



REDES NEURONALES 2024

Clase 23 parte 1

Lunes 11 de noviembre 2024

FAMAF, UNIVERSIDAD NACIONAL DE CÓRDOBA

INSTITUTO DE FÍSICA ENRIQUE GAVIOLA (UNC-CONICET)

5.- REGULARIZACIONES

Agregamos términos adicionales a la función error (o costo, o loss) que "limitan" los valores de los parámetros según nuestros deseos

hyperparámetros

$$E(\bar{w}, \bar{s}) = \sum_{j=1}^p \sum_i e(\bar{w}, \bar{s}_i) + \lambda R(f)$$

funció que tenemos que aprender

Regularización L₁ o Lasso (least shrinkage and selection operator)

$$\tilde{E}_L(\{\bar{w}\}, \{\bar{s}\}) = \frac{1}{2} \sum_{k=1}^p \sum_{i=1}^n (\bar{s}_i^k - \theta_i^k)^2 + \lambda \sum_{\alpha} |w_{\alpha}|$$

↑
Suma sobre todos los coeficientes y unidades

Regularización L₂ o Ridge

$$\tilde{E}_L(\{\bar{w}\}, \{\bar{s}\}) = \sum_{k=1}^p \sum_{i=1}^n (\bar{s}_i^k - \theta_i^k)^2 + \lambda \sum_{\alpha} w_{\alpha}^2$$

↑
Suma sobre todos los coeficientes y unidades

6.- Podemos reemplazar el ECM por otras funciones error o loss

Cross Entropy: La función error se ha definido como la suma de los cuadrados de la diferencia entre los valores deseados y la salida obtenida por la red. Por la naturaleza de esta función y dado que el método del descenso por el gradiente se definió proporcional al $\frac{\partial E}{\partial w}$, cuanto más próximos estemos al mínimo donde sea $E = C = 0$ menor será la velocidad (pendiente) con la que nos acerquemos. Si esta función costo fuera reemplazada por una función logarítmica, el descenso por el gradiente iría incrementando su velocidad de acercamiento al mínimo dado que el gradiente sería cada vez mayor. Esto se suele usar como una métrica para saber cuán cerca estamos de la solución correcta.

Una de las funciones loss más usada es la llamada **ENTROPÍA CRUZADA** (en inglés *cross entropy*). No es fácil entender porqué esta función funciona muy bien, sobre todo en los problemas de clasificación, pero lo intentaremos.

La idea entropía surgió en la formalización de las teorías sobre la termodinámica. El nombre fue introducido por primera vez por el físico alemán Rudolf Clasius en 1865 para referirse a la forma de energía en la cual inevitablemente toda energía se transforma: **el calor**.

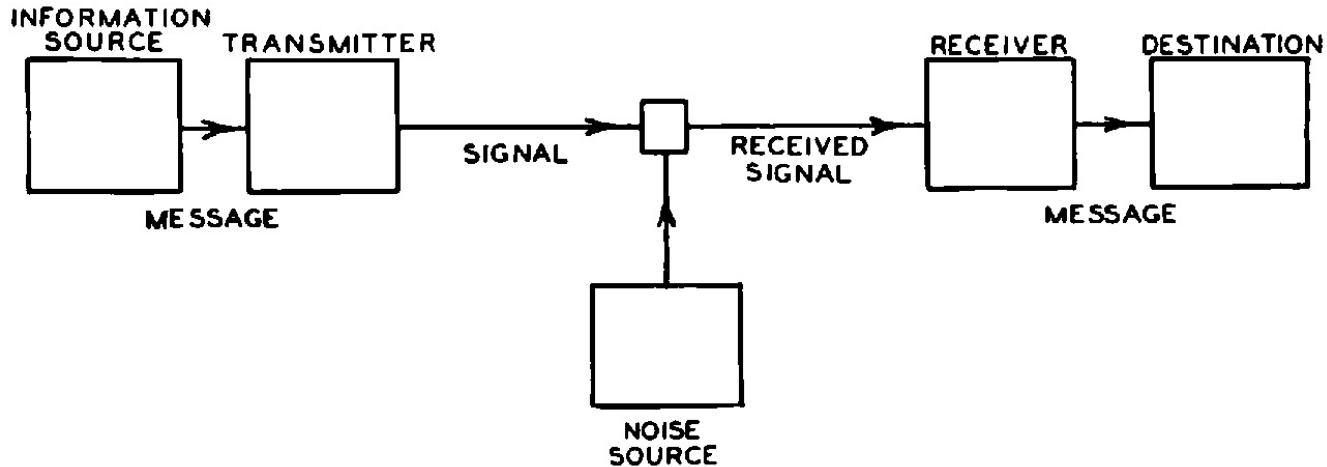
$$\Delta S = \int \frac{dQ}{T}$$

El mismo concepto fue usado por Boltzmann y Gibbs en forma independiente:

