# Note on the relationship between Error, Loss Function and Maximum Likelihood

Dr François Pitié

---

As we have seen in the handout on Least Squares, there is a very fundamental link between the distribution of the prediction error and the type of loss function you should consider.

Very early on, Gauss connected Least squares with the principles of probability and to the Gaussian distribution.

Recall that the linear model is:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

The error $\boldsymbol{\varepsilon}$ is the random variable that embodies the uncertainty of the model and explains the differences between the prediction $\mathbf{x}_i^\top \mathbf{w}$ and the outcome $y_i$.

Let's assume that the error follows a Gaussian distribution, *i.e.* that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{{\varepsilon_i}^2}{2\sigma^2}\right)$$

We can measure the **likelihood** to have $y_i$ given $\mathbf{x}_i$. It is given by:

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = p(\varepsilon_i = \mathbf{x}_i^\top \mathbf{w} - y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}\right)$$

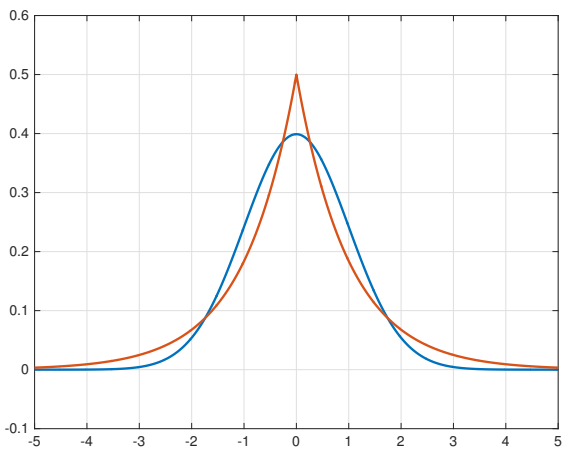


Figure 1: Propability Density Functions corresponding to a Gaussian and a Laplace error prediction distribution.

Assuming independence of the error terms $\epsilon_i$, the combined likelihood to have all outputs $\mathbf{y}$ given all data $\mathbf{X}$ is given by

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\mathbf{w}) &= \prod_{i=1}^{n} p(\varepsilon_i) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^{n}\frac{\left(\mathbf{x}_i^\top\mathbf{w}-y_i\right)^2}{2\sigma^2}\right) \end{aligned} \tag{1}$$

The **maximum likelihood** estimate $\hat{\mathbf{w}}_{ML}$ is simply the weight vector $\mathbf{w}$ that maximises the likelihood $p(\mathbf{y}|\mathbf{X},\mathbf{w})$:

$$\hat{\mathbf{w}}_{ML} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X},\mathbf{w})$$

A more practical, but equivalent, approach is to minimise the negative log likelihood:

$$\hat{\mathbf{w}}_{ML} = \arg\min_{\mathbf{w}} -\log\left(p(\mathbf{y}|\mathbf{X},\mathbf{w})\right) \tag{2}$$

$$= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top\mathbf{w}-y_i\right)^2 - n\log\left(\sqrt{2\pi\sigma^2}\right) \tag{3}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{n}\left(\mathbf{x}_i^\top\mathbf{w}-y_i\right)^2$$

Thus we've shown that the Least Square estimate is in fact the Maximum Likelihood solution if the error is assumed to be Gaussian.

Now, let's assume that the error follows a Laplace distribution: $\varepsilon_i \sim \text{Laplace}(0,\lambda)$.

$$p(\varepsilon_i) = \frac{1}{2\lambda}\exp\left(-\frac{|\varepsilon_i|}{\lambda}\right)$$

Assuming independence of the error terms $\epsilon_i$, the combined likelihood to have all outputs $\mathbf{y}$ given all data $\mathbf{X}$ is this time given by

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\mathbf{w}) &= \prod_{i=1}^{n} p(\varepsilon_i) \\ &= \left(\frac{1}{2\lambda}\right)^n \exp\left(-\frac{1}{\lambda}\sum_{i=1}^{n}|\mathbf{x}_i^\top\mathbf{w}-y_i|\right) \end{aligned} \tag{4}$$

From which we can derive that minimising the Mean Absolute Error (MAE) loss is identical

to finding the maximum likelihood solution, if the error follows a Laplace distribution.

Note that solving for the MAE loss is typically tricky. Convex optimisation techniques have been developed in the 2000s to solve for these kind of problems. The mathematics involved are beyond the scope of this module.

**Take Away.** The loss function is intimately related to the distribution of your errors. This can give us a way to check that we are using an appropriate loss function. Say you use Least Squares to find the Mean Square Error minimiser $\mathbf{w}_{MSE}$. If you compute the prediction errors for $\mathbf{w}_{MSE}$, you can then build a histogram of these errors and check that it is indeed close enough to a Gaussian distribution. If the error is far from Gaussian, it may be a good idea to use different loss function, or to go back to the dataset and remove any possible spurious outlier.