PERFORMANCE ASSESSMENT

Caren Marzban
http://www.nhn.ou.edu/˜marzban

Introduction

This handout is an expanded version of the lecture notes. Although in my previous two lectures/handouts I got into the business of performance evaluations, there is more to it than I said there. In this lecture, we'll delve deeper into that business.

The topic of performance assessment is extremely complex, but some very smart people have done some very good work on it. The one name that always pops up is Allan Murphy. In fact, almost all of the following can be found in his papers (referenced in the text, below). Here is how he sets up the problem:

Designate observations with x and forecasts with f. For example, x could be the number of daily highs collected over some location, and f would be the daily forecast of the highs. That constitutes an example of continuous observations and continuous forecasts. It may also be that the observations are categorical, say binary - labeled by 0 and 1 - while the forecasts are probabilities for the occurrence of the two categories (or classes). In fact, each of our basic quantities (x and f) can come in any of the following varieties: continuous, categorical, and probabilistic. So, you can imagine there are many permutations that can arise, and each one has its own specific tools for assessing the agreement between x and f.

We will not consider all the combinations, but only three examples:
A - Continuous observations, and continuous forecasts.
B - Binary observations, and binary forecasts.
C - Binary observations, and probabilistic forecasts.
By a continuous variable we mean one that takes a reasonably dense set of values, like temperature. Even if temperature is measured in integer values ranging from -100F to 100F, we would still consider it continuous. Now, if a continuous variable takes 10 integer values between 1 and 10, one could argue that one is dealing with a categorical variable with 10 categories. There are a number of other relevant concepts (e.g. nominal, ordinal, etc.) that we will not get into here. In statistics, the prediction of continuous variables falls under the topic of regression, and that of categorical variables is called a classification problem.

Before we get into the framework for handling these problems, let me say that it is an extremely difficult task to decide between the better of two forecasts (or models). First of all, “better” is easy to decide if the comparison between the models is done in terms of a single scalar measure. But what if the performance measure that's relevant to my problem is not relevant for you? The fact is that the quality of forecasts is a multifaceted thing. Performance has many faces as well. We can gauge it in terms of many different measures, and it's entirely possible that one model will outperform another model in terms of one measure but not another. In general, it's better to think ahead of time about what measure is relevant to ones problem. However, that's often easier said than done. In practice then, it's recommended to consider as many distinct measures as possible. In fact, it's better to employ diagrams rather than scalar measures (e.g., mean square error) because diagrams can display more of the multidimensionality of the underlying problem. This is what I'll do here.

Examples

In example A, suppose the daily temperature highs are measured and forecast in integers, say 32, 33, 34, ... Then, a scatterplot of the two will display almost everything we care for. Equivalently, one may display the scatterplot in terms of a (contingency) table:

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

where $n_{32,33}$, stands for the number of $x = 32$ and $f = 33$ days in the data set. In my convention, the rows (columns) correspond to observations (forecasts). Ideally, we would expect a scatterplot showing a tight scatter of points about a diagonal line of slope 1. In terms of the corresponding table, we would like to see a mostly diagonal table, with very small off-diagonal elements.

In example B, suppose we are forecasting tornadoes. In that case, we would represent the existence of a tornado with a 1, and its nonexistence with a 0. A scatterplot would not display much information, but the corresponding contingency table carries all necessary information. It looks like

$$\begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix}, \tag{1}$$

where n_{01} represents the number of nontornadoes incorrectly classified/predicted as tornadoes (i.e., false alarms). Etc.

In this tornado/nontornado problem, if the forecasts are probabilities, then we are dealing with example C. Of course, with probabilities ranging from 0 to 1 (or 0 to 100%), the forecasts are then continuous, but we can instead talk about probability categories. For example, the range 0-1 can be broken up into 11 intervals: $p < 0.05$, $0.05 \le p < 0.15$, ..., $0.85 \le p \le 0.95$, $p \ge 0.95$. Then the table can be written as

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

where $n_{0,1}$ represents the number of nontornadoes that have forecast probabilities in the range 0.05-0.15, etc.

