Precision-recall curves

Andreas Beger
Duke University

15 April 2016

Abstract: ROC curves and the area under them, AUC-ROC, are sometimes used to assess model fit in international relations and other research. At the same time, the outcomes of interest like civil war or interstate war onset are rare events, leading to data that mostly consists of 0's, with few 1's. AUC-ROC is misleading for such data and overstates the actual performance of a model because it does not capture what likely is low precision in the predictions–many false positives for every true positive. Precision-recall curves and the area under them, AUC-PR, are based on precision rather than the false positive rate, and thus better reflect model performance when predicting rare outcomes.

adbeger@gmail.com or @andybeega.

Code to plot PR curves, calculate AUC-PR, and replicate the examples are at https://github.com/andybega/auc-pr

ROC curves are a fairly standard way to evaluate model fit with binary outcomes, like (civil) war onset. I would be willing to bet that most if not all quantitative political scientists know what they are and how to construct one. Unlike simpler fit statistics like accuracy or percentage reduction in error (PRE), they do not depend on the particular threshold value used to divide probabilistic predictions into binary predictions, and thus give a better sense of the tradeoff between true and false positives inherent in any probabilistic model. The area under a ROC curve (AUC) can summarize a model's performance and has the somewhat intuitive alternative interpretation of representing the probability that a randomly picked positive outcome case will have been ranked higher by the model than a randomly picked negative outcome case. What I didn't realize until more recently though is that ROC curves are a misleading indication of model performance with kind of sparse data that happens to be the norm in conflict research.

	$p < \theta$	$p \geq \theta$
Y=0	True Neg.	False Pos.
Y=1	False Neg.	True Pos.

To recap, the basic situation is that we have a binary outcome, but a stream of predictions that as probabilities range between 0 and 1, and the challenge is how to map this onto the binary outcomes. We could calculate Brier scores and avoid the problem, or we could choose a particular threshold and calculate things like accuracy, percentage reduction in error, etc. These measures rely on how positive predictions match up with observed outcomes (see the confusion table above), but the drawback is that they depend on a particular threshold value, and will change as the threshold changes. A way around this is to record and plot all possible combinations over the range of possible threshold values, and this is essentially what ROC curves are.

Observed	Predicted
0	0.1

Observed	Predicted
0	0.6
1	0.4
1	0.8
…	…

To construct a ROC curve, you would pick all possible thresholds, bin predictions to 0 or 1, and then calculate the true positive rate and false positive rate associated with each threshold, giving you the data you would need to plot the ROC curve. The true and false positive rates are calculated as the ratio of true positives (cases the model got right) to overall positives in the data, and the ratio of false positives (cases the model predicted 1, but that did not have a positive outcome) to overall negatives in the data. Here's an example using some simulated data I'll discuss more below:

Threshold	TPR	FPR
0.99996	0.00000	0.00000
0.99995	0.00048	0.00000
0.99943	0.00048	0.00034
0.99894	0.00097	0.00034
0.99877	0.00145	0.00034
0.99870	0.00194	0.00034

Which gives the following ROC curve if we plot the TP and FP rates:

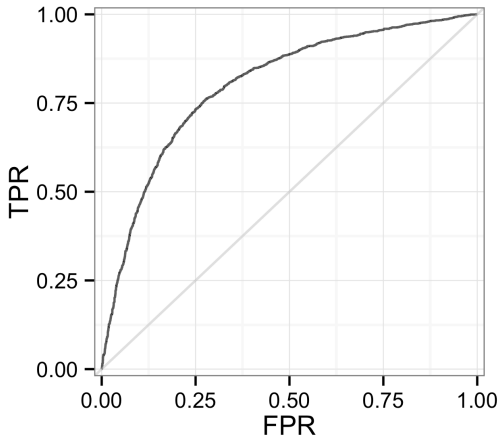

Figure 1: ROC curve for the example data.

In this example, about 40% of outcomes are positive, but this is rarely the case in international relations and conflict research in particular, where data tend to be sparse, with much fewer positive outcomes for things like war or civil war onset and occurrence. Fearon and Laitin's (2003) paper on civil war onset has 167 positives per 10,000, and the two projects I mostly work on these days have rates of 17 per

10,000 for country-month irregular leadership changes (Beger et al., 2016)[1] and 1-5 per 10,000 for IEDs in Afghanistan (Beger et al., 2015).[2]

With sparse data it becomes pretty easy for any model to correctly predict negatives. Because the ROC curve in part plots false positive rates that are calculated with the resulting large number of true negatives in the denominator, by that metric we will seem to be doing pretty well. The table below shows a contrived example using numbers similar to what one might get with the Fearon & Laitin 2003 data and a reasonably good model. With these number we get 50% recall (recall is the same as TPR) and a false positive rate of only 9%. In the ROC world, we are doing well. Except that looking at the table it is obvious that the model predictions are still problematic: for every correct positive prediction our model makes, there are 10 false positives.

