Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation

Afshin Gholamy[1], Vladik Kreinovich[2], and Olga Kosheleva[3]
[1]Department of Geological Sciences
[2]Department of Computer Science
[3]Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
afshingholamy@gmail.com, vladik@utep.edu
olgak@utep.edu

Abstract

When learning a dependence from data, to avoid overfitting, it is important to divide the data into the training set and the testing set. We first train our model on the training set, and then we use the data from the testing set to gauge the accuracy of the resulting model. Empirical studies show that the best results are obtained if we use 20-30% of the data for testing, and the remaining 70-80% of the data for training. In this paper, we provide a possible explanation for this empirical result.

1 Formulation of the Problem

Training a model: a general problem. In many practical situations, we have a model for a physical phenomenon, a model that includes several unknown parameters. These parameters need to be determined from the known observations; this determination is known as *training* the model.

Need to divide data into training set and testing set. In statistics in general, the more data points we use, the more accurate are the resulting estimates. From this viewpoint, it may seem that the best way to determine the parameters of the model is to use all the available data points in this determination. This is indeed a good idea if we are absolutely certain that our model adequately describes the corresponding phenomenon.

In practice, however, we are often not absolutely sure that the current model is indeed adequate. In such situations, if we simply use all the available data to determine the parameters of the model, we often get *overfitting* – when the model describes all the data perfectly well without being actually adequate. For example, if we observe some quantity x at n different moments of time, then it is always possible to find a polynomial $f(t) = a_0 + a_1 \cdot t + a_2 \cdot t^2 + \ldots + a_{n-1} \cdot t^{n-1}$ that will fit all the data points perfectly well — to find such a polynomial, it is sufficient to solve the corresponding system of n linear equations with n unknowns $a_0, \ldots, a_{n-1}$:

$$a_0 + a_1 \cdot t_1 + a_2 \cdot t_i^2 + \ldots + a_{n-1} \cdot t_i^{n-1}, \quad i = 1, \ldots, n.$$

This does not mean that the resulting model is adequate, i.e., that the resulting polynomial can be used to predict the values $x(t)$ for all t: one can easily show that if we start with noisy data, the resulting polynomial will be very different from the actual values of $x(t)$. For example, if $n = 1$ and the actual value of $x(t)$ is a constant, then, due to noise, the resulting polynomial $x(t) = a_0 + a_1 \cdot t$ will be a linear function with $a_1 \neq 0$. Thus, for large t, we will have $x(t) \to \infty$, so the predicted values will be very different from the actual (constant) value of the signal.

To avoid overfitting, it is recommended that we divide the observations into training and testing data:

- First, we use the training data to determine the parameters of the model.

- After that, we compare the model's predictions for all the testing data points with what we actually observed, and use this comparison to gauge the accuracy of our model.

Which proportion of data should we allocate for testing? Empirical analysis has shown that the best results are attained if we allocate 20-20% of the original data points for testing, and use the remaining 70-80% for training.

For this division, we get accuracy estimates which are:

- valid – in the sense that they do not overestimate the accuracy (i.e., do not underestimate the approximation error), and

- are the more accurate among the valid estimates – i.e., their overestimation of the approximation error is the smallest possible.

What we do in this paper. In this paper, we provide a possible explanation for this empirical fact.

2 Formal Description and Analysis of the Problem

Training and testing: towards a formal description. Our goal is to find the dependence of the resider quantity y on the corresponding inputs $x_1, \ldots, x_n$. To be more specific, we assume that the dependence has the form

$$y = f(a_1, \ldots, a_m, x_1, \ldots, x_n),$$

for some parameters $a_1, \ldots, a_m$. For example, we can assume that the dependence is linear, in which case $m = n + 1$ and

$$y = a_1 \cdot x_1 + \ldots + a_n \cdot x_n + a_{m+1}.$$

We can assume that the dependence is quadratic, or sinusoidal, etc.

To find this dependence, we use the available data, i.e., i.e., we use N situations $k = 1, \ldots, n$ in each of which we know both the values of the inputs $x_1^{(k)}, \ldots, x_n^{(k)}$ and the corresponding output $y^{(k)}$.

Let p denote the fraction of the data that goes into the training set. This means that out of the original N patterns $\left(x_1^{(k)}, \ldots, x_n^{(k)}, y^{(k)}\right)$:

- $N \cdot p$ patterns form a training set, and
- the remaining $(1 - p) \cdot N$ patterns form a testing set.

