1. R^2 *or the log-likelihood?*

> **Info:** This post may be of interest to scholars and practitioners who have already used/heard about the R^2 metric and are familiar with maximum likelihood estimation.

R^2 is useful when your model outputs fixed uncertainty. But log-likelihood is broader —it can also deal with outputs with non fixed uncertainty!

Let's take a concrete example — this will help us show that log-likelihood is a broader measure than R^2. Imagine we have the following test set (data points we have hidden away to assess our model):

Index	Temperature (°C)	Date
m+1	6	12-11-02
m+2	4.3	13-11-02
m+3	5.7	14-11-02
m+4	6.7	15-11-02
m+5	3.9	16-11-02

which represent the factitious temperatures (°C) measured over five consecutive days in London in November 2002. $m+5$ is the size of our full dataset, m being the size of the training set and $m+1,\dots m+5$ the indices of the test set. For conciseness, we introduce some mathematical notation: take the recorded temperature values at date $t = m+1\dots m+5$ to be denoted by $y_t \in \mathbb{R}$; e.g, $y_{m+3} = 5.7$ is the temperature recorded on 14-11-02. Our prediction task is shown in figure 1.1, where we are given the pairs

$(t=1, y_1), \dots (t=m, y_m)$ as training data and we are asked to predict the question marks for $t = [m+1, \dots, m+5]$.

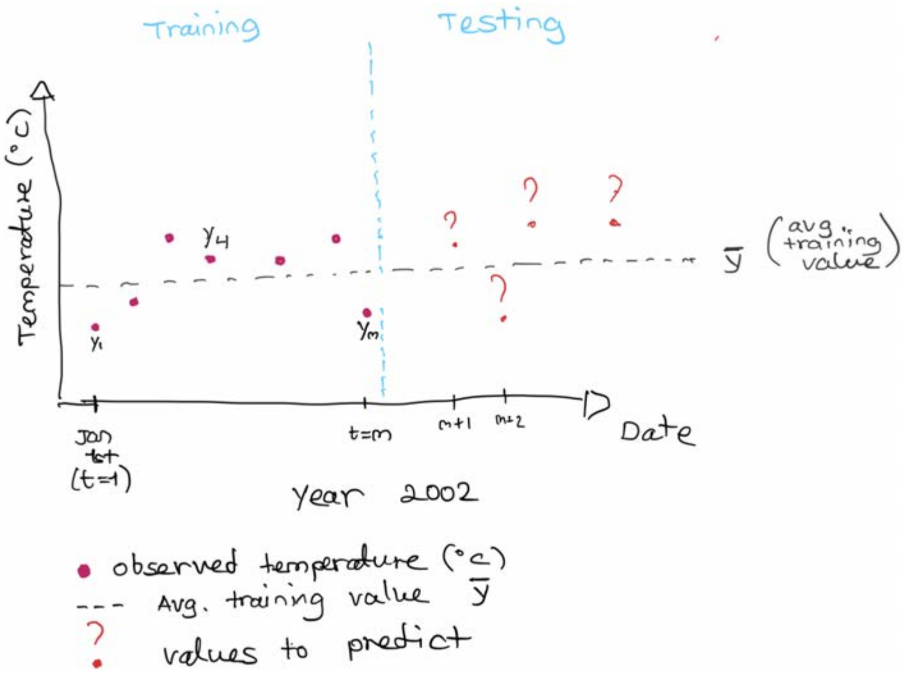

Figure 1.1

We have also constructed a model to predict those temperature values in the test set. Let's denote the predicted temperature values as $\hat{y}_t \in \mathbb{R}$. Constructing a table with the math notation and inserting the predicted values we have:

t	y_t	$\hat{y}_t$
m+1	6	5.5
m+2	7.3	5.9
m+3	4.1	5
m+4	5.7	6.1
m+5	5.9	5.9

and we can use the values in the table and just plug them in the formula for

MSE

R^2:

$$1-\frac{\boxed{\frac{1}{N}\sum_t (y_t-\hat{y}_t)^2}}{\frac{1}{N}\sum_t (y_t-\bar{y})^2}$$

where $\bar{y}$ is the average of all the recorded y_t values used for training our model. In figure 1.2 we can see the denominator terms in the R^2, and in figure 1.3 both the denominator and numerator terms.

