REDES NEURONALES 2024

Clase 17 parte 1

Jueves 17 de octubre 2024

FAMAF, Universidad Nacional de Córdoba

Instituto de Física Enrique Gaviola (UNC-CONICET)

M24/10/23 c18p2

El método del descenso por el gradiente

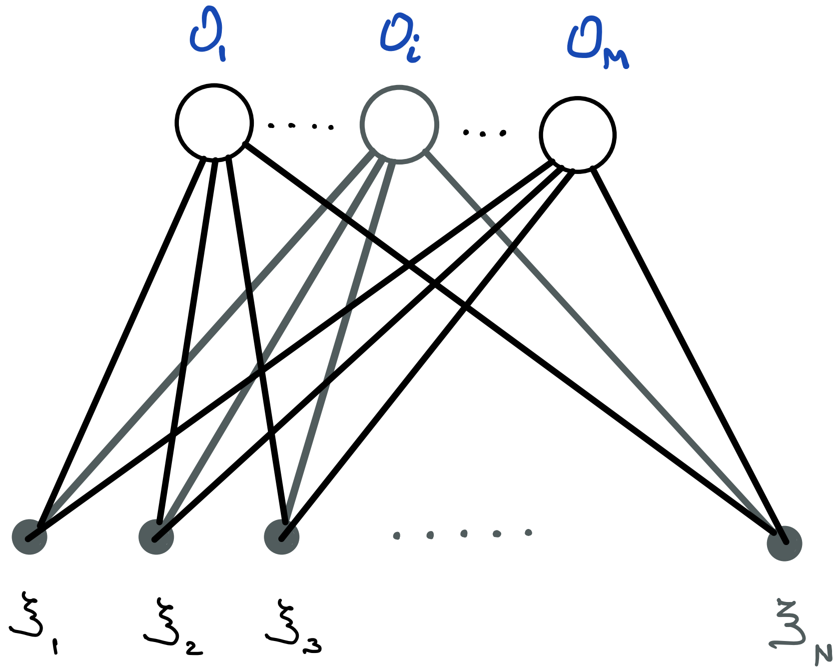
Hoy presentaremos por primera vez en este curso un elemento central en el aprendizaje supervisado moderno, el cual nos acompaña todos los días en nuestras investigaciones y nuestras aplicaciones de IA. Veremos en esta parte la estrecha relación existente entre aprender y optimizar una función, o sea, encontrar sus valores mínimos, locales y globales.

Definir como función error o función costo (loss function, en inglés).

$$E(\bar{w}) = \sum_{\mu=1}^{p} \left(\sum_{i=1}^{M} \frac{1}{2} \left(\zeta_i^{\mu} - O_i^{\mu} \right)^2 \right)$$

Suma sobre la p elementos del conjunto de entrenamiento

Suma sobre la M neuronas de la capa de salida



Cuando la presentamos la entrada μ obtenemos el resultado dado por la regla del perceptrón lineal:

$$O_i^\mu = g(h_i^\mu) = h_i^\mu = \sum_k^N w_{ik}\,\xi_k^\mu = \underbrace{\bar{W}_i \cdot \bar{\xi}^\mu}$$

La idea es comenzar con un vector inicial arbitrario W, como hacíamos en el caso de la neurona de salida binaria y presentarle un ejemplo, el μ del conjunto de entrenamiento:

$$E : \mathbb{R}^{N\times M} \longrightarrow \mathbb{R}$$

$$\overrightarrow{\nabla E} : \mathbb{R}^{N\times M} \longrightarrow \mathbb{R}^{N\times M} \qquad \nabla E_{ik} = \frac{\partial E}{\partial w_{ik}}$$

$$W_{ik}^{nuevo} = W_{ik}^{anterior} + \Delta W_{ik}$$

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

$$= -\eta \frac{\partial}{\partial W_{ik}} \frac{1}{2} \sum_{\mu=1}^{p} \left(\zeta_i^\mu - O_i^\mu \right)^2$$

$$= -\eta \sum_{\mu=1}^{p} \frac{1}{2} \frac{\partial}{\partial W_{ik}} \left(\zeta_i^\mu - \sum_j w_{ij} \xi_j^\mu \right)^2$$

$$= -\eta \sum_{\mu=1}^{p} \frac{2}{\cancel{2}} \left(\zeta_i^\mu - \sum_j w_{ij} \xi_j^\mu \right) \frac{\partial}{\partial W_{ik}} \left(- \sum_j w_{ij} \xi_j^\mu \right)$$

$$= \eta \sum_{\mu=1}^{p} \left(\zeta_i^\mu - O_i^\mu \right) \xi_k^\mu$$

$$\boxed{\Delta W_{ik} = \eta \sum_{\mu=1}^{p} \left(\zeta_i^\mu - O_i^\mu \right) \xi_k^\mu}$$

$$W_{ik}^{nuevo} = W_{ik}^{anterior} + \Delta W_{ik}$$

Si definimos:

$$d_i^\mu = \zeta_i^\mu - O_i^\mu$$

entonces:

$$\boxed{W_{ik}^{nuevo} = W_{ik}^{anterior} + \eta \sum_{\mu=1}^{p} d_i^\mu \xi_k^\mu}$$

Esta regla nos permite buscar un mínimo local de una función de muchas variables. En nuestro caso buscamos el mínimo de la función *Error Cuadrático Medio* sobre todos los posibles valores de las sinapsis y umbrales. En otras palabras, se trata de encontrar el juego de sinapsis y umbrales que produce el menor de aprendizaje sobre el conjunto de entrenamiento.

