The Complexity of Forecast Testing

LANCE FORTNOW
and
RAKESH V. VOHRA
Northwestern University

Consider a weather forecaster predicting the probability of rain for the next day. We consider tests that given a finite sequence of forecast predictions and outcomes will either pass or fail the forecaster. It is known that any test which passes a forecaster who knows the distribution of nature can also be probabilistically passed by a forecaster with no knowledge of future events. This note summarizes and examines the computational complexity of such forecasters.

Categories and Subject Descriptors: J.4 [**Computer Applications**]: Social and Behavior Sciences—*Economics*

General Terms: Economics, Measurement, Theory, Verification

Additional Key Words and Phrases: Forecast Testing, Prediction, Bounded Rationality

1. INTRODUCTION

Suppose one is asked to forecast the probability of rain on successive days. In the absence of any knowledge about the distribution that governs the change in weather, how should one measure the accuracy of the forecast? Clearly, the question is relevant not just for weather forecasting, but to any kind of probability forecast. For example, the probability forecasts one obtains from prediction markets.

A popular and well studied criterion for judging the effectiveness of a probability forecast is called calibration. Dawid (1982) offers the following intuitive definition of calibration:

> Suppose that, in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of

To appear in *Econometrica*. It was presented at the 9th ACM Conference on Electronic Commerce.

Lance Fortnow, Department of Electrical Engineering and Computer Science, Northwestern University, Evanston IL 60208.
Rakesh Vohra, Department of Managerial Economics and Decision Sciences, Kellogg Graduate School of Management, Northwestern University, Evanston IL 60208. Research supported in part by NSF grant ITR IIS-0121678.

> precipitation was, say, close to some given value ω and (assuming these form an infinite sequence) determine the long run proportion p of such days on which the forecast event (rain) in fact occurred. The plot of p against ω is termed the forecaster's *empirical calibration curve*. If the curve is the diagonal $p = \omega$, the forecaster may be termed (empirically) *well calibrated*.

If a forecaster knows the actual probability distribution that governs the weather, then by reporting the correct conditional probability of rain each day, she will be well calibrated. On the other hand, as described in Foster and Vohra (1993), it is possible to be well calibrated without any knowledge of the weather at all. Specifically, they describe a randomized forecasting algorithm that with no knowledge whatsoever of the distribution that governs the weather, will with high probability generate forecasts that are well calibrated. Thus, a forecaster with no meteorological knowledge would be indistinguishable from one who knew the distribution that governs the change in weather.[1] In one sense this result is unsurprising because calibration is a weak criterion. Different forecasts can be well calibrated with respect to the same set of data. For example, suppose it rains on alternate days, i.e., wet, dry, wet, dry, etc. A forecast of 50 % probability of rain on each day would be well calibrated. So, would a forecast of 100% chance of rain on the first day, 0 % chance of rain on the second day, 100 % chance of rain on the third day, etc. Thus, if we interpet the result of Foster and Vohra (1993) as criticizing calibration as a test to evaluate probability forecasts, what test should one use?

Rather than generate a list of possible tests, let us identify properties that one would want a test of a probability forecast to satisfy. Formally, a test takes as input a forecasting algorithm, a sequence of outcomes and after some period accepts the forecast (PASS) or rejects it (FAIL). Sandroni (2003) proposed two properties that such a test should have. The first is that the test should declare PASS/FAIL after a finite number of periods. This seems unavoidable for a practical test. Second, suppose the forecast is indeed correct i.e., accurately gives the probability of nature in each round. Then, the test should declare PASS with high probability. We call this second condition "passing the truth." Call a test that satisfies these two conditions a *good* test. A test based on calibration is an example of a good test. Perhaps, there are good tests that cannot be 'gamed' in the way a calibration test can. Formally, a forecaster with no knowledge of the underlying distribution that can pass a good test with high probability on all sequences of data is said to have *ignorantly* passed the test. Since this randomized forecast can pass the test for all distributions it must be independent of the underlying (if any) distribution being forecasted. Hence, in some sense, this forecast provides no information at all about the process being forecasted.

