

From Least Squares to Cross Entropy

j.p.lewis

first draft beware of typos

Start with least squares,

$$\min_{y_k} \sum_k (y_k - x_k)^2 \tag{1}$$

where x_k are the given data and y_k are the corresponding points estimated by the model. This can be related to cross-entropy in two steps: 1) convert into a likelihood, 2) convert to KL-divergence between the data and model probabilities. In the “conversion” process there are several steps that transform through a log or exp, or by negating and flipping max/min. These are monotonic transformation and do not alter the location of the solution.

1. Least squares to likelihood

To convert into a likelihood, we need to maximize something rather than minimize. Negate, and switch the min to a max:

$$\max_{y_k} \sum_k -(y_k - x_k)^2$$

Note that the location of the maximum is the same as the location of the minimum in (??), we have just flipped the quadratic bowl upside-down.

Next take the exponential of this, and recall the rule “the log of a product is the sum of the logs”.

$$= \max \prod_k \exp(-(y_k - x_k)^2)$$

This has the form of a product of Gaussians. If the data has different error variances at different points this could be generalized to

$$= \max \prod_k \exp\left(-\frac{(y_k - x_k)^2}{\sigma_k^2}\right)$$

This has the form of a likelihood in the case where the errors are independent and therefore factor as a product over the individual data points:

$$P(\mathbf{x}|\theta) = \prod_k P(x_k|\theta)$$

where $P(x_k|\theta) \propto \exp(-(y_k - x_k)^2/\sigma^2)$.

2. Likelihood to KL divergence

Usually instead of maximizing the likelihood, the negative of the log likelihood is minimized. This goes backwards a few steps.

$$\max \prod_k P(x_k|\theta) \quad \Rightarrow \quad \min - \sum_k \log P(x_k|\theta)$$

(σ_k is dropped for simplicity).

Replace the sum with an average.

$$\min -\frac{1}{N} \sum \log P(x_k|\theta)$$

A tricky part (when going from likelihood to KL divergence rather than in the other direction), rewrite this as

$$\min -\sum_k \left[\frac{1}{N} \sum_i \delta(x_k - x_i) \right] \log P(x_k|\theta)$$

and the term in brackets is the data probability distribution, $P_D(x) = \frac{1}{N} \sum_i \delta(x - x_i)$.

Rewrite using $P_D(x)$

$$\min -\sum_k P_D(x_k) \log P(x_k|\theta)$$

This is now in “entropy land”.

The minimum is unchanged if we add any term that does not involve the model parameters θ . The term to add is the negative entropy of the probability of the data,

$$\begin{aligned} &= \min \left[\sum_k P_D(x_k) \log(P_D(x_k)) \right] - \sum_k P_D(x_k) \log P(x_k|\theta) \\ &= \min \sum_k P_D(x_k) \log \frac{P_D(x_k)}{P(x_k|\theta)} \\ &= \min \text{KL} [P_D(x) \| P(x|\theta)] \end{aligned}$$

To give a more conventional appearance, rewrite $P(x|\theta) \Rightarrow P_\theta(x)$,

$$= \min \text{KL} [P_D(x) \| P_\theta(x)]$$

I.e. the result is to minimize the KL divergence between the data and model probabilities.

The cross-entropy is $-\sum_k P_D(x_k) \log P(x_k|\theta)$. It appeared above by dropping the negative data entropy term after noting that the latter does not affect the location of the minimum.

Recapping, a general statement of model fitting is to minimize the KL divergence between the data and model probabilities. Cross-entropy appears in ignoring a term that does not depend on the model parameters and thus is not used in the computation. Least squares is obtained when the model assumes independent Gaussian-distributed errors.