PROF. FRANCESCO A. N. PALMIERI
UNIV. DELLA CAMPANIA Nov. 2023

APPENDIX

CROSS ENTROPY

Give two discrete distributions p and q,

$p = \{p_1, \dots, p_n\} \quad 0 \le p_i \le 1 \quad \sum_{i=1}^{n} p_i = 1$

$q = \{q_1 \dots q_n\} \quad 0 \le q_i \le 1 \quad \sum_{i=1}^{n} q_i = 1$

the cross-entropy between p and q is

$$H(p,q) \triangleq \sum_{i=1}^{n} p_i \lg \frac{1}{q_i}$$

Note that is not symmetric: $H(p,q) \neq H(q,p)$

The cross-entropy is derived from the KL-DIVERGENCE (Kullback-Leibler) between p and q

$$KL(p\|q) \triangleq \sum_{i=1}^{n} p_i \lg \frac{p_i}{q_i}$$

KL is well-known in information theory and it is positive $KL(p\|q) \ge 0 \quad \forall p,q$ and non symmetric $KL(p\|q) \neq KL(q\|p)$ (see any introductory text on information theory). Rewriting

$$KL(p\|q) = \sum_{i=1}^{n} p_i \lg \frac{1}{q_i} - \sum_{i=1}^{n} p_i \lg \frac{1}{p_i},$$

we recognize the entropy of p, therefore

$$KL(p\|q) = H(p,q) - H(p),$$

and

$$H(p,q) = KL(p\|q) + H(p) \ge 0$$

because both terms are positive (the entropy is always ≥ 0).

The cross-entropy is used as a measure of "distance" between p and q even if when $p=q$, $KL(p\|q)=0$ and $H(p,q)=H(p)$. This is acceptable because usually the cross-entropy is used to find q, with p fixed:

$$\min_q H(p,q) = \min_q \left(KL(p\|q) + H(p)\right) = \min_q KL(p\|q)$$

therefore minimization of $H(p,q)$ with respect to q is equivalent to KL minimization with respect to q.

More about cross-entropy and KL divergence can be found in any text of information theory. We limit ourselves here to what is directly relevant to machine learning.

GRADIENT OF CROSS-ENTROPY

In learning algorithms it is useful to consider the gradient of $H(p,q)$ with respect to $\{q_1, q_2 \dots q_n\}$

$$\nabla_q H(p,q) = \begin{bmatrix} \frac{\partial H(p,q)}{\partial q_1} \\ \vdots \\ \frac{\partial H(p,q)}{\partial q_n} \end{bmatrix}$$

$$\frac{\partial}{\partial q_\ell} H(p,q) = \frac{\partial}{\partial q_\ell} \sum_{i=1}^{n} p_i \lg \frac{1}{q_i} = -\frac{\partial}{\partial q_\ell} \sum_{i=1}^{n} p_i \lg q_i = -\frac{p_\ell}{q_\ell}$$

$$\nabla_q H(q,q) = -\begin{bmatrix} \frac{p_1}{q_1} \\ \frac{p_2}{q_2} \\ \vdots \\ \frac{p_n}{q_n} \end{bmatrix}$$

BINARY CROSS ENTROPY

The binary case, as usual, deserves special attention
Here p and q are binary distributions

$$p = \{\alpha, 1-\alpha\} \quad q = \{\beta, 1-\beta\}$$

the cross entropy is

$$H(p,q) = \alpha \log\frac{1}{\beta} + (1-\alpha)\log\frac{1}{1-\beta}$$

Now, if p is fixed and we look for the gradient with respect to q, we only need the derivative of $H(p,q)$ with respect to β which is

$$\frac{\partial H(p,q)}{\partial \beta} = -\frac{\partial}{\partial \beta}\left[\alpha \log \beta + (1-\alpha)\log(1-\beta)\right]$$

$$= -\frac{\alpha}{\beta} + \frac{(1-\alpha)}{1-\beta} = \frac{-\alpha(1-\beta) + \beta(1-\alpha)}{\beta(1-\beta)} = \frac{-\alpha + \cancel{\alpha\beta} + \beta - \cancel{\alpha\beta}}{\beta(1-\beta)}$$

$$= \frac{\beta - \alpha}{\beta(1-\beta)}$$