Joint Distributions

It may not be obvious, but in all of these examples we are talking about the so-called *joint* probability density of observations and forecasts, denoted $p(x, f)$. Let's write the elements of the contingency table as n_{ij}, with i=1,2,...I, and j=1,2,...,J. If we allow for $I \neq J$, then all of the above examples can be accommodated. Then

$$p(x = i, f = j) = \frac{n_{ij}}{n_{..}},$$

where a "." indicates a summed index. For example, $n_{.1} = n_{01} + n_{11}$, and $n_{..} = n_{00} + n_{01} + n_{10} + n_{11}$.

All the necessary information is carried in $p(x, f)$. However, it turns out to be useful to break down $p(x, f)$ into *conditional* probabilities. The following equation follows from the basic laws of probability:

$$p(x, f) = p(x|f)p(f) = p(f|x)p(x) \ \ ,$$

where $p(x|f)$ is the (conditional) probability of observation x, given that the forecast is f. For example, $p(x = 0|f = 1)$ is the probability of a nontornado, given that the forecast is for a tornado. Generally, one can think of $p(x = i|f = j)$ as the *belief* that a class-i event was assigned to class j. $p(x)$ is the climatological probability of x, i.e., the probability of x when nothing is known about forecasts. Note that whereas $p(x, f) = p(f, x)$, $p(x|f)$ is not equal to $p(f|x)$. With a data set at hand, all of these probabilities can be estimated as ratios. Specifically, a little bit of thought can justify that

$$p(x = i|f = j) = \frac{n_{ij}}{n_{.j}} \ \ , \ \ p(f = j|x = i) = \frac{n_{ij}}{n_{i.}} \ \ , \qquad (2)$$

and

$$p(x = i) = \frac{n_{i.}}{n_{..}} \ \ , \ \ p(f = j) = \frac{n_{.j}}{n_{..}} \ \ . \qquad (3)$$

Keep in mind that these probabilities can still be written as tables or matrices. In fact, $p(x, f)$, $p(f|x)$, and $p(x|f)$ are often called the performance matrix, percent confusion matrix, and the belief matrix, respectively. For instance, in example C we can write

$$p(x = i|f = j) = \begin{pmatrix} n_{00}/n_{.0} & n_{01}/n_{.1} & ... & n_{0J}/n_{.J} \\ n_{10}/n_{.0} & n_{11}/n_{.1} & ... & n_{1J}/n_{.J} \end{pmatrix} \ \ ,$$

where $J = 11$ (corresponding to the 11 intervals defined above). To make contact with some of the measures employed in my previous lectures, note that the plot of $p(x = 1|f = j)$ (i.e., the second row) as a function of j itself is nothing but the reliability diagram. More on this, later.

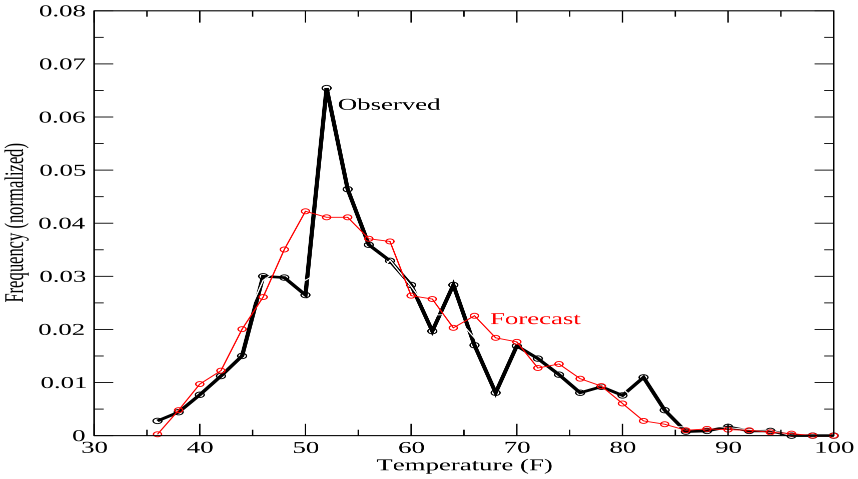
Note that although once in a while I specialize to the case of binary observations or forecasts, the idea of assessing performance in terms of $p(x, f)$ or its 4 "factors"

$(p(x|f), p(f|x), p(f), p(x))$ is completely general. One more point worth mentioning is that of the 4 factors, one is independent of the forecasts, i.e., $p(x)$. It is an attribute of the problem itself and not of the forecaster. So, we need to look only at the other 3 probability matrices.

