Performance Evaluation Measures

Caren Marzban

Applied Physics Lab. and Department of Statistics
Univ. of Washington, Seattle, WA, USA 98195

Abstract

We begin by discussing some general issues in performance evaluation (a.k.a. verification). The discussion specializes to specific measures of the quality of forecasts, followed by a review of the behavior of these measures in rare-event (extreme?) situations. Some old and new statistical models proposed for understanding that behavior are examined. The talk ends by providing specific suggestions.

Generalities

This is mostly a review, and so, incomplete.
A lot of generalities, and so, a lot of talking.

Murphy (1993): performance is multifaceted - Consistency, Quality, Value.

Quality itself is multifaceted: Bias, association, accuracy, skill, reliability, resolution, sharpness, discrimination, and uncertainty.

Verification FAQ, Jolliffe and Stephenson (2003), and Casati, et al. (2008) give lots of reasons for evaluating performance.

My lesson 1: Do not evaluate!

Because there are too many issues:
- Explaining/diagnosing, but not much we can do to fix errors.
- Model comparison/selection, but:
- Model A better than B in terms of measure X but not Y.
- Model A better than B in Summer but not Winter.
- Model A better than B in East but not West.
- Measure may be inequitable, inducing hedging.
- Model may be an overfit, i.e. good training err but poor prediction err.
- Even on homogeneous data, measure is subject to sampling variability.
- Even the choice of measure(s) is not obvious.
- E.g., which model is better:

Model A: correctly predicts 98% of observed tornados;
Model B: 98% of tornadic forecasts are in fact correct.

Extremness/rareness exacerbates almost every one of these issues.

My lesson 2: Must evaluate! So,
- Make choices, where applicable.
- Rely on Statistics.
- E.g., to check overfitting, estimate prediction/generalization error.

See Chapter2.pdf, and R Code. More, below.

Types of Forecasts and Observations.

Are the observed and/or forecast parameters discrete or continuous?
E.g., Yes/no tornado, or temperature (in 0.1 degress).

Is forecast deterministic or probabilistic?
E.g., Yes tornado, temperature=32,
prob(tornado)=0.9, prob(temp=32) = 0.9

Example: binary x, binary f

E.g.: $x =$ non/existence of tornado $= 0, 1$, $f = 0, 1$
Best tool: Contingency Table (C-table).

$$\text{C-table} = \begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix},$$
$$= \begin{pmatrix} . & \text{false alarms} \\ \text{misses} & \text{hits} \end{pmatrix}.$$

$N_0 = n_{00} + n_{01} = n_{0.}$, $N_1 = n_{10} + n_{11} = n_{1.}$, $N = N_0 + N_1$.

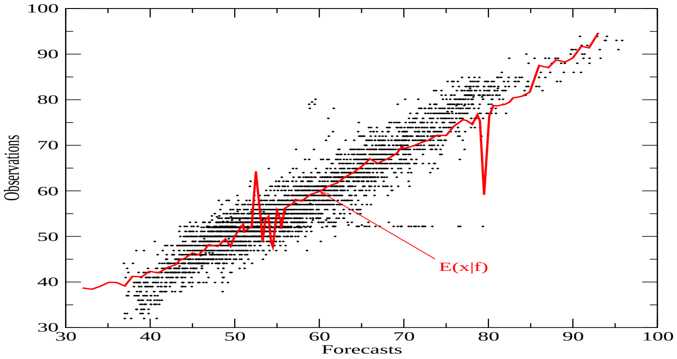
Example: continuous x, continuous f

E.g.: x, $f =$ daily T highs = 32, 33, 34, ...

Best tool: Scatterplot of f vs. x. Effectively C-table.

$$\begin{pmatrix} n_{32,32} & n_{32,33} & \cdots \\ n_{33,32} & n_{33,33} & \cdots \\ \cdots & & \end{pmatrix},$$

$n_{32,33} =$ no. of days with $x = 32$, $f = 33$.