$$S(A) = k_B \log_2 (\#M_A)$$

k_B es una constante universal (llamada constante de Boltzmann). Claude Shannon a su vez usó este concepto para aplicarlo a la teoría de la información. Un transmisor codifica información en una señal que es corrompida por ruido para ser luego decodificada por el receptor. El propuso separar y modelar la señal y el ruido de forma probabilística.





Shannon definió dos conceptos importantes:

El mínimo número de bits por segundo para representar la información y lo llamó *razón de entropía H*. Este número cuantifica la incertezza involucrada al querer determinar cuál es el mensaje generado por la fuente. Cuanto menor es la razón de entropía, menor es la incertezza y es más simple codificar el mensaje. Si mandamos por minuto 100 letras de un texto necesitamos 470 bits, pues:

$$26^{100} \approx 2^{470}$$

$$\log_2(26^{100}) \approx \log_2(2^{470}) = 470$$

2. Definió el máximo número de bits por unidad de tiempo que puede transferirse con confianza en la presencia de ruido y la llamó *capacidad de información C*.

Finalmente él dedujo que es posible enviar información confiable a menos que:

$$H < C$$

Treatemos de entender las ideas de la entropía cruzada.

- A) Si leyo una muestra de sus adulteros y les puse el mensaje **SALIO CARA** el mensaje les da cierta información. ¿Cuánta información?

Sale cara (probabilidad $\frac{1}{2}$): información = 1 bit

- B) Me levanto en la mañana (tarde) y les puse el siguiente mensaje: **AMANECÍO**. Ustedes van a considerar mi mensaje inicial pues todos los días el sol aparece y el día amanece. No estoy enviando información.

Amaneció (probabilidad = 1): información = 0 bits

- C) Un día me levanto antes que ustedes, y descubro con horror que el sol no salió. Algo astromóricularmente catástrofico sucedió. Es una super noticia:

No sucede ($\text{probabilidad} = 0$): información = ∞ bits

La información contenida en un mensaje sobre cualquier evento está fuertemente vinculada a la incertidumbre o a la sorpresa del evento. Que sucede un evento improbable nos da más información que la ocurrencia de un evento usual o esperado.

Shannon definió la información como

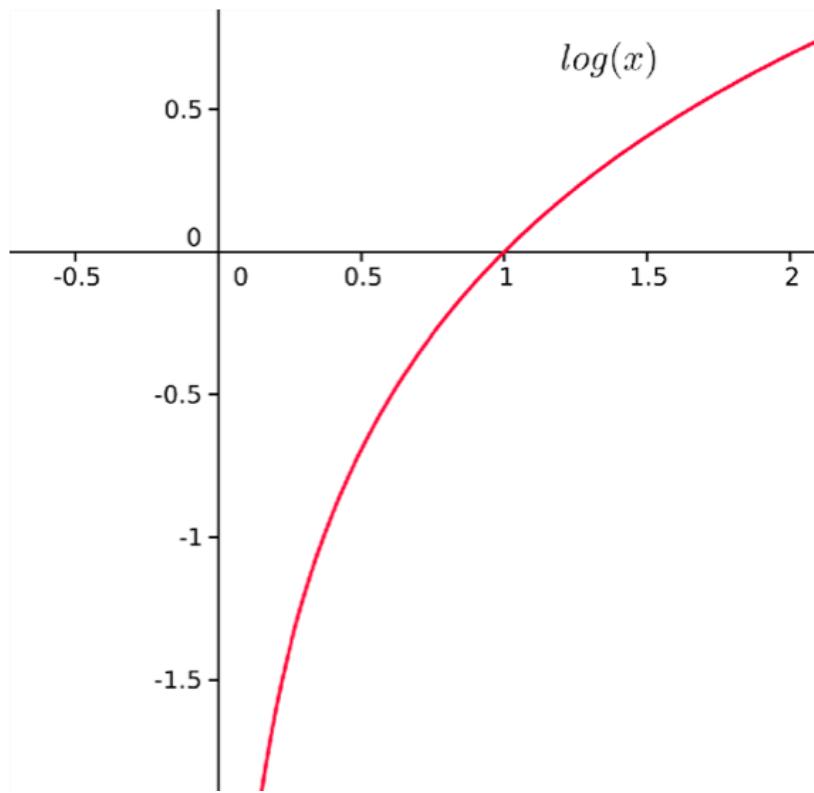
$$I(x) = -\log_2(P(x))$$

donde $P(x)$ es la probabilidad de ocurrencia de x . Noten que con esta definición