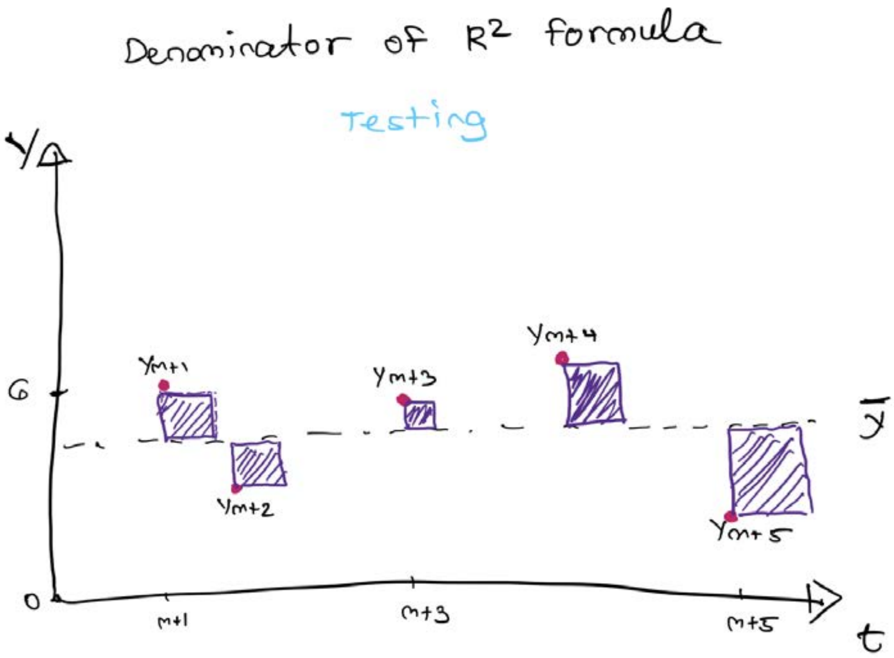

Figure 1.2

Let's explore the terms in the formula for R^2 through a probabilistic lens. Then, we can readily see what is R^2 useful for and what it is not. The numerator of the second term in the formula, namely $\sum_t \frac{1}{N}(y_t-\hat{y}_t)^2$, is the mean squared error (MSE). The MSE can be viewed as a recipe to assess errors, in which the square operation is placed to avoid errors canceling each other; or, it can be viewed probabilisticly as the result of minimising the negative log-likelihood of the variance term in the following model (call it M1):