Example: binary x, probabilistic f

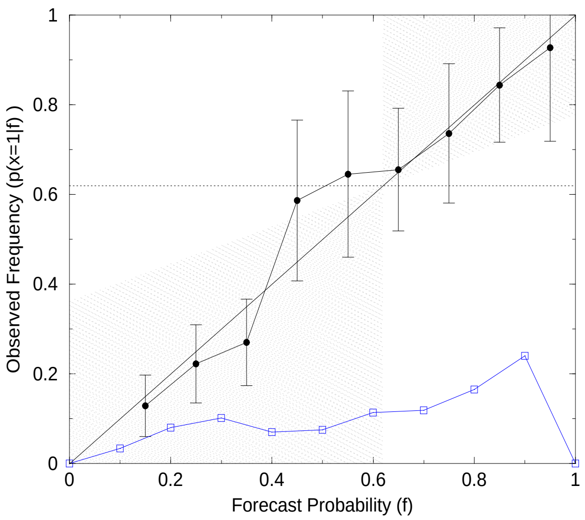
E.g.: x = non/existence of tornado = 0, 1, f = prob. of tornado.

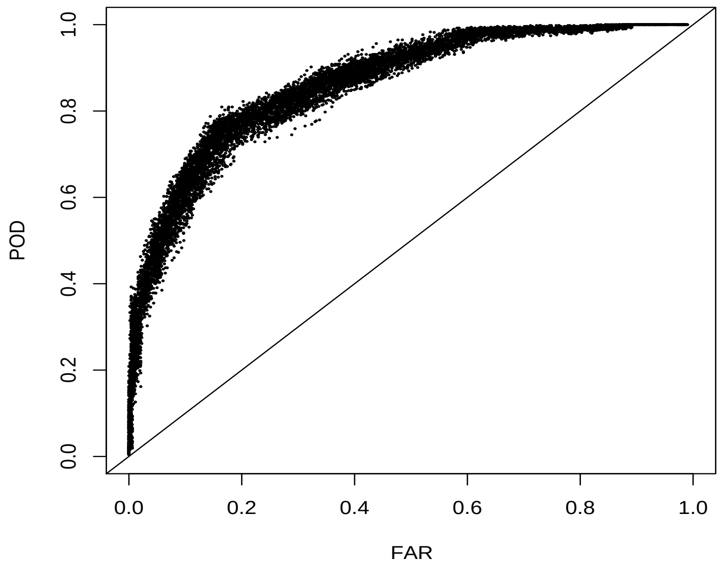
If prob = intervals 0.0-0.05, 0.05-0.15, 0.95-1.0, then again C-table:

$$\begin{pmatrix} n_{0,0} & n_{0,1} & \dots & n_{0,10} \\ n_{1,0} & n_{1,1} & \dots & \end{pmatrix},$$

$n_{0,1}$ no. nontornadoes with forecast probs 0.05-0.15. Etc.

Two good, and complementary measures:





Often 2 × 2 is enough, especially for extreme events;
even for continuous vars one can do peaks over threshold, block maxima, or quantiles.

Measures of Performance

Three common (scalar) measures:

$$\mathrm{H} = \mathrm{POD} = \frac{n_{11}}{N_1}$$

$$\text{False Alarm Ratio} = \frac{n_{01}}{n_{01} + n_{11}}$$

$$\text{False Alarm Rate (FAR)} = \frac{n_{01}}{N_0}$$

A measure of skill:

$$\mathrm{HSS} = \frac{2(n_{00}n_{11} - n_{01}n_{10})}{n_{0.}n_{.1} + n_{1.}n_{.0}}.$$

A measure of (frequency) Bias:

$$\mathrm{B} = \frac{n_{.1}}{n_{1.}}.$$

There are many more.

Rare-event Situation

Here is the problem with most of them:

$$\text{Fraction Correct (FRC)} = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{1 + \frac{n_{11}}{n_{00}}}{1 + \frac{n_{01}+n_{10}+n_{11}}{n_{00}}} \to \frac{1+0}{1+0} \to 1.$$

Q: Why is this bad?
A: It's noninformative

Many measures are also inquitable (allow for "hedging.")

Doswell et al. (1990), Marzban (1998), Stephenson et al. (2008), et al.:
"all" measures are sick in some way.
From Marzban (1998): $N = N_0 + N_1, p = N_1/N =$ base rate.