si $P(x) = \frac{1}{2}$ $I(x) = -\log_2\left(\frac{1}{2}\right) = -(-\log_2(2)) = 1$

si $P(x) = 1$ $I(x) = -\log_2(1) = 0$

si $P(x) = 0$ $I(x) = -\log_2(0) = +\infty$



¿Cómo generalizamos este definición a una probabilidad de eventos?

Ejemplo: supongamos que el experimento consiste en sacar una pelotilla de una caja. En la caja hay nueve (9) pelotillas de diferentes colores:

2 pelotillas rojas

3 pelotillas verdes

4 pelotillas azules

¿ Cuánta información obtenemos cuando recibimos una pelotita ?

$$P(\text{sacar 1 pelotita roja}) = \frac{2}{9}$$

$$I(\text{sacar 1 pelotita roja}) = -\log_2\left(\frac{2}{9}\right)$$

$$P(\text{sacar 1 pelotita verde}) = \frac{3}{9} = \frac{1}{3}$$

$$I(\text{sacar 1 pelotita verde}) = -\log_2\left(\frac{1}{3}\right)$$

$$P(\text{sacar 1 pelotita azul}) = \frac{4}{9}$$

$$I(\text{sacar 1 pelotita azul}) = -\log_2\left(\frac{4}{9}\right)$$

Ahora podemos definir la entropía del evento
SACAR UNA PELOTITA DE ESTA CAJA EN
ESPECIAL como el "valor esperado" de la
información I que surge de realizar el
experimento y sacar una pelotita de color x .

$$\text{Entropía} = E[I(\text{todas las pelotitas})]$$

$$= -\left(\frac{2}{9}\right) \log_2\left(\frac{2}{9}\right) - \left(\frac{1}{3}\right) \log_2\left(\frac{1}{3}\right) - \left(\frac{4}{9}\right) \log_2\left(\frac{4}{9}\right)$$

$$H(P) \equiv \text{Entropía} = \mathbb{E}_{x \sim P} (-\log_2 (P(x)))$$

$X \sim P$: x obtenidos con probabilidad $P(x)$

Con este definición de Información y Entropía de la información podemos definir la entropía cruzada

Entropía cruzada (cross entropy)

La entropía cruzada mide la entropía relativa entre dos distribuciones de probabilidad diferentes, $P(x)$ y $Q(x)$ definidas sobre el mismo conjunto de eventos X .

$$H(P, Q) = \mathbb{E}_{x \sim P} [-\log (Q(x))]$$

O sea, calcula las informaciones con Q para los eventos con P .

$$H(P, Q) = \sum_{x \sim P} P(x) (-\log_2 (Q(x)))$$

$$H(P, Q) = \int dx P(x) \left(\log_2(Q(x)) \right)$$

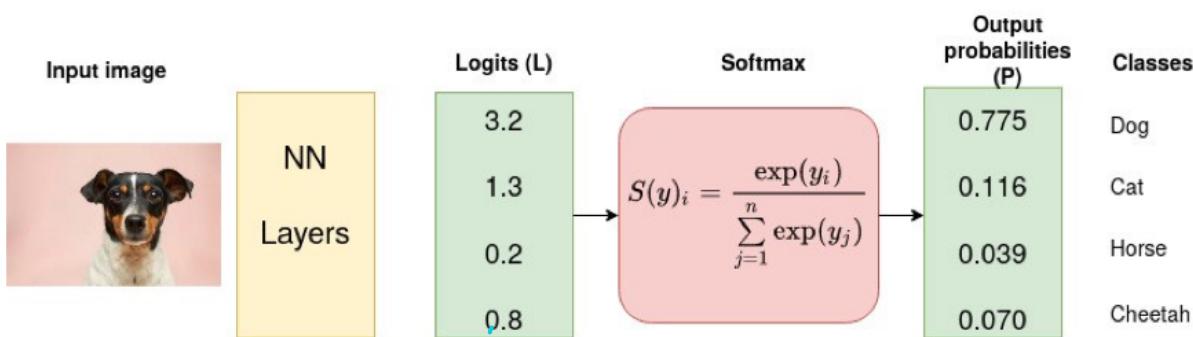
Miremos un ejemplo, imaginando que tenemos dos experimentos similares con la caja I y II con distintos números de perlas de cada color.