Back to Example A

How do we formulate the three examples (A, B, C) within this framework of the joint distribution of forecasts and observations?

In example A, both the forecasts and observations are continuous. Let's again consider daily temperature highs. $p(x)$ would be the easiest one illustrate. Just compute a histogram of the observed temperatures. The overall shape of this histogram is a good representation of $p(x)$. [1] The next easy factor is $p(f)$. This can be represented with a histogram of the forecasts. Here is an example of $p(x)$ at Sacramento and $p(f)$ I produced (from a neural net):

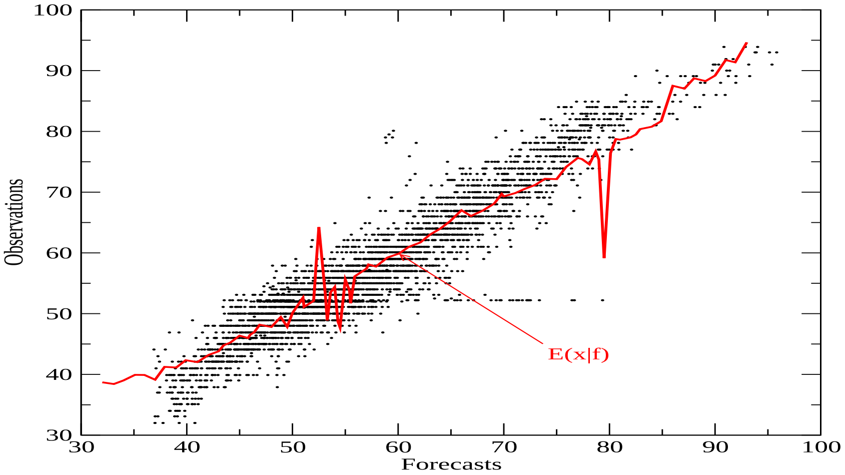


You can see that for temperature, both of these (marginal) probabilities are somewhat bell-shaped. Although based on this example it's hard to see why, $p(f)$ is said to measure the refinement or sharpness of the forecasts. Meanwhile, notice that $p(f)$ is quite similar to $p(x)$, which is a good thing. If one were shifted with respect to the other, I would say that the forecasts were biased. But, here, they are unbiased.

Now, $p(x|f)$ and $p(f|x)$ are harder to draw, because each one is effectively an infinite

[1] The shape of a histogram can depend a lot on the bin size. If you have plenty of data, then it's not a big problem. But for small sample sizes, you'll have to be careful about what you call the shape of the histogram. At least, experiment with different bin sizes to get a feeling of how much the shape changes.

dimensional matrix. Take $p(x|f)$ for example. We could compute the reliability diagram mentioned above, i.e., a plot of $p(x = 32|f)$ as a function of f itself. But, then we had better look at $p(x = 33|f)$, too. And $p(x = 34|f)$, etc. In other words, even though we are looking at x only as integers, we would still have many reliability plots to make. Things become quickly unwieldy. It's more practical to look at $E(x|f)$, which is the expected value (or average) of x, given f, and plot that as a function of f. Reliable forecasts would then yield a diagonal line. In other words, forecasts are reliable if for any forecast, the average of the corresponding observations is equal to the forecast itself. So, I'm going to take all the forecasts of 32, and average the corresponding observations. Then, repeat that for the next value of the forecast. Etc. The result is the solid (red) line in the following plot. I've superimposed this reliability curve on the scatterplot itself so that you can see how much of a spread there is about the reliability curve.