	$\begin{pmatrix} N_0 & 0 \\ 0 & N_1 \end{pmatrix}$	$\begin{pmatrix} 0 & N_0 \\ 0 & N_1 \end{pmatrix}$	$\begin{pmatrix} N_0 & 0 \\ N_1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & N_0 \\ N_1 & 0 \end{pmatrix}$	Random.
PRD	1	p	0	0	...
AVG	1	$\frac{1}{2}(1+p)$	$0, \frac{1}{2}$	0	...
FRC	1	p	$1-p$	0	...
EFF	1	0	0	0	...
CSI	1	p	0	0	...
TSS	1	0	0	-1	0
HSS	1	0	0	...	0
GSS	1	0	0	...	0
CSS	1	$(p-1), (1-p)$	$(1-p), -p$	-1	0
DSS	1	0	0	1	0
DIS	∞	1	1	∞	1
θ	0	0	0	$\pi/2$	...
ϕ	0	...	...	0	...
Bias	1	$1+\frac{N_0}{N_1}$	0	$\frac{N_0}{N_1}$	...

Conclusion: No good measures.

Some limits are 0/0, or ambiguous.
To tame them, I assumed forecast quantity is conditionally Normal.
That's a probability model.

Conclusion: Same!

Stat 101: What's a probability model?

E.g., Is a coin fair? I.e., is $\pi = 0.5$?

Prob model:
x = no. of Hs out of 100 coins is binomial, with parameter π:

$$p(x) = \begin{pmatrix} 100 \\ x \end{pmatrix} \pi^x \, (1-\pi)^{100-x}$$

Now, collect data,
i.e., toss 100 coins, record no. of Hs, say 48.

Q: What is the value of π that maximizes the likelihood of this data?

A1: Not obvious, but fully anticipated, it's $48/100 = 0.48$.

A2: More importantly, 95% confident that π is in

$$0.48 \pm 1.96 \sqrt{\frac{(0.48)(1-0.48)}{100}} \quad ,$$

$$95\% \text{ CI for } \pi : (0.38, 0.56) \quad .$$

Conclusion: No evidence that coin is not fair.

Fast-Forward 10 years:

Ferro (2007) used a better probability model.
Better because it assumes a better distribution (than Normal) for extremes.
See Richard Smith's lecture.

Very briefly, suppose $x \sim$ normal.
Then a sample of size 100 will look normal.
But what about the largest of the 100 numbers, x_{max}?
Q: What is its distribution?
A1: Gumbel: $G(x) = e^{-e^{-\frac{x-a}{b}}}$, with parameters a, b.
A2: More generally, Generalized Extreme Value family.

Ferro adopts a distribution with two parameters: κ, η;
to be estimated from data.
He found the following C-table/N:

$$\begin{pmatrix} 1 - 2\,p + \kappa\,p^{1/\eta} & p - \kappa\,p^{1/\eta} \\ p - \kappa\,p^{1/\eta} & \kappa\,p^{1/\eta} \end{pmatrix}$$

$p = \frac{N_1}{N}$.
Note: $H = \kappa\,p^{\frac{1}{\eta}-1}$.

Stephenson, et al. (2008) compute measures from this C-table.
Conclusion: Most measures, still sick.

However, one seems OK: Extreme Dependency Score (Coles et al. (1999)):

$$EDS = \frac{2\,\log((n_{11} + n_{01})/n_{..})}{\log(n_{11}/n_{..})} - 1 = \frac{2\,\eta\,\log(p)}{\log(p) + \eta\,\log(\kappa)} - 1$$

Moreoever, upon making a few more assumptions, the 95% CI is:

$$EDS \pm 1.96\,\frac{2\,\log(p)}{H\,(\log(p\,H))^2}\sqrt{\frac{H(1-H)}{n\,p}}.$$

Note the resemblance to the previous coin example.
EDS does not depend on frequency Bias. So, supplement with B.

If unhappy about those CI assumptions, can use Resampling (e.g., bootstrap).

