Conceptos previos acerca de redes neuronales

1. Minimización del error cuadrático.

Cuando se entrena una red neuronal, el objetivo principal es obtener el menor error posible a la hora de realizar una predicción de clases. Esta predicción de clases viene dada por una distribución de probabilidad:

$$P(x_1, x_2 \dots x_n|\sigma) = \prod_{i=1}^{n} P(x_i|\sigma)$$

El objetivo principal es obtener un vector:

$$(\sigma_1, \sigma_2 \dots \sigma_n)$$

Este vector va a estar formado por los conocidos **estimadores de verosimilitud**, que se corresponderán con los pesos de las conexiones, y que tiene como objetivo maximizar la distribución probabilidad que rige la **función de verosimilitud** (*likelihood function*). Maximizar esta función de verosimilitud va a implicar forzosamente que se minimice el **error cuadrático medio** (si aumentamos *f(x)*, disminuimos *-f(x)*).

Es importante decir también que estas distribuciones de probabilidad vienen dadas por una función de separación de clases, ya sea lineal o no lineal. En este caso supondremos que la función de separación es una recta con ecuación:

$$y = wx + b$$

Para maximizar (o minimizar) una función, simplemente hay que estimar su derivada e igualarla a cero (de esta forma se obtienen los máximos y mínimos)

1.1 Datos que siguen una distribución continua (Gaussiana)

En este caso estudiaremos el caso en que los datos siguen una distribución continua (no suele ser así). Cuando es así, podemos expresar la función de separación de clases de la siguiente forma:

$$y \approx N(wx, b)$$

donde:

$$\mu = wx$$
$$\sigma = b(cte)$$

De esta forma la distribución de probabilidad queda de la siguiente forma:

$$P(x_i|\sigma) = P(y|wx)$$

Como hemos mencionado antes, el objetivo es minimizar el error de nuestra red o lo que es lo mismo, maximizar la función de verosimilitud de la distribución que siguen los datos. Por tanto debemos estimar la derivada de nuestra distribución de probabilidad:

$$\sigma = 1$$

$$\prod_{i=1}^{n} P(x_i|\sigma) = \sum_{i=1}^{n} log(P(x_i|\mu_i, \sigma)) =$$

$$\sum_{i=1}^{n} log\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma}(y_i - \mu_i)^2)$$

Hemos supuesto el valor cte 1 para que al derivar el cuadrado se anule con *1/2* y de esta forma la expresión final sea más sencilla. Podemos ver que la exponencial y el logaritmo se anulan. Nuestra expresión final a derivar es la siguiente:

$$\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2}(y_i - \mu_i)^2$$

A continuación derivamos esta expresión:

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2}(x_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \mu_i)$$

Por tanto la igualdad que buscábamos es:

$$\sum_{i=1}^{n}(y_i - \mu_i) = 0$$

$$\mu_i = w_i x_i$$

$$\sum_{i=1}^{n}(y_i - \mu_i) = \sum_{i=1}^{n}(y_i - w_i x_i)$$

$$\sum_{i=1}^{n}(y_i - w_i x_i) = 0$$

1.2 Datos que siguen una distribución discreta (Bernoulli)

Otra posibilidad (más frecuente) es que los datos sigan una distribución discreta. En este caso asumiremos que la distribución discreta que siguen es una Bernoulli.

En este caso la función de probabilidad cambia, puesto que la Bernoulli sigue la siguiente función de probabilidad.

$$\mu^{x_i}(1-\mu)^{1-x_i}$$

En este caso la distribución acepta solo dos valores (0,1). Nuestro objetivo sin embargo es el mismo que el anterior. Estimar la derivada de la función de verosimilitud para maximizarla.

$$\prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i} = \sum_{i=1}^{n} x_i log(\mu) + \sum_{i=1}^{n}(1-x_i)log(1-\mu) =$$

$$\sum_{x_i}^{n} x_i \frac{1}{log(\mu)} - \sum_{x_i}^{n}(1-x_n)\frac{1}{log(1-\mu)} = 0$$

Como antes, tenemos una función de separación de clases, en este caso lineal, con la misma forma que antes. Así, tenemos:

$$\mu = wx_i$$

$$\sum_{x_i}^{n} x_i \frac{1}{log(wx_i)} - \sum_{x_i}^{n}(1-x_n)\frac{1}{log(1-wx_i)} = 0$$

Finalmente obtenemos la función de coste, que es la siguiente:

$$ylog(wx) + (1-y)log(1-wx)$$

Esta expresión nos da la **función de coste de la regresión logística**:

$$J = -\frac{1}{n}\sum_{i=1}^{n} y_i log(w_i x_i) + (1-y_i)log(1-wx_i)$$

O lo que es lo mismo, la expresión anterior con **signo contrario**