WHITE PAPER: VALIDATING A FORECASTING NEURAL NETWORK

by Suzanne M. Rodriguez

According to my not-so-new edition of Webster's New Collegiate Dictionary, one definition of valid is "appropriate to the end in view." Establishing whether a neural network lives up to this simple definition, however, can be quite challenging. It may surprise (and relieve) you to know that the challenge is not primarily a statistical one. As discussed briefly later in this White Paper, the techniques presented in an introductory-level statistics text are sufficient to validate a forecasting neural network. The true challenge arises in the deceptively simple phrase "appropriate to the end in view."

Let's start by considering the definition of "appropriate." A forecast is appropriate when it is reasonably correct. Statistics can be used to analyze forecast errors, but first you will have to decide how to measure forecast error. For example, if a neural network is to be used for sales forecasting, we could measure forecast error as the number of units difference between a forecast and actual sales. One possible problem with this measurement is that it doesn't distinguish a 100 unit forecast error on unit sales of 200 from the same 100 unit forecast error on sales of 2,000. Both errors appear equally important. But if sales continue through the next forecasting period at similar rates, the 100 unit overstock in the first case will remain in inventory far longer than the 100 unit overstock in the second case.

In order to consider the relative magnitude of an error instead of its absolute size, we could measure forecast error as a percentage of the actual sales. From this standpoint, the forecast error of 100 units on sales of 200 units is a 50% forecast error. The same forecast error of 100 units on sales of 2,000 units is only a 5% forecast error. Now the first error appears more serious than the second.

Well, that certainly paints a different picture, but is it the right picture? What if the item we have been discussing with sales of 2,000 units actually costs 50 times more than the slower selling item? Measuring the impact of each forecast error on inventory carrying costs, we might now consider the forecast error of 100 units on sales of 2,000 units more serious than the forecast error of 100 units on sales of 2,000 units.

These are certainly not the only scenarios we could consider, but they are sufficient to illustrate the need for careful consideration of alternative measures of neural network performance. If you aren't careful, you'll get what you measure. The good news is that you don't have to choose just one performance measure. In fact, using two or more performance measures has definite advantages. Each measure provides a unique perspective on neural network performance which can reveal strengths and weaknesses which would not be apparent using a different measure of performance.

Once you have determined "appropriate" measures of neural network performance, you are ready to begin the statistical analysis by computing basic summary statistics such as the mean, standard deviation and range of forecast errors. The information needed for decision-making, however, can only be produced by considering these statistics with respect to "the end in view." Is an average forecast error of 100 units good or bad? The end in view determines the standard that will be applied to evaluate neural network performance.

The least useful performance benchmark is a nice round number pulled out of thin air. The problem with using an abstract benchmark is that it is not possible to predict with accuracy the inherent difficulty of forecasting in a specific context. Forecasting sales of a newly introduced "fad item" like a big purple dinosaur is inherently more difficult than forecasting sales of toothpaste. An abstract benchmark can kill a project if it sets an unrealistically high performance standard. When a forecasting technique fails to meet the standard, no one can establish to what extent the observed error levels reflect the amount of "noise" in the forecasting environment as opposed to a weakness in the forecasting technique.

The most useful performance benchmark is a set of forecasts produced by a currently used and/or competing forecasting method (see the March/April 1993 issue of PC AI magazine for a discussion of linear regression models as benchmarks). The error levels associated with a set of alternative forecasts provide a realistic frame of reference as to the difficulty of the forecasting problem. A set of alternative forecasts can be used to test whether, for example, the average forecasting error is significantly lower using a neural network. Inferential statistics can also be used to establish a confidence interval which describes the likely difference between two forecasting methods. One of the side benefits of the validation process can be providing information about existing methods so that abstract benchmarks are no longer the only benchmarks available.

Statistical tests can be used to indicate whether the neural network forecasts are better, the same, or worse than comparison forecasts. The "end in view" determines how the neural network performance needs to stack up against the benchmark. For example, if a sales forecasting neural network is being considered as a replacement for buyers' forecasts, you might demand that the network perform better than the buyers by at least some specified margin. If the neural network is being considered as an alternative to hiring and training new buyers, network performance equivalent to the buyers might be acceptable.

The validation process can provide a wealth of information beyond whether or not the neural network's overall performance is appropriate to the "end in view." What you learn from the network's mistakes can be just as important as what you learn about overall performance. Exploratory analysis of forecast errors can reveal patterns which allow you to improve the network design or specify the situations in which forecasts are weak. For example, the neural network may forecast better than a buyer for low and moderately-priced products while forecasting worse than a buyer for high-priced products. Other than suggesting possible redesign of the network with respect to the product pricing dimension, this result may suggest that you modify "the end in view." What if the neural network provided forecasts for low- and moderately-priced products so that buyers could restrict their attention to the high-priced products? A successful validation process not only suggests whether the neural network is appropriate to the end in view, but also how the end in view might be modified in light of the network's strengths, weaknesses and performance levels.

If all of this sounds like it requires some time and advance planning - it does. Validation data should be identified at the same time as training data. Keep in mind that statistics, like neural networks, gain strength in numbers. The project budget for a neural network should include time for planning and conducting the validation and reporting the results. At least one person with solid, basic statistics skills will be required to lead the validation effort. The statistical analysis can be performed on a spreadsheet, but I wouldn't recommend using a spreadsheet for validation (unless the person responsible for the calculations has done something particularly loathsome to you recently). A statistical package which provides support for analyzing outliers and goodness of fit is useful in determining the various ways in which your data set "misbehaves" (and it will) prior to performing the analysis.

If you expect magic to happen as soon as the statistical package arrives, you will be disappointed. The formula for a successful validation combines statistical knowledge, a statistical package, and most importantly, information about how to measure performance in the context of your specific project. Only then will you be able to establish whether a neural network is "appropriate to the end in view."