Redes Neuronales profundas

Introducción

Gamaliel Moreno

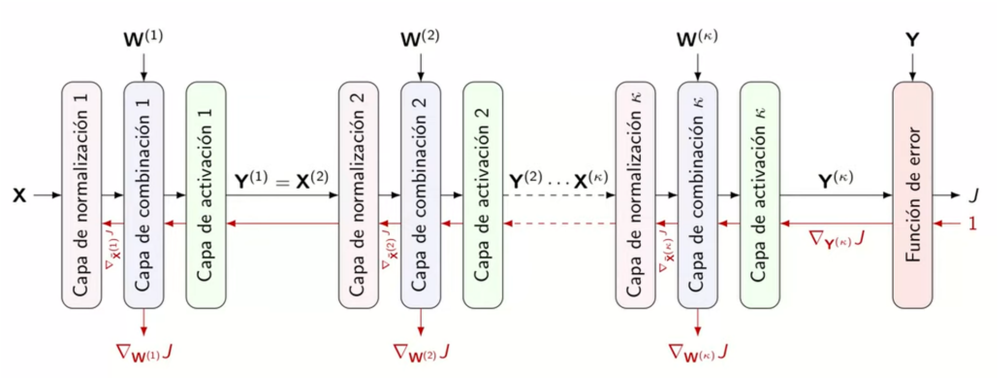
gamalielmch@uaz.edu.mx

Enero-julio 2021

Contenido

- Estructuras de redes profundas
- Capas de combinación
- Capas de activación
  - Capas simples
  - Softmax
- Normalización
- Funciones de pérdida
  - MSE
  - Entropía cruzada

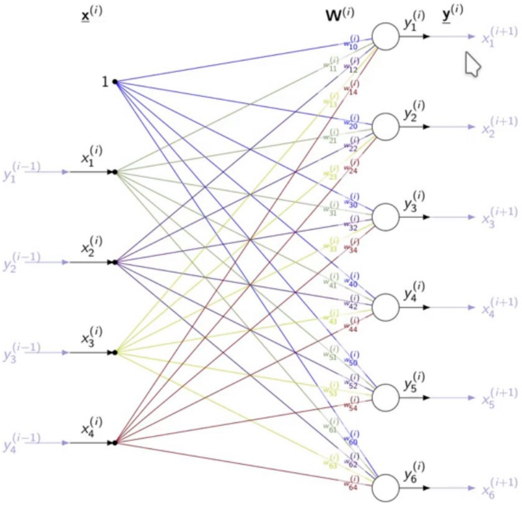
Estructura genérica de red neuronal profunda



Contenido

- Capas de combinación o transferencia
  - Capas totalmente conectadas
  - Capas convolucionales
- Capas de activación
  - Capas simples
    - Problemas : asimetría postiva, desvanecimiento del gradiente
- Normalización
- Funciones de pérdida
  - MSE
  - Entropía cruzada

Capa totalmente conectada o densa



$$\underline{\mathbf{y}}^{(i)} = \underline{\mathbf{g}}(\mathbf{W}^{(i)}\underline{\mathbf{x}}^{(i)})$$

$$\mathbf{W}^{(i)} = \begin{bmatrix} w_{10}^{(i)} & w_{11}^{(i)} & w_{12}^{(i)} & w_{13}^{(i)} & w_{14}^{(i)} \\ w_{20}^{(i)} & w_{21}^{(i)} & w_{22}^{(i)} & w_{23}^{(i)} & w_{24}^{(i)} \\ w_{30}^{(i)} & w_{31}^{(i)} & w_{32}^{(i)} & w_{33}^{(i)} & w_{34}^{(i)} \\ w_{40}^{(i)} & w_{41}^{(i)} & w_{42}^{(i)} & w_{43}^{(i)} & w_{44}^{(i)} \\ w_{50}^{(i)} & w_{51}^{(i)} & w_{52}^{(i)} & w_{53}^{(i)} & w_{54}^{(i)} \\ w_{60}^{(i)} & w_{61}^{(i)} & w_{62}^{(i)} & w_{63}^{(i)} & w_{64}^{(i)} \end{bmatrix}$$

$\underline{\mathbf{w}}_{:,0}^{(i)}$ es vector de **sesgo**/*bias*

Capa totalmente conectada o densa

- Estas capas producen $\boldsymbol{y}$ con una entrada $\boldsymbol{x}$ como combinación lineal :

$$\boldsymbol{y} = \boldsymbol{W} \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix}$$

- Concepto se generaliza para procesar mini-lotes para matriz de diseño $\boldsymbol{X}$ con los datos en sus filas (y salidas en filas de $\boldsymbol{Y}$)

$$\boldsymbol{Y} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix} \boldsymbol{W}^T$$