We use the training set to find estimates $\widehat{a}_1, \ldots, \widehat{a}_m$ of the parameters $a_1, \ldots, a_m$. Then, for each pattern $\left(x_1^{(k)}, \ldots, x_n^{(k)}, y^{(k)}\right)$ from the testing set, we compare the desired output $y^{(k)}$ with the result

$$\widehat{y}^{(k)} = f\left(\widehat{a}_1, \ldots, \widehat{a}_m, x_1^{(k)}, \ldots, x_n^{(k)}\right)$$

of applying the trained model to the inputs. Based on the differences

$$d_k \stackrel{\text{def}}{=} y^{(k)} - \widehat{y}^{(k)},$$

we gauge the accuracy of the trained model.

How do we gauge the accuracy of the model. Many different factors influence the fact that the resulting model is not perfect, such as measurement errors, approximate character of the model itself, etc.

It is known that under reasonable assumptions, the distribution of a joint effect of many independent factors s close to Gaussian (normal) – the corresponding mathematical result is known as the *Central Limit Theorem*; see, e.g., [1]. Thus, we can safely assume that the differences d_k are normally distributed.

It is known that a 1-D normal distribution is uniquely determined by two parameters: mean value μ and standard deviation σ. Thus, based on the differences d_k, we can estimate:

- the mean value (bias) of the trained model, and
- the standard deviation σ describing the accuracy of the trained model.

A general fact from statistics: reminder. In statistics, it is known that when we use M values to estimate a parameter, the standard deviation of the estimate decreases by a factor of $\sqrt{M}$.

Example. The factor-of-$\sqrt{M}$ decrease is the easiest to explain on the simplest example when have a single quantity q, and we perform several measurements of this quantity by using a measuring instrument for which the standard deviation of the measurement error is σ_0. As a result, we get M measurement results $q_1,\ldots,q_M$. As an estimate for q, it is reasonable to take the arithmetic mean

$$\widehat{q}=\frac{q_1+\ldots+q_M}{M}.$$

Then, the resulting estimation error $\widehat{q}-q$, i.e., the difference between this estimate and the actual (unknown) value q of the quantity of interest has the form

$$\widehat{q}-q=\frac{q_1+\ldots+q_M}{M}-q=\frac{(q_1-q)+\ldots+(q_M-q)}{M}.$$

By definition, for each difference q_i-q, the standard deviation is equal to σ_0. and thus, the variance is equal to σ_0^2.

Measurement errors corresponding to different measurements are usually independent. It is known that the variance of the sum of independent random variables is equal to the sum of the variances. Thus, the variance of the sum $(q_1-q)+\ldots+(q_M-q)$ is equal to $M\cdot\sigma_0^2$, and the corresponding standard deviation is equal to $\sqrt{M\cdot\sigma_0^2}=\sqrt{M}\cdot\sigma_0$. When we divide the sum by M, the standard deviation also divides by the same factor. So, the standard deviation of the difference $\widehat{q}-q$ is equal to $\dfrac{\sqrt{M}\cdot\sigma_0}{M}=\dfrac{\sigma_0}{\sqrt{M}}$.

Let us use the general fact from statistics. We estimate the parameters of the model based on the training set, with $p\cdot N$ elements. Thus, the standard deviation of the corresponding model is proportional to $\dfrac{1}{\sqrt{p\cdot N}}$.

When we gauge the accuracy of the model, we compare the trained model with the data from the testing set. Even if the trained model was exact, because of the measurement errors, we would not get the exact match. Instead, based on $(1-p)\cdot N$ measurements, we would get the standard deviation proportional to $\dfrac{1}{\sqrt{(1-p)\cdot N}}$.

We want to estimate the difference d_k between the trained model and the testing data. It is reasonable to assume that, in general, the errors corresponding to the training set and to the testing set are independent – we may get positive correlation in some cases, negative correlation in others, so, on average, the correlation is 0. For independence random variables, the variance is equal to the sum of the variances. Thus, on average, this variance is proportional to

$$\left(\frac{1}{\sqrt{p\cdot N}}\right)^2+\left(\frac{1}{\sqrt{(1-p)\cdot N}}\right)^2=\frac{1}{p\cdot N}+\frac{1}{(1-p)\cdot N}=\frac{1}{(p\cdot(1-p))\cdot N}.$$

Thus, to get the smallest possible estimate for the approximation error, then, out of all possible values p, we need to select the value p for which the product $p\cdot(1-p)$ is the largest possible.

Which values p are possible? The only remaining question is now: which values p are possible?