		rojos	verdes	azules
CAJA I	P	2	3	4
CAJA II	Q	4	4	1

$$H(P, Q) = - \left(\frac{2}{9} \right) \log_2 \left(\frac{4}{9} \right) - \left(\frac{3}{9} \right) \log_2 \left(\frac{4}{9} \right) - \left(\frac{4}{9} \right) \log_2 \left(\frac{1}{9} \right)$$

Entropía cruzada como loss

Vemos ahora como usar la entropía cruzada en inteligencia artificial. Pero en concreto con un ejemplo de clasificación



En la entrada tenemos imágenes de 4 tipos de animales (perros, gatos, caballos y leones). Una red neuronal feed-forward calcula 4 neuronas de salida ($\text{logit} = \log(x) / \log(1-x)$) y una cuenta simple. Clasifica softmax los lleva a las cuatro variables de salida

$S(y_1)$: prob. de perro

$S(y_2)$: prob. de gato

$S(y_3)$: prob. de caballo

$S(y_4)$: prob. de león

de forma tal que

$$0 \leq S(y_i) \leq 1$$

$$S(y_1) + S(y_2) + S(y_3) + S(y_4) = 1$$

No son estrictamente probabilidades sino estimados

En el ejemplo, la red nos dice que es más probable que sea un perro ($S(y_1) = 0.775$) y muy poco probable que sea un chico ($S(y_4) = 0.070$).

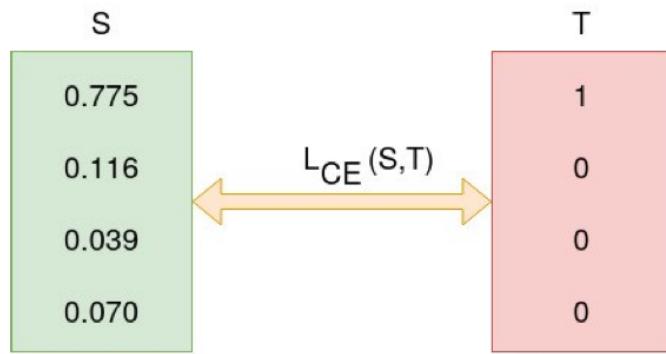
Pero supongamos que el experto que etiquetó el conjunto de entrenamiento sabe que es un perro, entonces nosotros conocemos la verdadera (T) distribución del experimento "mostrar este foto a la red y clasificarla"

Probab. de que sea un perro $t_1 = 1$

Probab. de que sea un gato $t_2 = 0$

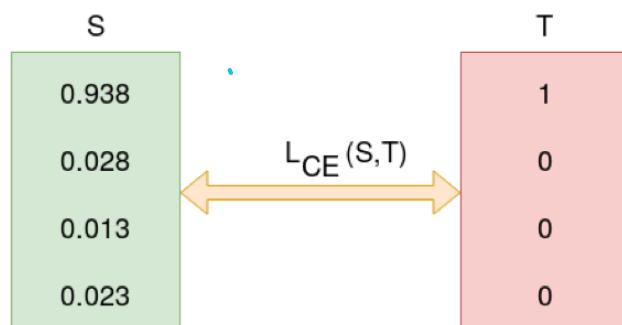
Probab. de que sea un caballo $t_3 = 0$

Probab. de que sea un chico $t_4 = 0$



$$\begin{aligned}
 L_{CE} = H(S, T) &= \sum_{i=1}^4 t_i \left[-\log_2(s_i) \right] \\
 &= - [1 \cdot \log_2(0.775) + 0 \cdot \log_2(0.116) + \\
 &\quad 0 \cdot \log_2(0.039) + 0 \cdot \log_2(0.070)] \\
 &= -\log_2(0.775) = 0.3677
 \end{aligned}$$

Supongamos que le mostramos la misma foto pero con las cifras después y obtenemos:



$$L_{CE} = -\log_2(0.938) = 0.095$$

Pore el caso de clasificación binaria con aprendizaje supervisado, si la recta es sigmoidal

$$\sigma_i^{\mu} = \frac{1}{1 + e^{-h_i^{\mu}}}$$

$$E(\{w\}) = - \sum_{i=1}^{M,p} y_i^m \log_2(\sigma_i^m) + (1-y_i^m) \log_2(1-\sigma_i^m)$$

$$E \geq 0 \quad \bar{J}^{\mu} \neq \bar{J}^{\mu}$$

$$E = 0 \quad \bar{J}^{\mu} = \bar{J}^{\mu}$$

