From Least Squares to Cross Entropy

j.p.lewis

first draft beware of typos

Comment: it is more sensible to start with KL divergence, the more fundamental quantity, and derive least squares as a special case. This note is for people who are familiar with least squares but less so with entropy.

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 (1), 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 equal to the sum of the logs".

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

This has the form of a product of Gaussians. (Note, for simplicity this is ignoring both the $2\sigma^2$ in the argument to $\exp$ in the Gaussian formula, and the normalizing constant in front of the Gaussian. Including the normalizing constant results in an additional regularizing term that prefers σ to be 1.) If the data has different error variances at different points this could be generalized to

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

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.

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

(σ_k is dropped for simplicity).

Replace the sum with an average.

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

A tricky part (when going from likelihood to KL divergence rather than in the other direction). Consider a toy dataset where the data has values $x_k = \{1, 2, 2, 7, 4\}$. The sum above is then

$$\frac{1}{5} (\log P(1|\theta) + \log P(2|\theta) + \log P(2|\theta) + \log P(7|\theta) + \log P(4|\theta))$$

where (as a reminder) $P(1|\theta)$ is the likelihood of the value 1 under the model with parameters θ. This can be re-written as a sum over the unique values of the data, rather than over the data,

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

Here x_i indexes the unique values $\{1, 2, 7, 4\}$, skipping the repeated 2. The sum in brackets is a loop over all the data, This is the empirical (data) probability distribution, $P_D(x) = \frac{1}{N} \sum_i \delta(x - x_k)$. The bracketed term gives $1/N$ for each non-repeated datum, or $2/N$ for a data item that has one duplicate, etc.

Rewrite using $P_D(x)$

$$\min \quad -\sum_i P_D(x_i) \log P(x_i|\theta)$$

This is now in “entropy land” – it is the cross entropy of (data,model).

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,

$$= \min \quad \left[\sum_i P_D(x_i) \log(P_D(x_i)) \right] - \sum_i P_D(x_i) \log P(x_i|\theta)$$

$$= \min \quad \sum P_D(x_i) \log \frac{P_D(x_i)}{P(x_i|\theta)}$$

$$= \min \quad KL\left[P_D(x) \| P(x|\theta)\right]$$

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

$$= \min \quad KL\left[P_D(x) \| P_\theta(x)\right]$$

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

The cross-entropy is $-\sum_i P_D(x_i) \log P(x_i|\theta)$. 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 diverergence 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.