$$\hat{y}_t \sim \text{Normal}(f(t), \sigma^2)$$

where we chose to model the mean temperature values $(f(t))$ based only on

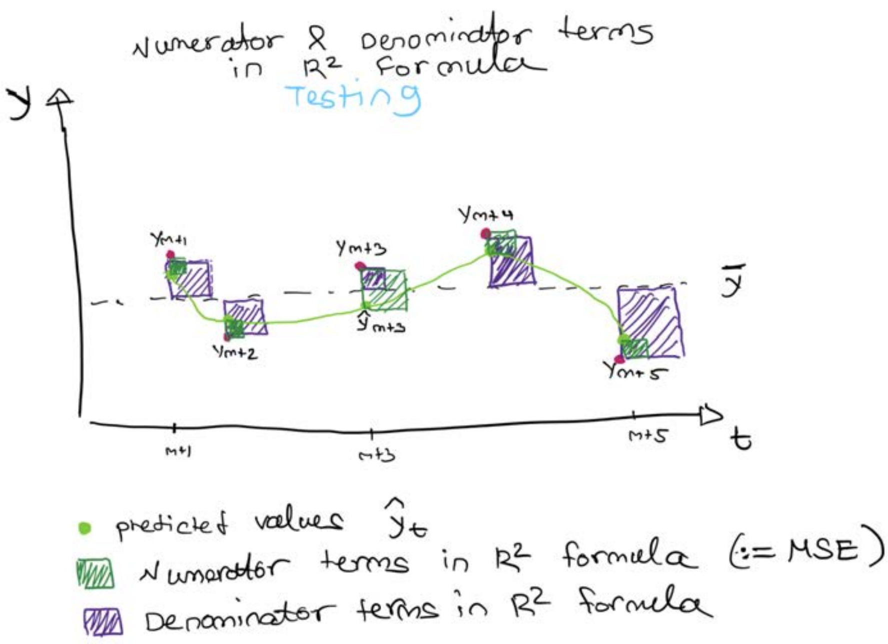

Figure 1.3

time t. This result is not specific to our model, but to any model with a normally distributed random variable with a fixed variance. So, for these models we can write that MSE $= \min_{\sigma^2}$ - Likelihood$(y_1, \dots, y_m|\text{M1})$ (hereafter Likelihood will be denoted with ℓ).

$$\min_{\sigma^2} - \left[\log \ell(y_1, \dots, y_m | M_1)\right] \underset{\text{likelihood def.}}{=} \min_{\sigma^2} - \left[\log \overbrace{p(y_1, \dots, y_m)}^{\text{joint dist. of training data}}\right] =$$

$$\underset{\substack{\forall i \neq j\ y_i \text{ indep.} \\ \text{of } y_j}}{=} \min_{\sigma} \log p(y_1) \cdot \dots \cdot p(y_m) \underset{\text{log. rules}}{=} \min_{\sigma} \log p(y_1) + \dots + \log p(y_m) =$$

$$\underset{\text{normal dist.}}{=} \min_{\sigma^2} - \left[\log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1 - f(1)}{\sigma}\right)^2} + \dots + \log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_m - f(m)}{\sigma}\right)^2}\right] =$$

$$\underset{\text{log rules}}{=} \min_{\sigma^2} - \left[\sum_{i=1}^{m} \log 1 - \log \sigma - \log \sqrt{2\pi} - \frac{1}{2}\left(\frac{y_i - f(i)}{\sigma}\right)^2\right].$$

In order to minimize σ^2, we take derivative and equate to zero. So,

$$-\sum_{i=1}^{m} \frac{1}{\sigma} - \frac{1}{2} \cdot \frac{-2\cancel{\sigma} \cdot (y_i - f(i))^2}{\sigma^{\cancel{4}3}} = 0 \underset{\text{if \& only if}}{\iff} \sum_{i=1}^{m} \frac{-\sigma^2 + (y_i - f(i))^2}{\sigma^3} = 0$$

$$\iff -m\sigma^2 = -\sum_{i=1}^{m} (y_i - f(i))^2 \iff \boxed{-\sigma^2 = -\frac{1}{m}\sum_{i=1}^{m} (y_i - f(i))^2 = -\text{MSE}}$$

In the same manner we can think of the denominator of the second term representing the model (call it M2)

$$\hat{y}_t \sim \text{Normal}(\bar{y}, \rho^2)$$

and hence calculating R^2 is the reciprocal value to the maximum likelihood ratio between the fixed variances (in our case σ and ρ) of normally distributed random variables. In other words

$$R^2 = 1 - \frac{\min_{\sigma^2} -\ell(y_1, \dots, y_m | M1)}{\min_{\rho^2} -\ell(y_1, \dots, y_m | M2)}$$

, and if you model your problem with a fixed variance, then R^2 is just another way to communicate the likelihood ratio. However, if you use a model with non fixed variance, for example $\hat{y}_t \sim \text{Normal}(f(t), \sigma(t)^2)$, evaluating it with R^2 will completely ignore the fact that you have a non fixed variance. An example of such a model is Gaussian Process. Instead, **you can just use the likelihood of your model, rather than just plugging the values into the R^2 formula**.