So, now we see that my forecasts are quite reliable. Don't expect me to quantify what I mean by "quite." As I mentioned in the beginning, we must refrain from quantifying things too much. This reliability diagram conveys a lot more information than a number would. In fact, take a look at the paper by Murphy, Brown, and Chen (*Weather and Forecasting*, 1989, **4**, p. 485) to see how we can go even beyond just the (conditional) average of the observations. They consider different quantiles of the observations to assess the "spread" of the reliability curve. Finally, one can and should look at $E(f|x)$, but in this case the plot looks very similar to the above, and so, I won't show it. If the forecasts were more poor, it would be worthwhile.

Since I mentioned the scatterplot, I might as well say that it is one of the best ways of displaying the quality of forecasts. Any deviation from a cloud of dots scattered about a diagonal line of slope 1, means something has gone wrong with the forecasts.

A related plot is the residual plot, i.e. a plot of the difference between forecasts and observations versus the forecasts. This one should look like a cloud of dots scattered about the horizontal line. The real use of this plot is not the plot itself, but the histogram of the residuals. That histogram reflects the distribution of the errors. It is this distribution that is often assumed to be normal (or gaussian) if you are testing some hypothesis. So, this would be a good place to post-check the validity of that assumption. Plot the histogram of the residuals and make sure it is at least bell-shaped.

Finally, let's get over one more very important notion - that of bias vs. variance. One of the reasons for its importance is that it offers a diagnostic way of evaluating the forecasts. It goes without saying that in evaluating continuous forecasts against continuous observations the most common summary measure is the mean squared error (MSE). But MSE can be decomposed as

$$\mathrm{MSE} = \frac{1}{N}\sum_i^N (x_i - f_i)^2 = (\bar{x} - \bar{f})^2 + \sigma^2_{(x-f)} \quad ,$$

where $\sigma_y^2 = \frac{1}{N}\sum_i^N (y_i - \bar{y})^2$ is the variance of a variable y. The over-bar indicates average over the N cases. The difference $\bar{x} - \bar{f}$ is called bias. The second term on the right hand side is the variance of the difference between the forecasts and observations (i.e. the individual errors or residuals). So, in words, MSE is equal to bias (squared) plus variance. This decomposition is important because it can tell you how much of the overall error (MSE) is due to an overall bias, and how much is due to the spread of the individual errors. I can assure you that if you decompose the MSE in this way, you will be handsomely rewarded, because it can often tell you what is wrong with the forecasts.

Back to Example B

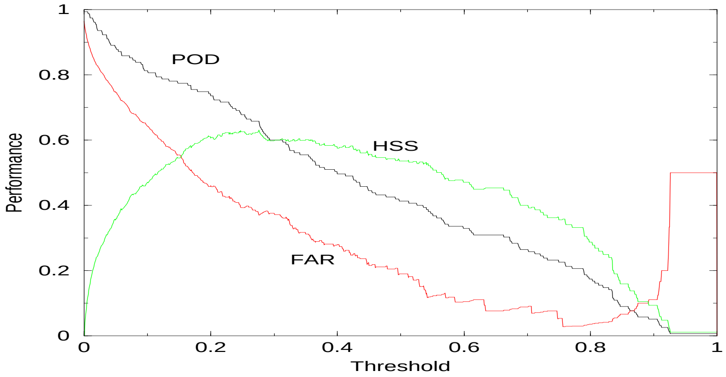
In this example, we consider the case of binary observations and forecasts, e.g., yes/no forecasts of rain/no-rain. This is probably the most common example, because even when neither forecasts nor observations are binary, one can always reduce them to binary, albeit at the cost of some information loss.

The elemental entity is still the joint distribution $p(x, f)$, and its factors. x is either 1 or 0, depending on whether an event (e.g., rain) was observed or not. And the forecast f is also 1 or 0, with a similar meaning. As mentioned above, all three matrices can be computed from the contingency table, n (see eqn 1). Three common summary (scalar) measures are Probability of Detection (POD), False Alarm Ratio, and False Alarm Rate:

$$\text{Probability of Detection} = \frac{n_{11}}{n_{1.}} \quad , \quad \text{False Alarm Ratio} = \frac{n_{01}}{n_{.1}} \quad , \quad \text{False Alarm Rate} = \frac{n_{01}}{n_{0.}} \quad .$$