- Generalmente utilizadas con datos de información distribuida en todas las componentes (salida múltiples de sensores)

Inicialización de pesos

- ¿Cómo inicializamos las matrices de pesos $\boldsymbol{W}$?
- Dependerá de los tipos de capas de activación que usemos
- Usualmente serán números aleatorios distribuidos de cierta forma.
- El rango de valores utilizables es limitado
- Si se eligen muy pequeños, entrenamiento los aniquila ($\to 0$)
- Si se eligen muy grandes, explotan (NaN)
- Buena elección acelera convergencia
- Mala elección frena convergencia
- Marcos de trabajo ofrecen alternativas "estándar", pero cada método tiene un uso particular

Inicialización en capa totalmente conectada

- Asumamos que
  - entrada $\boldsymbol{x}$ tiene componentes i.i.d. $x_i \sim \mathrm{N}(0, \alpha_x)$
  - pesos $\boldsymbol{W}$ tiene componentes i.i.d. $w_ij \sim \mathrm{N}(0, \alpha_w)$
- Si $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ ¿cómo se distribuyen los y_i ?
- Para responder esto necesitamos dos propiedades :

$$E\left[\sum_i a_i\right] = \sum_i E\left[a_i\right]$$

$$E\left[\prod_i a_i\right] = \prod_i E\left[a_i\right]$$

  siempre que a_i sean independientes entre sí
- Se cumple que $y_i = \boldsymbol{w}_{i,:}^T \boldsymbol{x} = \sum_{k=1}^{n} w_{ik} x_k$

- El valor esperado para y_i es entonces

$$E\left[y_i\right] = E\left[\sum_{k=1}^{n} w_{ik} x_k\right] = \sum_{k=1}^{n} E\left[w_{ik} x_k\right] = \sum_{k=1}^{n} E\left[w_{ik}\right] E\left[x_k\right] = 0$$

  es decir, el valor medio sigue siendo cero.
- La varianza de y_i es (considerando que $E\left[y_i\right] = 0$)

$$Var\left[y_i\right] = E\left[(y_i - E\left[y_i)^2\right])\right] = E\left[y_i)^2\right] = E\left[\left(\sum_{k=1}^{n} w_{ik} x_k\right)^2\right]$$

$$= E\left[\left(\sum_{k=1}^{n} w_{ik} x_k\right)\left(\sum_{l=1}^{n} w_{il} x_l\right)\right] = \sum_{k=1}^{n}\sum_{l=1}^{n} E\left[w_{ik} x_k\right] E\left[w_{il} x_l\right]$$

$$= E\left[\left(\sum_{k=1}^{n} w_{ik}x_k\right)\left(\sum_{l=1}^{n} w_{il}x_l\right)\right] = \sum_{k=1}^{n}\sum_{l=1}^{n} E\left[w_{ik}x_k\right] E\left[w_{il}x_l\right]$$

$$= \sum_{k=1}^{n}\sum_{l=1}^{n} E\left[w_{ik}w_{il}x_k x_l\right] = \sum_{k=1}^{n}\sum_{l=1}^{n} E\left[w_{ik}w_{il}\right] E\left[x_k x_l\right]$$

donde x_k y x_l son i.i.d. solo si $k \neq l$, y lo mismo para w_{ik} y w_{il}

Inicialización en capa totalmente conectada

- Esto quiere decir que si $k \neq l$ entonces $E\left[w_{ik}w_{il}\right] = 0$ y $E\left[x_k x_l\right] = 0$

$$Var\left[y_i\right] = \sum_{k=1}^{n}\sum_{l=1}^{n} E\left[w_{ik}w_{il}\right] E\left[x_k x_l\right] = \sum_{k=1}^{n} E\left[w_{ik}^2\right] E\left[x_{ik}^2\right] = \sum_{k=1}^{n} \sigma_w^2 \sigma_x^2$$

..

$$\sum_{k=1}^{n} \sigma_w^2 \sigma_x^2 = n\sigma_w^2 \sigma_x^2$$

- si la entrada tiene $\sigma_x = 1$, entonces $Var[y_i] = n\sigma_w^2$
- Si pegáramos varias capas, la varianza !crece con ! (números posibles van a crecer)
- Luego de varias capas, esos números alcanzarían NaN
- Si queremos que $Var[y_i] = 1$ entonces tenemos que elegir $\sigma_w = 1/\sqrt{n}$

Inicialización en capa totalmente conectada

- El análisis anterior solo considera la capa conectada aislada
- Falta considerar la capa de activación
- Salida de esa capa debería tener varianza unitaria (entrada a siguiente capa)

Capas convolucionales