Towards Predictive Statistics: A Pedagogical Explanation

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
vladik@utep.edu

Abstract

In statistics application area, lately, several publications appeared that warn about the dangers of the inappropriate application of statistics and remind the users of the recall that prediction is the ultimate objective of the statistical analysis. This trend is known as *predictive statistics*. However, while the intended message is aimed at the very general audience of practitioners and researchers who apply statistics, many of these papers are not easy to read since they are either too technical and/or too philosophical for the general reader. In this short paper, we describe the main ideas and recommendation of predictive statistics in – hopefully – clear terms.

1 Limitations of the Traditional Statistics and Need for Predictive Approach

Prediction is important. One of the main objectives of science and engineering is to predict future events – i.e., to predict what will happen in general, and, specifically, predict what will happen if we undertake a certain action.

Prediction in science and engineering. From this viewpoint, the progress of science and engineering is usually made as follows:

- we analyze the existing data, and come up with formulas connecting different quantities,
- we then use these formulas to predict new phenomena and/or future values of different quantities.

If the prediction is successful, i.e., if the observed future values are indeed close to the predictions, then the theory is confirmed.

This is how Mendeleev's periodic table became an accepted theory: when Mendeleev used the observed periodicity to predict several new elements, and

these elements were actually observed. This is how Einstein's General Relativity Theory became an accepted theory: when this theory predicted how the path of light will be curved when the light ray goes near the Sun, and these predictions were experimentally confirmed.

What can we do if it is not feasible to set new experiments. In some practical situations, it is not feasible to perform new experiments, we have to deal with whatever data we have. In such situations, a reasonable idea is to divide all the observed data into the training set and the testing set.

We then use apply some methodology to the training set to find the dependence between the corresponding quantities. If this same dependence holds for the testing set as well, this means that we have indeed uncovered a meaningful law.

This process simulates real predictions: at the moment of time when we only had training data, we would be able to uncover this dependence, and then test in on the future testing data. This analogy shows that for temporal data, it is a good idea to select some moment of time, so that:

- all the observations made before this moment of time are selected as the training set, while
- all the observations made after this moment of time are selected as the testing set.

This is a usual procedure, e.g., in machine learning; see, e.g., [1].

Usual statistical approach is different. The usual statistical approach to data processing is different. Instead of dividing the data into two sets, practitioners usually consider the whole data, and try to see if they can extract some dependence from this data.

How do we know that the resulting dependence can help with prediction? How do we know that his a meaningful dependence, a dependence that can help us with predictions?

In practice, we only have finitely many observations, so we can always fit a polynomial of appropriate order that will match all these observations exactly – it does not mean, of course, that this same dependence will hold for any new data. This phenomenon is known as *over-fitting*.

In general, the more parameters we use for our model, the better it we can fit with the data. We therefore need to balance the quality of a fit with the number of parameters. There exist several methods for describing this balance such as AIC, BIC, etc. The problem is that different methods sometimes lead to different results, and it is not clear which of these methods we should use.

2 Predictive Approach: Ideal Statistical Case

Alternative idea: use predictive statistics. A natural alternative idea is to apply the same idea which is successfully used in science and engineering (and in machine learning);

- divide the observations into training set and testing set,
- apply statistical techniques to the training set to find the dependence between the corresponding quantities, and then
- check, on observations from the testing set, whether these dependence holds; as a criterion for the quality of fit, we can use, e.g., the mean square deviation between the observed values and the model's predictions.

In this case, a natural criterion is how accurately the dependence holds on the testing set.

Additional advantages of predictive approach to statistics. There is no need to have criteria like AIC: if we use too many parameters in the model determined based on the training set, it will simply not fit with the testing set.

For example, let us assume that the actual dependence between quantities x_i and y_i is linear: $y_i = a_0 + a_1 \cdot x_i + \varepsilon_i$, for some noise ε_i. Then, if we apply the Least Squares technique to estimate a_0 and a_1 based on the training set, the resulting estimates $\widehat{a}_0$ and $\widehat{a}_1$ will be close to a_0 and a_1, and thus, for each testing pair (x_j, y_j) the estimated linear dependence $y_j = \widehat{a}_0 + \widehat{a}_1 \cdot x_j$ will be a good fit.

However, if we find a high-order polynomial $y = a_0 + a_1 \cdot x + \ldots + a_n \cdot x^n$ that fits exactly all the observations from the training set, we do not expect that this polynomial formula will be any good in describing observations from the testing set.

This, in a nutshell, is what is proposed in predictive statistics; see, e.g., [2, 3, 6] To check how well the dependence uncovered on the training set fits the testing set, we can still use p-values and other traditional techniques, the question is not so much which criteria to use, but rather what paradigm to use: instead of determining the model based on all available observations, the idea is:

- to determine the model based on some of the observations, and then
- to verify this model by checking how well it fits the remaining observations.

3 General Case: Calibration Approach

Need to go beyond simple predictive statistics. It would be great if we had a model that perfectly fits the training data – and then statistically significantly fits the testing data. In some cases, we have such models. However, in many practical situations, no such models is available.

Such a situation is typical in science and engineering: many real-life processes are very complex, no models has a perfect fit with observations. In such situations, what scientists do is come up with a model which – at least for the quantities of interest – provide a better fits than previous known models. In economic applications, this approach is known as *calibration*; see, e.g., [4, 5].

What is calibration: a brief description. We start with a training data, and we use some methodology to find a model which provides, for quantities of

interest, a better fit that previously known models. This model can be obtained, e.g., by equating relevant parameters of the model and empirical estimates for these parameters: e.g., by comparing moments, or – if we are predicting economic cycles, the average length of the cycle, etc.

As a criterion for the quality of fit, we can use, e.g., the mean square deviation between the observed values and the model's predictions.

Once a model is formulated, it is then tested on the testing data. If on the testing data, the model also provides a better fit that previous known models (or if no model previously explained the phenomenon at all), then this model indeed provides a new insight into the analyzed phenomenon.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] L. Brieman, "Statistical modeling: the two cultures", *Statistical Science*, 2001, Vol. 16, No. 3, pp. 199–231.

[3] W. Broggs, *Uncertainty: The Soul of Modeling, Probability, and Statistics*, Springer Verlag, Cham., Switrzerland, 2016.

[4] Th. F. Cooley, "Calibtade models", *Oxford Review of Economic Policy*, 1997, Vol. 13, No. 3, pp. 55–69.

[5] D. N. DeJong and C. Dave, *Structural Macroeconometrics*, Princeton University Press, Princeton, New Jersey, and Oxford, UK, 2007.

[6] G. Shmueli, "To explain or to predict?", *Statistical Science*, 2010, Vol. 25, No. 3, pp. 289–310.