Now if what you have to assess is just one set of binary forecasts and observations, then all you can construct is one n_{ij} matrix. And from that you compute the different probability matrices, and some summary measures, and you would be done. At least, done in the sense that the matrices $p(x), p(f), p(x|f)$, and $p(f|x)$ could all be computed and written down concisely. The only thing you would still want to do is to compare your results to something else, for example to forecasts based on climatology alone. Take a look at Marzban (1998, *Weather and Forecasting*, **13**, 753-763)to see how to do that. The matrix that you would get based on climatology is called the expected matrix (or table), and is called E in that paper. There is some ambiguity that gives rise to slightly different notions of E, and that's why there is also an E^* in that paper. By the way, when a performance measure is defined to take into some base-rate (like climatology), it is called a *skill* score. We'll see one of these later.

That's if you have only one set of binary forecasts and observations. Often, though, the observations are binary, but the forecast is derived from some continuous quantity. For example, suppose I want to predict tornado vs. no tornado based on wind speed. Then, what I can do is to place a (decision) threshold on wind speed, and forecast everything above that threshold as tornado and everything below it as nontornado. For a given threshold, then, I would have one set of binary forecasts and observations. But now I can increment the threshold by some amount, and recompute the probability matrices. As such, I'll end up with a large number of matrices, depending on how small my threshold increments are. Now, I can plot some summary measure as a function of the threshold to see at what decision threshold I could maximize my performance. Here is an example of what it would look like.



The reason the x-axis is between 0 and 1 is not that wind speed varies in that range. I simply scaled my wind speeds to lie in that range - an unnecessary step. The POD on the plot is Probability of Detection, the FAR is the false alarm ratio (not rate), and HSS is a summary measure called Heidke's skill score, defined as

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

HSS is designed to be zero for climatological forecasts, and for that reason it's a skill score. In this example, we clearly want to place a threshold at around 0.25 to obtain maximum HSS. The corresponding values of POD and FARatio are then about 70% and 40%, respectively. We will return to this threshold approach, later, when we are dealing with probabilistic forecasts.

Scalar Measures

Speaking of scalar measures of performance (like HSS), the probability matrices can be employed to construct scalar/summary measures of performance. As I have already said numerous times, I do not recommend this, because not only too much information is lost, but also it gives the false impression that performance can be assessed in terms of only one measure. Still, there are times when one simply needs a single scalar measure. So, let's talk about them just a bit anyway.

In statistics, one takes a weighted sum of the "belief" probabilities, and calls the result the class-conditional expected risk:

$$R_i = \sum_j L_{ij} p(f=j|x=i) \ \ ,$$

where the weights L_{ij} are called the loss matrix. The loss matrix is supposed to be supplied by the user. Another weighted sum of these guys gives something called the Bayes risk

$$R = \sum_i p(x=i) R_i \ \ ,$$

where the weights $p(x=i)$ are the climatological probabilities, not supplied by the user. In meteorological circles, Bayes risk is called the expected score, and the (transpose of) the loss matrix is called the scoring matrix. You can see that the probability matrices can be employed in different ways to arrive at different scalar measures of performance (or risk); even just by changing the loss matrix, for example.

There have been numerous attempts to single out a unique scalar measure of performance as better than others. Even the giants of the field have faltered. Gandin and Murphy (1992: *Mon. Wea. Rev.*, **120**, 361-370) show that if the loss matrix is symmetric, then only one measure - the True Skill Score (or Kuipers' performance index) - is an "equitable" measure. However, later we showed (Marzban and Lakshmanan, 1999: *Monthly Weather Review*, **127** 1134-1136) that the assumption of a symmetric loss matrix is too severe and unrealistic, and that without it the uniqueness of the True Skill Score goes away. In short, there does not exist a single good measure, however you define good. The choice of a measure of performance is contingent on the user.