[1] Lehrer (1997), Sandroni, Smorodinsky and Vohra (2003) and Vovk and Shafer (2005) give generalizations of this result.

Remarkably, Sandroni (2003) showed that for every good test there exists a randomized forecasting algorithm that will ignorantly pass the test. Therefore, no good test can distinguish between a forecaster who knows the underlying distribution of the process being forecast from one who simply 'games' the test. To see what drives this result, consider a good test that must decide PASS/FAIL after one observation only. Since the test must pass the truth, with one observation alone it must pass every forecast! The jump from one observation to a finite number of observations is not large because there are an infinite number of scenarios that agree on the first n observations.

How then is one to get around the impossibility result of Sandroni (2003)? Dekel and Feinberg (2006) as well as Olszewski and Sandroni (2007) do so by relaxing one of the properties of a good test. For example, allowing the test to declare PASS/FAIL at 'infinity', allowing the test to declare FAIL in a finite number of periods but PASS 'at infinity' or relaxing the condition that the test always passes the truth. These tests can often be made efficient in the sense that they can run in time linear in the length of the current sequence but the number of forecasts before a bad forecaster is failed could be extremely large as a function of the forecaster.

Olszewski and Sandroni (2008) have noted that the tests considered by Dekel and Feinberg (2006) and Olszewski and Sandroni (2007) rely on counterfactual information. Specifically, the test can use the predictions the forecast would have made along sequences that did not materialize because the test has access to the forecasting algorithm itself. As noted by Olszewski and Sandroni (2008) this is at variance with practice. For this reason they consider tests that are not permitted to make use of counterfactual predictions on the part of the forecaster but relax the condition that the test must decide in finite time. Formally, two different forecasting algorithms that produce the same forecast on a realization must be treated in the same way. If such tests pass the truth with high probability they show that for each such test, there is a forecasting algorithm that can ignorantly pass the test.

Al-Najjar, Sandroni, Smorodinsky and Weinstein (2008), take issue with the assumption that the test has no prior whatsoever of the underlying distribution. So, they examine what happens if the test knows that the distribution is drawn from some suitably rich class of distributions. Can this prior knowledge be used to construct a test that cannot be ignorantly passed? Their paper shows that yes, this is the case. Essentially the authors have relaxed the condition that the test must always pass the truth. So, if the forecaster makes a forecast inconsistent with a distribution from this class, the forecaster is failed. Otherwise, the forecast is evaluated according to the test proposed in Al-Najjar, Sandroni, Smorodinsky and Weinstein (2008).

It is natural to ask if a test, using a proper scoring rule[2] like *log-loss*, can cir-

[2]Assuming the forecaster is compensated on the basis of the scores associated with the rule, a proper scoring rule gives the forecaster the incentive to reveal his/her true beliefs. See Good (1952).

cumvent these difficulties. Here one penalizes the forecaster $\log p$ if the forecaster predicts a probability p of rain and it rains and a penalty of $\log(1-p)$ if it doesn't rain. The lowest possible score that can be obtained is the long-run average entropy of the distribution. One could imagine the test passing the forecaster if the log loss matches the entropy. However, such a test would need to know the entropy of the distribution. As noted in the introduction, we are concerned with tests which operate without any prior knowledge of the distribution. Proper scoring rules are good methods to compare two forecasters but are not useful for testing the validity of a single forecaster against an unknown distribution of nature.

2. COMPUTATIONALLY-BOUNDED FORECASTERS

Our paper approaches these questions by examining the consequences of imposing computational limits on both forecaster and the test. We measure the complexity as a function of the length of the history so far.

