# Tarea 5 - Redes neuronales

Sofía Chávez Bastidas

## P1

Considere el siguiente dataset de clasificación multiclase $\{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, con $y^{(i)} \in \{c_1, ..., c_k\}$. Dadas $p$ y $q$ dos distribuciones de probabilidad discretas, la **entropía cruzada** está dada por:

$$H(p, q) = \sum_{j=1}^{K} p(c_j) \log\left(\frac{1}{q(c_j)}\right) \tag{1}$$

Se define entonces la **función de pérdida de entropía cruzada** mediante:

$$L(q, p) = \frac{1}{N} \sum_{i=1}^{N} H(p_i, q_i) \tag{2}$$

Demuestre que

$$\hat{\theta} = \arg\min_{\theta} L(p_\theta, p) \tag{3}$$

donde $\hat{\theta}$ es el estimador de máxima verosimilitud y $p$ es la distribución de probabilidad empírica, es decir, la distribución que se puede "observar" en los datos, la cual, en el caso no condicional puede ser vista como:

$$p_e(y) = \frac{1}{N} \sum_{i=1}^{N} \delta(y - y^{(i)}) \tag{4}$$

con $\delta$ la función delta de Dirac.

---

Consideramos un conjunto de datos de clasificación multiclase $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, donde $y^{(i)} \in \{c_1, ..., c_k\}$.

Definimos:

- $p_e$: La distribución empírica de las clases en los datos, dada por $p_e(y) = \frac{1}{N} \sum_{i=1}^{N} \delta(y - y^{(i)})$.
- $p_\theta$: La distribución predicha por el modelo parametrizado por $\theta$.

Para encontrar el estimador de máxima verosimilitud $\hat{\theta}$, buscamos:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} p(\mathcal{D}|\theta) \tag{5}$$

Asumiendo que los datos son iid, podemos construir $\mathcal{L}(\theta)$ como:

$$\mathcal{L}(\theta) = p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p_\theta(y^{(i)}) \tag{6}$$

Con lo cual podemos, equivalentemente, maximizar la log likelihood $\ell$:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \mathcal{L} = \arg\max_{\theta} \ell = \arg\max_{\theta} \log(\mathcal{L}) \\ &= \arg\max_{\theta} \log\left(\prod_{i=1}^{N} p_\theta(y^{(i)})\right) = \arg\max_{\theta} \sum_{i=1}^{N} \log(p_\theta(y^{(i)})) \end{aligned} \tag{7}$$

Por otro lado, podemos desarrollar $L(p_\theta, p)$ considerando que en la entropía cruzada $H$, $p_i$ es la distribución empírica $p_e$ y $q_i$ es la distribución predicha por el modelo con parámetro $\theta$, es decir $p_i = p_e$ y $q_i = p_\theta$:

$$L(p_\theta, p) = \frac{1}{N} \sum_{i=1}^{N} H(p_e, p_\theta) \tag{8}$$

Con lo cual

$$\hat{\theta} = \arg\min_{\theta} L(p_\theta, p) \tag{9}$$

Para demostrar esto, consideramos la definición de la función de pérdida de entropía cruzada y la desarrollamos:

$$L(p_\theta, p) = \frac{1}{N} \sum_{i=1}^{N} H(p_e, p_\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_e(c_j) \log\left(\frac{1}{p_\theta(c_j)}\right) \tag{10}$$

Dado que $p_e$ es la distribución empírica y $\sum_{i=1}^{N}\sum_{j=1}^{K} p_e(c_j)\log\left(\frac{1}{p_\theta(c_j)}\right)$ equivale a $\sum_{i=1}^{N}\log\left(\frac{1}{p_\theta(c_j)}\right)$ cuando $c_j$ es la clase real, es decir, $y^{(i)}$, podemos simplificar:

$$L(p_\theta, p) = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{p_\theta(y^{(i)})}\right) = \frac{-1}{N}\sum_{i=1}^{N}\log(p_\theta(y^{(i)})) \tag{11}$$

Con lo cual

$$\begin{aligned}
\hat{\theta} &= \arg\min_\theta L(p_\theta, p) \\
&= \arg\min_\theta \frac{-1}{N}\sum_{i=1}^{N}\log(p_\theta(y^{(i)})) \\
&= \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N}\log(p_\theta(y^{(i)})) \\
&= \arg\max_\theta \sum_{i=1}^{N}\log(p_\theta(y^{(i)}))
\end{aligned} \tag{12}$$