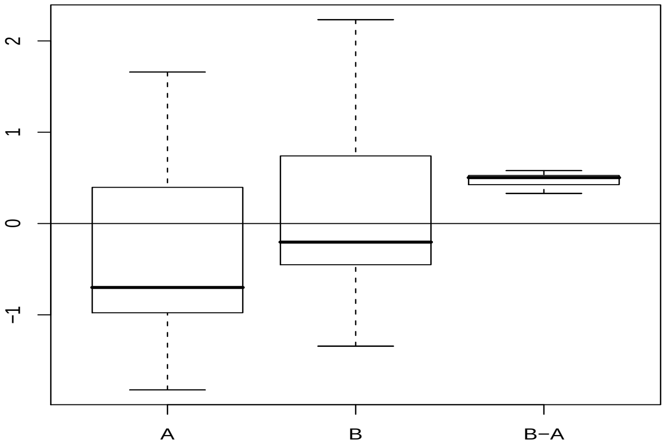
My suggestion: look at the actual empirical sampling distribution of EDS.
E.g., Instead of plotting EDS and its confidence band vs. p,
plot boxplots of EDS vs. p.
Then,
1) Don't have to worry about violating assumptions underlying CI, and
2) Boxplot conveys more information than CI.

Warning: May need double-bootstrap (p.28 of Chapter2.pdf, Tian et al. 2007):
Recall we need *prediction* error, e.g. EDS on *independent* data.
Do one round of resampling to (point-)estimate it (trialin in my R code).
Do another round of resampling to get *its distribution* (trialout).

One Technical Point

Suppose you want to compare Model A with Model B.
Suppose no. of (outer) bootstrap trials = 10.
So, you have 10 prediction values of EDS from each model.

Trial	EDS from A	EDS from B
1	-0.8254514	-0.2458394
2	-0.7292382	-0.3318763
3	0.2141812	0.7405906
4	0.3961835	0.7258984
5	-0.6709980	-0.1610687
6	-1.8215229	-1.3428739
7	1.3792800	1.8051315
8	-0.9831074	-0.4873622
9	-0.9769927	-0.4511232
10	1.6592086	2.2323081



According to above advice, you should look at the boxplots:

Supposedly, based on the significant overlap between Model A and B, they are statistically equivalent.
However, you must look at the boxplot of the *difference*, because the data is so-called paired.
And now you see that Model A is significantly better than B.

Epilogue (Take Aways)

0) Subliminal message: Avoid performance evaluation.

1) Evaluate in terms of figures. e.g., scatterplots, conditional histograms, ...
See Chapter2.pdf (Marzban, in Haupt, Pasini, and Marzban (2008)) on CD.

2) If you must use scalar measures, use many.
See Chapter2.pdf and handout_2004.pdf on CD.

3) Beware of the behavior of the measures in rare-event situations.
For example, examine your favorite measure under Ferro's model.

4) For now, EDS and frequency Bias seem good. Caution: EDS is new.
It appears to be unqiue/special,
but so did TSS, until 1995 (Marzban and Lakshmanan 1995).

5) Examine the sampling variability of the measure, either with CIs, or boxplots from double-bootstrap.

6) Issues rarely covered for rare-events:
- Value; extreme is usually associated with (high) cost.
- Spatial forecasts; Special issue of WAF and Intercomparison Project (http://www.ral.ucar.edu/projects/icp/).
- Probabilistic forecasts (from ensemble, or otherwise).

References

Ferro C. A. T., 2007: A probability model for verifying deterministic forecasts of extreme events. *Wea. Forecasting*, **22**, 10891100.

Haupt, S.E, A. Pasini, and C. Marzban (2008): *Artificial Intelligence Methods in the Environmental Sciences,* Springer. ISBN 1402091176

Jolliffe I.T., Stephenson D.B., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons: Chichester, UK.

Marzban C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753763.

Marzban, C., V. Lakshmanan, 1999: On the uniqueness of Gandin and Murphy's equitable performance measures. Monthly Weather Review, Vol. 127, No.6, 1134-1136.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Wea. Forecasting, 8, 281-293.

Stephenson, D. B., B. Casati, C. A. T. Ferroa, and C. A. Wilson, 2008: The extreme dependency score: a non-vanishing score for forecasts of rare events. *Meteorol. Appl.*, **15**, 4150.

Tian, B. L., T. Cai, E. Goetghebeur, L.J. Wei, 2007: Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, **94** (2), 297-311, doi:10.1093/biomet/asm036 .

Verification FAQ: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_pa