Most practical tests have a complexity that is polynomial in the length of the history, so it seems reasonable to restrict attention to good tests that have a complexity that is polynomial in the length of the history. Restricting the test in this way, should make it 'easier' to be ignorantly passed. It seems natural to conjecture that for every polynomial time good test, there exists a polynomial time randomized forecasting algorithm that will ignorantly pass the test. However, as we show, this is not the case. We exhibit a good linear time test that would require the forecaster to factor numbers under a specific distribution, or fail the test. The existence of an efficient (i.e. probabilistic polynomial time) algorithm for factoring composite numbers is considered unlikely. Indeed, many commercial available cryptographic schemes are based on just this premise. This result suggests that the 'ignorant' forecaster of Sandroni (2003) must have a complexity at least exponential in n. Hence, the 'ignorant' forecaster must be significantly more complex than the test. In particular its complexity may depend on the complexity of nature's distribution.

To prove this result, we interpret the observed sequence of 0-1's as encoding a number followed by a list of its possible factors. A sequence that correctly encodes a list of factors is called correct. The test fails any forecaster that does not assign high probability to these correct sequences when they are realized. Consider now the distribution that puts most of its weight on correct sequences. If the forecaster can ignorantly pass the test, it must be able to identify sequences that correspond to correct answers to our computational question. We also create an efficiently samplable distribution that no forecaster can ignorantly pass without being able to factor quickly on average.

The factoring proof does not generalize to all NP search problems, because we need a unique witness (solution) in order to guarantee that the test always passes the truth. Witness reduction techniques like Valiant-Vazirani (1986) don't appear to help.

Our second result strengthens the previous one by exhibiting a good test that re-

quires the forecaster to solve PSPACE-hard problems by building on the structure of specific interactive proof systems. In both cases the tests are deterministic. Furthermore, they use only the realized outcomes and forecasts to render judgement.

In addition we consider the possibility that the test may have more computational power than the forecaster. If we restrict ourselves to forecasters using time $O(t(n))$ there is a test T using time $O(n^{O(1)}t(n))$ with the following properties.

(1) For all distributions of nature μ, T will pass, with high probability, a forecaster forecasting μ.

(2) For some distribution τ of nature, for every forecaster F running in time $O(t(n))$, T will fail F with high probability.

If one takes a highly non-computable distribution τ, a forecaster would not be able to forecast τ well, but T could not test this in general if T is also required to always pass the truth. In a nutshell, no forecasting algorithm can ignorantly pass a good test that is more complex than itself.

REFERENCES

Al-Najjar, N., A. Sandroni, R. Smorodinsky and J. Weinstein (2008) "Testing Theories with Learnable and Predictive Representations," manuscript, Northwestern University.

Dawid, A.P. (1982) "The Well Calibrated Bayesian," *Journal of the American Statistical Association*, **77**, 379, 605–613.

Dekel, E. and Y. Feinberg (2006) "Non-Bayesian testing of a Stochastic Prediction," *Review of Economic Studies*, **73**, 893-906.

Foster, D.P. and R.V. Vohra (1993) "Asymptotic Calibration," *Biometrika*, 85-2, 379-390.

Good, I. J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society*, Series B 14, 107-114.

Lehrer, E. (2001) "Any Inspection Rule is Manipulable," *Econometrica,* **69**-5, 1333-1347.

Olszewski, W. and A. Sandroni (2007) "A Non-Manipulable Test", *Annals of Statistics*, forthcoming.

Olszewski, W. and A. Sandroni (2008) "Manipulability of Future Independent Tests ", *Econometrica*, forthcoming.

Sandroni, A. (2003) "The Reproducible Properties of Correct Forecasts," *International Journal of Game Theory*, 32-1, 151-159.

Sandroni, A., R. Smorodinsky and R. Vohra (2003) "Calibration with Many Checking Rules," *Mathematics of Operations Research* 28-1, 141-153.

L. Valiant and V. Vazirani. (1986) "NP is as easy as detecting unique solutions," *Theoretical Computer Science* 47(1), 85-93.

V. Vovk and G. Shafer. (2005) "Good sequential probability forecasting is always possible," *Journal of the Royal Statistical Society: Series B* **67**-5, 747-763.