Our requirement was that we should select p for which the gauged accuracy is guaranteed not to overestimate the accuracy. In precise terms, this means that the standard deviation of the trained model – i.e., the standard deviation of the estimate $\widehat{y}^{(k)}$ – should be smaller than or equal to the standard deviation of the difference d_k by which we gauge the model's accuracy:

$$\sigma\left[\widehat{y}^{(k)}\right] \le \sigma[d_k].$$

Here, $d_k = \widehat{y}^{(k)} - y^{(k)}$ is the difference between:

- the estimate $\widehat{y}^{(k)}$ whose inaccuracy is cased by the measurement errors of the training set and
- the value $y^{(k)}$ whose inaccuracy is cased by the measurement errors of the testing set.

So, we must have

$$\sigma\left[\widehat{y}^{(k)}\right] \le \sigma\left[\widehat{y}^{(k)} - y^{(k)}\right].$$

In general, for two random variables r_1 and r_2 with standard deviations $\sigma[r_1]$ and $\sigma[r_2]$, the smallest possible value of the standard deviation of the difference is $|\sigma[r_1 -]\sigma[r_2]|$ (see, e.g., [1]):

$$\sigma[r_1 - r_2] \ge |\sigma[r_1] - \sigma[r_2]|.$$

In particular, for the difference $d_k = \widehat{y}^{(k)} - y^{(k)}$, the smallest possible value of its standard deviation $\sigma\left[\widehat{y}^{(k)} - y^{(k)}\right]$ is

$$\left|\sigma\left[\widehat{y}^{(k)}\right] - \sigma\left[y^{(k)}\right]\right|.$$

Thus, to make sure that we do not underestimate the measurement error, we must guarantee that

$$\sigma\left[\widehat{y}^{(k)}\right] \le \left|\sigma\left[\widehat{y}^{(k)}\right] - \sigma\left[y^{(k)}\right]\right|,$$

i.e., that $a \le |a - b|$, where we denoted $a \stackrel{\text{def}}{=} \sigma\left[\widehat{y}^{(k)}\right]$ and $b \stackrel{\text{def}}{=} \sigma\left[y^{(k)}\right]$.

In principle, we can have two different cases: $a \le b$ and $b \le a$. Let us consider these two cases one by one.

- If $a \ge b$, then the desired inequality takes the form $a \le a - b$, which for $b > 0$ is impossible.
- Thus, we must have $b \le a$. In this case, the above inequality takes the form $a \le b - a$, i.e., equivalently, $2a \le b$.

Thus, we must have

$$2\sigma\left[\widehat{y}^{(k)}\right] \le \sigma\left[y^{(k)}\right].$$

Since the inaccuracy of the estimate $\widehat{y}^{(k)}$ comes only from measurement errors of the training set, with $p \cdot N$ elements, we have

$$\sigma\left[\widehat{y}^{(k)}\right] = \frac{\sigma_0}{\sqrt{p \cdot N}}$$

for some σ_0. Similarly, since the inaccuracy of the estimate $y^{(k)}$ comes only from measurement errors of the testing set, with $(1-p) \cdot N$ elements, we have

$$\sigma\left[y^{(k)}\right] = \frac{\sigma_0}{\sqrt{(1-p) \cdot N}}.$$

Thus, the above inequality takes the form

$$2 \cdot \frac{\sigma_0}{\sqrt{p \cdot N}} \le \frac{\sigma_0}{\sqrt{(1-p) \cdot N}}.$$

Dividing both sides of this inequality by σ_0 and multiplying by $\sqrt{N}$, we conclude that

$$\frac{2}{\sqrt{p}} \le \frac{1}{\sqrt{1-p}}.$$

Squaring both sides, we get

$$\frac{4}{p} \le \frac{1}{1-p}.$$

By bringing both sides to the common denomination, we get $4 - 4p \le p$, i.e., $4 \le 4p + p = 5p$ and $p \ge 0.8$.

Thus, to make sure that our estimates do not overestimate accuracy, we need to select the values $p \ge 0.8$.

Towards the final conclusion. As we have mentioned earlier, out of all possible values p, we need to select a pone for which the product $p \cdot (1-p)$ is the largest possible. For $p \ge 0.8$, the function $p \cdot (1-p)$ is decreasing. Thus, its largest values is attained when the value p is the smallest possible – i.e., when $p = 0.8$.

So, we have indeed explained why $p \approx 80\%$ is empirically the best division into the training and the testing sets.

Acknowledgments

This work was supported in part by the US National Science Foundation grant HRD-1242122.

References

[1] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.