) y_{H-2j}^p$$

- Al separar las dependencias funcionales de la neurona (i) y de la conexión (j), se tendría que:

$$\Delta_p w_{ij}^{H-1} = \gamma \delta_{H-1i}^p y_{H-2j}^p$$

- Donde se obtienen otros términos “delta” característicos de cada neurona: (y de la capa)

$$\delta_{H-1i}^p = F'(u_{H-1i}^p) \sum_{k=0}^{N[H]-1} \delta_{Hk}^p w_{ki}^H$$

Aprendizaje (IV)

- Flujo de información en dos fases:
  - Hacia adelante:
    - Respuesta de la red para una muestra:
      - Se propaga la salida de una capa a la siguiente
  - Hacia atrás:
    - Cálculo del error cuadrático medio:
      - Obtención de los términos delta
      - Actualización de los pesos de la capa de salida
      - Propagación hacia atrás de los términos delta
      - Actualización de los pesos de la capa oculta

Aprendizaje (V)

- Fase hacia delante:

$$y_{0i}^p = F(u_{0i}^p); \quad \text{donde} \quad u_{0i}^p = \left(\sum_{j=0}^{N0-1} w_{ij}^0 I_j^p \right) + w_{iN0}^0$$

$$y_{hi}^p = F(u_{hi}^p); \quad \text{donde} \quad u_{hi}^p = \left(\sum_{j=0}^{N[h-1]-1} w_{ij}^h y_{h-1j}^p \right) + w_{iN[h-1]}^h \quad \forall h = 1,2,\ldots H$$

- Fase hacia atrás:

Términos delta:

$$\delta_{Hi}^p = (d_i^p - y_{Hi}^p) F'(u_{Hi}^p)$$

$$\delta_{hi}^p = F'(u_{hi}^p) \sum_{k=0}^{N[h+1]-1} \delta_{h+1k}^p w_{ki}^{h+1} \quad \forall h = 0,1,\ldots H-1$$

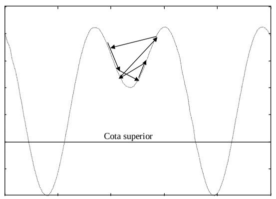
Actualización de los pesos:

$$\Delta_p w_{ij}^h = \gamma \delta_{hi}^p y_{h-1j}^p \quad \forall h = 1,2,\ldots H$$

$$\Delta_p w_{ij}^0 = \gamma \delta_{0i}^p I_j^p$$

Aprendizaje (VI)

- Imposibilidad de salir de mínimos locales:



- Adición de término de inercia o momento:

$$\Delta_p w_{ij}^h = \gamma \delta_{hi}^p y_{h-1j}^p + \alpha \Delta_{p-1} w_{ij}^h \quad \forall h = 1,2,\ldots H$$

$$\Delta_p w_{ij}^0 = \gamma \delta_{0i}^p I_j^p + \alpha \Delta_{p-1} w_{ij}^0$$

α es el coeficiente del momento y γ el de aprendizaje

Aprendizaje (VII)

- Dos conjuntos de muestras:
  - Aprendizaje (80%) y verificación (20%)
  - Conjuntos disjuntos
  - Ambos representativos del conjunto total

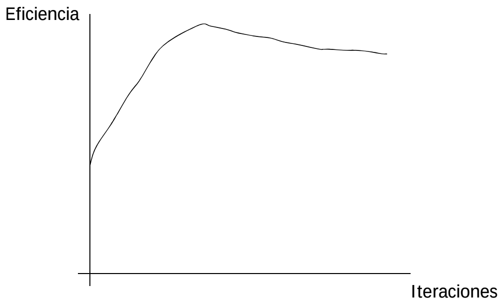
Para cada muestra de aprendizaje (*)
    Salida de la capa oculta
    Salida de la capa de salida
    Términos delta de capa de salida
    Actualización de los pesos de salida
    Términos delta de la capa oculta
    Acualización de los pesos capa oculta
Para cada muestra de verificación
    Respuesta de red
Comprobación de la condición de final de aprendizaje
    Caso negativo vuelta al inicio (*)

Aprendizaje (VIII)

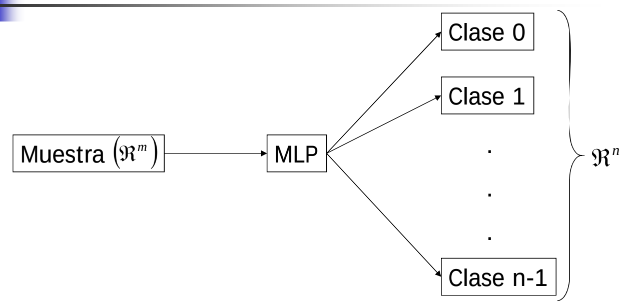
- Condición de final de aprendizaje:
  - Depende de cada caso particular:
    - Tasa de aciertos mínima
    - Cota superior para el error máximo por muestra
    - Cota superior para el error total
- En cualquier caso, el aprendizaje del MLP es siempre por épocas enteras.
    - Una época es una pasada por todas las muestras de un experimento

Aprendizaje (IX)

- Sobre-entrenamiento (overtraining):



Reconocimiento de patrones estáticos (I)



Salida deseada para la clase i:
(0, 0, ….., 0, 1, 0, …..0)
i

Reconocimiento de patrones estáticos (II)

- Problemas con la saturación de la derivada de la función sigmoide:
  - Cuando su valor es uno o cero, su derivada es cero, lo que implica la nula modificación de pesos, aun siendo opuestas la respuesta y la deseada.
  - Los ceros y unos de la salida deseada se sustituyen por valores próximos: 0.05 y 0.95

- El criterio para considerar acierto:
  - Coincidencia entre la neurona ganadora y la neurona asociada a la clase.