Another defect of most (if not all) scalar measures is that they convey the wrong amount of the true performance in some situations. One common situation is where the number of events (e.g., tornadoes) is much smaller than the number of nonevents. This is referred to as the rare event situation, and requires special handling. The details of the story can be found in Marzban (1998: *Weather and Forecasting*, **13**, 753-763). Although the paper tries to identify the healthier of several popular measures, the only important point to learn is that in rare event situations, all scalar measures incorrectly assess performance. The simplest way of seeing this is to examine one very popular measure, namely Fraction Correct. In terms of our contingency table, this would be $(n_{00} + n_{11})/n_{..}$. A rare event situation means the number of events is much smaller than the number of nonevents, i.e., $n_{1.} << n_{0.}$. Writing this out, we get $n_{10} + n_{11} << n_{00} + n_{01}$. If the forecaster has any skill at all, what happens is that the 00 element dominates the other three terms, i.e., $n_{00} >> n_{10}, n_{01}, n_{11}$. Then, the Fraction Correct can be written as

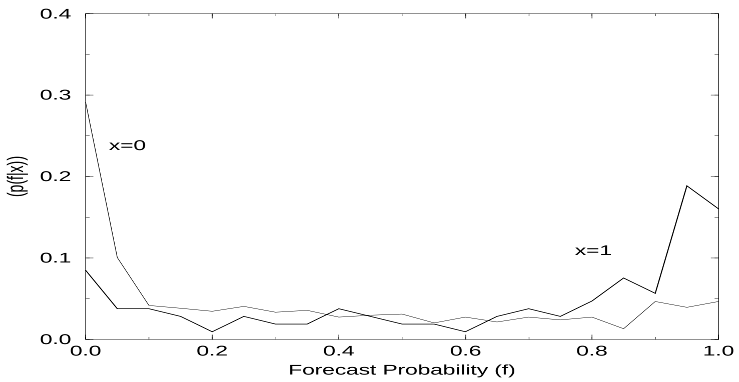
$$FC = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{1 + \frac{n_{11}}{n_{00}}}{1 + \frac{n_{01}+n_{10}+n_{11}}{n_{00}}} \to \frac{1+0}{1+0} \to 1.$$

In short, if events are rare, even a modestly skilled forecaster can reach near 1 (i.e., 100%) values for fraction correct. What this means is that FC is not a good measure in rare event situations. But as shown in the above mentioned paper, FC is not alone. All scalar measures have some kind of a pathological behavior, about which little can be done. At the least, one should check to assure that one's measure of choice is not too sick. As shown in the above paper, HSS is one of the relatively healthy ones. But, please, do try to avoid scalar measures if you can.

Back to Example C

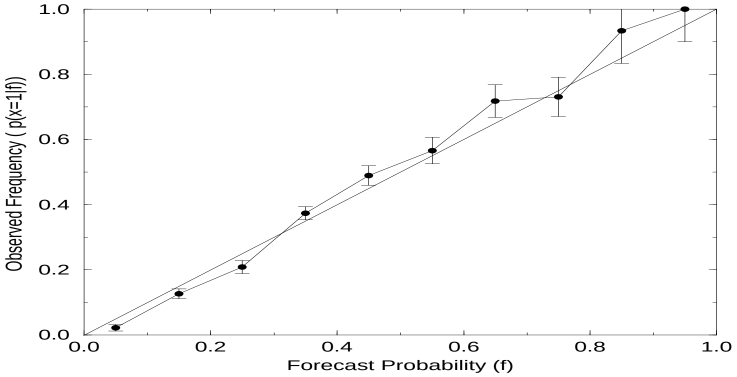
Now we come to the example where the observations are binary, and the forecasts are probabilities. How do we assess the quality of these probabilistic forecasts? We have already laid down all the relevant quantities: $p(f), p(f|x)$, and $p(x|f)$. There is no point in thinking much about $p(x)$, because $p(x = 1)$ is just the climatological probability of the class 1 event (e.g. tornado), and $p(x = 0) = 1 - p(x = 1)$. Like all probabilities in this note, you compute them as ratios of numbers (see equations 2,3).