Esta regla fue deducida y bautizada en forma independiente varías veces:

- **Regla ADELINE o Delta (Widrow y Hoff, 1960)**
- **Regla Least Means Square (1986)**
- **Regla Rescola wagner (1976)**

Algoritmo de actualización en línea (pseudo código)

Elegimos lo *N* parámetros ***W*** al azar

Sea *época=1*

Repetimos hasta que *época=época_maxima*

Sea mu =1

Repetimos, hasta que *mu=p*

$$O_i^\mu = \bar{W}_i \cdot \bar{\xi}^\mu \qquad i = 1, 2, \ldots, M$$

Calculamos la loss E para el ejemplo mu y obténemos los M valores de salida O

$$\Delta w_{ik} = -\eta \frac{\partial E}{\partial w_{ik}} = \eta\, \delta_i^\mu \xi_k^\mu$$

Calculamos los MxN valores de los incrementos de cada parámetro

$$\bar{W}_i = \bar{W}_i + \Delta \bar{W}_i$$

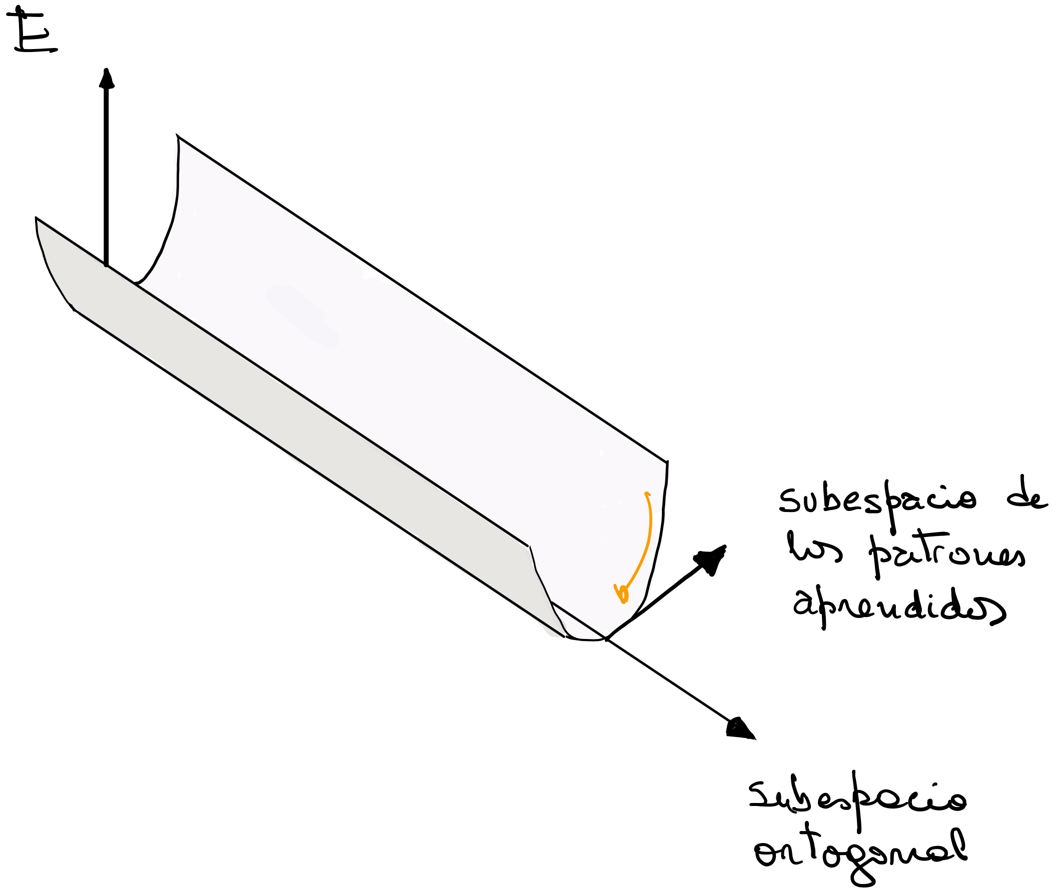
Actualizamos los MxN

mu = mu+1

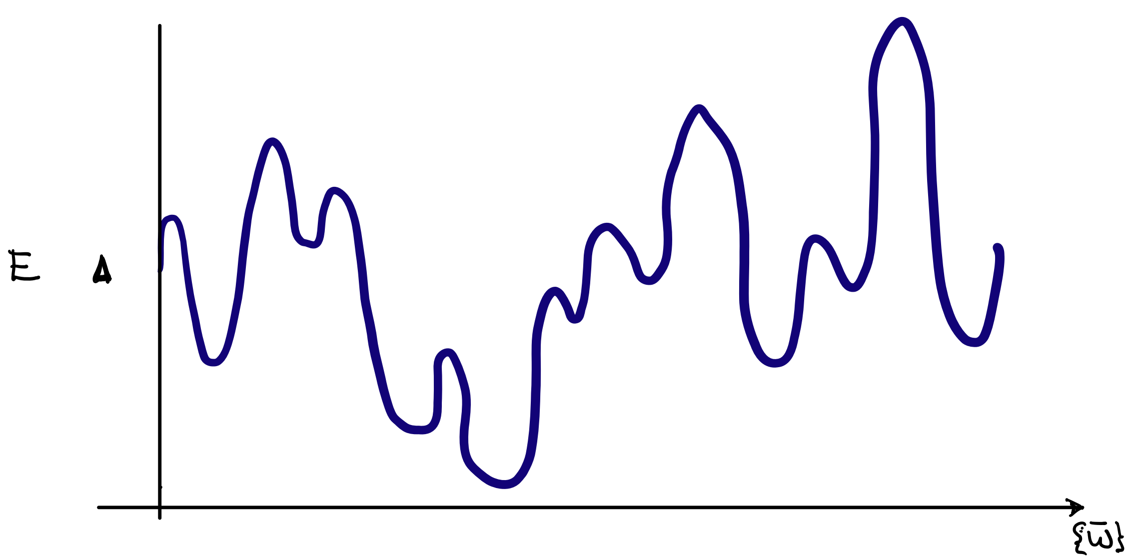
Volvemos a “Repetimos (mu)”

época = época + 1

Volvemos a “Repetimos (época)”

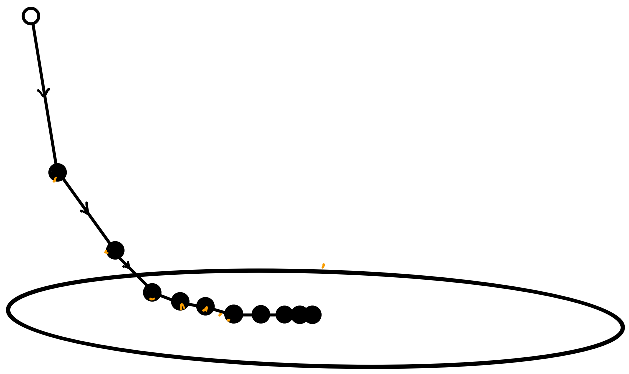


Ahora no vale la separabilidad lineal. Tenemos que intentar aprender descendiendo por la compleja función loss, la cual es bastante

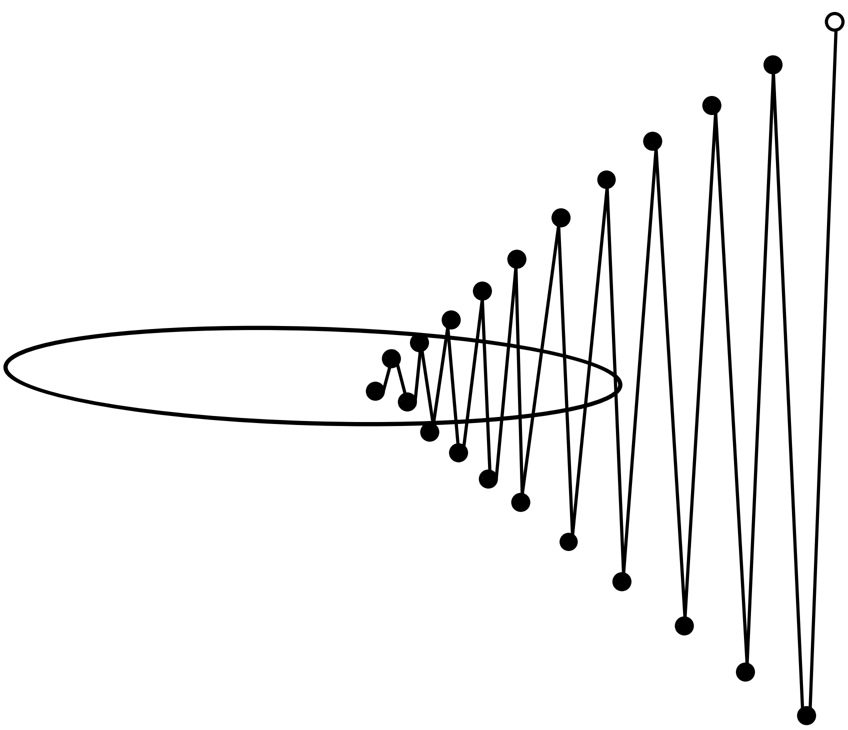


Ejemplo:

$$E = x^2 + 20y^2$$



$\eta = 0.02$



$\eta = 0.476$