First, let's look at $p(f|x)$. Since x is binary, we can look at $p(f|x = 0)$ and $p(f|x = 1)$. In other words, identify all the nonevents $(x = 0)$, and plot a histogram of their forecasts (f). Then repeat for all $x = 1$ cases. You'll get something looking like this



Ideally, you would want to get two semi-bell-shaped curves with no overlap, in which case you would say that your forecasts can discriminate perfectly between the events and nonevents. In reality, though, there is always some amount of overlap. Nevertheless, this type of plot (i.e., $p(f|x)$) offers a measure of the discrimination capability of the forecasts. Of course, we can quantify this by, say, computing the area of the overlap, but let's not.

Now, how about $p(x|f)$? We've seen this one many times already. A plot of $p(x = 1|f)$ versus f itself is the reliability plot. It looks like this

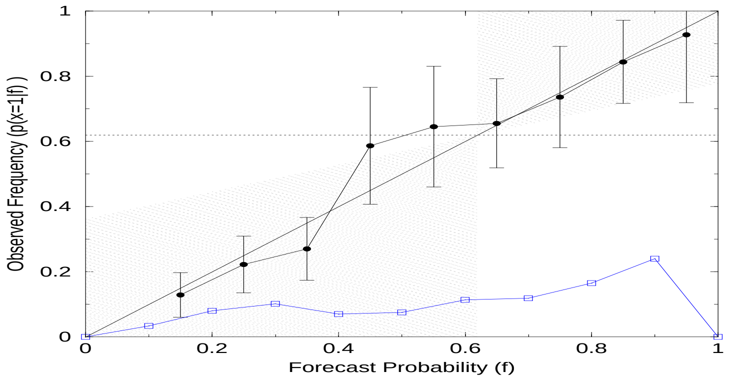


It measures the degree to which the forecasts agree with the observed frequency of events. In other words, we want a tornado forecast of 10% to mean that exactly 10% of such forecasts are in fact tornadic. Same for 20%, 30%, etc. So, ideally, the plot should be a diagonal line. In this case, the figure shows that another one of my NNs produces forecasts that are in fact perfectly reliable (or calibrated). How can I claim *perfect* reliability, when the points don't fall exactly on the diagonal line? Because of the error-bars. Computing them is a different can of worms, but you should at least keep in mind that all of these things we are computing have error bars regardless of whether or not we display them. That way, we won't be abandoning the forecasts leading to the above reliability plot and spend years trying to improve them. They cannot be improved (unless one has more data, of course).

Finally, $p(f)$. It's plot is called a refinement plot, and it gauges how sharp the forecasts are. If the forecasts are always the same number or in the same range, then the forecasts are not sharp. If on the other hand, the forecasts are only 0 or 1, then they are said to be very sharp. Clearly, good forecasts reside somewhere in the middle. Its plot looks very much like that of $p(f|x)$ if you just ignore the difference between $x = 0$ and $x = 1$. You'll see an example of this below.

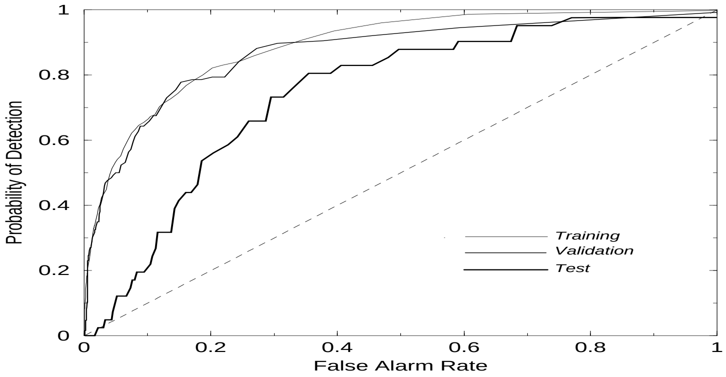
That covers all the relevant factors of the joint distribution. But there is one more plot that conveys even more information; it's called an attributes diagram. It's basically a reliability plot but with some extras. This is the way to produce it:

1) Plot the reliability diagram.
2) Draw a horizontal line at y-value $p(x = 1)$.
3) Draw a vertical line at x-value $p(x = 1)$.
4) At the point where the two meet, draw the bisector of the angle made by the horizontal and the *diagonal*.
5) Now, shade the area outlined by this bisector and the vertical line.
This may all sound very *ad hoc*, but the fact is this: Points on the reliability plot that happen to fall in this shaded region contribute positively to the Brier Score. This score is just the mean squared error of the forecasts and observations, where the former are probabilities, and the latter binary. This diagram is important, because recall that it is possible to have a perfectly reliable set of forecasts with no skill at all. The attributes diagram allows you to see which forecasts are reliable *and* skillful simultaneously. For the derivation of all these statements, see the 1992 paper of Murphy and Winkler (p. 443) and page 263 of the book by Wilks (Statistical Methods in the Atmospheric Sciences, Academic Press, 1995). As if the attributed diagram doesn't already convey too much information, I like to superimpose the refinement plot on it too. Here is an example of the whole thing:

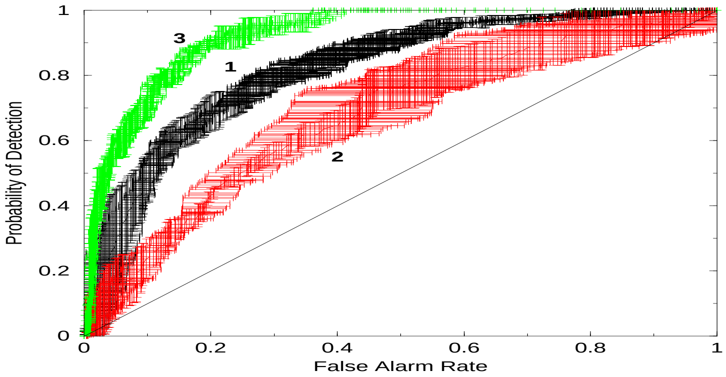


If we were to ignore the error-bars for a moment, you can see that the mid-range forecasts (45% and 55%), although pretty reliable, are skill-less, because they reside outside the shaded region. The refinement curve shows a nice spread across the whole range, with a slight bump at 0.3 and 0.9. There is also something called the resolution of the forecasts that can be read from the attributes diagram, but look that one up in the Murphy/Winkler paper or in Wilks' book.

You might think that this should end the discussion of example C. But there is (at least) one more thing, namely the Receiver's Operating Characteristic (ROC) Curve. The ROC is popular because it allows for an easy comparison between different forecasts or models. But, as advertised it is still based on the probability matrices we have been talking about. In fact, it is based on the Probability of Detection (POD) and the False Alarm Rate (not ratio), defined above. This is how you make it: Put a threshold on the forecast probability, say at 0.01. Then call everything above that threshold as an event, and everything below, as nonevent. That gives you a contingency table from which you can compute the POD and FARate. Then plot those two numbers on a plot of POD vs. FARate. Now, increment the threshold, and repeat the above steps to get another POD and FARate. Continue this until the threshold reaches 1. Now, the locus of all the points you have been plotting is called the ROC curve. Here is one example, showing three ROC curves.



In case I haven't said it already, The more the ROC curve bows above the diagonal, the better the forecasts. In fact, the diagonal line corresponds to random forecasts. You can probably guess that the forecasts underlying this ROC plot are from a neural net again. You can also see that the training and validation curves are somewhat similar, but the ROC curve for the test data (i.e., the truly independent data) is quite a bit worse (but still good). So, one question is how can we assess the difference between the curves? What I do is to come-up with some way - any way - to put some error bars on the curve. Of course, the error-bars would have to be in both x and y directions because both POD and FARate need separate error bars. Here is one:



In this case, I estimated the error bars with bootstrapping (see my previous lectures). You can come-up with other ways of putting error bars on the curve. In fact, I would say it's OK to have the wrong error bars as long as they are there, because they will remind you that two ROC curves may in fact be statistically equivalent. In that case, your funders cannot cut the funding for the *apparently* worse forecasts in favor of the *apparently* better ones.

Conclusion

I'm running out of time for turning in this handout, and so, I won't try to recapitulate what I've said above. Hopefully, I can do that in the lecture itself. And hopefully, if/when you get your hands dirtied with these things, it'll all start to make sense.