	$p < \theta$	$p \geq \theta$
Y=0	10,000	1,000
Y=1	100	100

Since it becomes easier to predict negatives as they become more common, looking at false positive rates with sparse data might not be that informative. Instead, let's plot something else. The only option, assuming that we do care about positives and hence the true positive rate or recall, would be to compare false positives to the overall number of positive predictions made by a model for a given threshold. This is called precision, and I think of it as how believable a model's predictions are ("My model says 1, what are the actual chances this is true?").

The plot below is a precision-recall curve that does this, for the same example as before. Instead of FPR we now have precision, and I've also flipped the axes as it seems to be convention to plot recall on the x-axis.

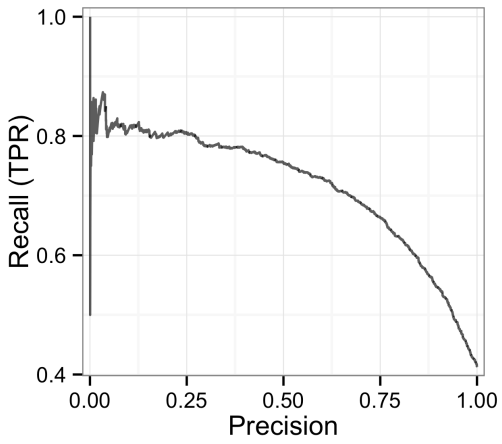

Figure 2: Precision-recall curve for the same example data with 0.4 positives.

[1] Also http://predictiveheuristics.com/2014/05/22/the-coup-in-thailand-and-progress-in-forecasting/

[2] And http://andybeger.com/2014/08/08/modeling-and-predicting-ieds/

Simulations!

Since the example I used had a positive rate of 0.4, the plot doesn't really make it obvious why one would want to look at precision-recall curves for sparse data. To illustrate that better, below are two plots from a simulation where I created 3 data sets with decreasing positive rates (0.4, 0.1, 0.01) and for each data set then created 3 models designed to achieve a particular AUC-ROC value (0.8, 0.9, 0.95).

The first series of plots are the ROC curves. Since the models are meant to match a given AUC, these shouldn't really look different, and they don't, as we move to the increasingly sparse datasets. The curves get a little bit edgier on the right, but this is just because there are less positive outcomes based on which recall/TPR are calculated. All models are doing equally well if we use ROC curves and AUC as our metric.

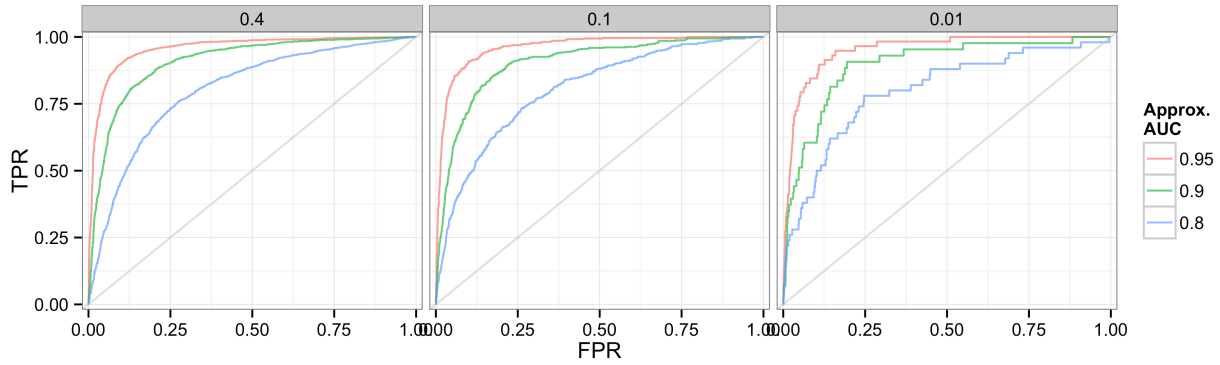

Figure 3: ROC curves for 9 models designed to achieve a given AUC, across 3 data sets with the positive ratios varying from 0.4, 0.1 to 0.01. By design, the curves are largely similar, except for some grainy-ness as the number of positive cases decreases.

The corresponding precision-recall plots on the other hand show the loss of precision as one moves to sparser data, and here it becomes more obvious that the sparse data present more of a challenge. On the right, even the 0.95 AUC model barely touches on 0.5 precision (1 correct positive for 1 false positive), and if we were to calculate the area under the PR curves (AUC-PR) we'd get values much lower, 0.25 and less.

Precision-recall curves for the same 9 models of 3 increasingly sparse datasets. The loss of precision as the data become more sparse is apparent, even though all models have the same AUC.

A lot of conflict research is in the world of the rightmost plot, maybe somewhere between the two rightmost plots if you are working with occurrence and country-year data. AUC values that I always thought were great, 0.8, even 0.9 or higher, actually can hide a lot of imprecision–"room for growth" as I like to tell myself in consolation.

Another thing that stands out from these plots is that you can always increase model recall. Just lower your threshold, everything will light up as a potential positive, and incidentally you will capture most if not all actual positive outcomes. Getting high precision on the other hand is much more difficult, and realistically there are

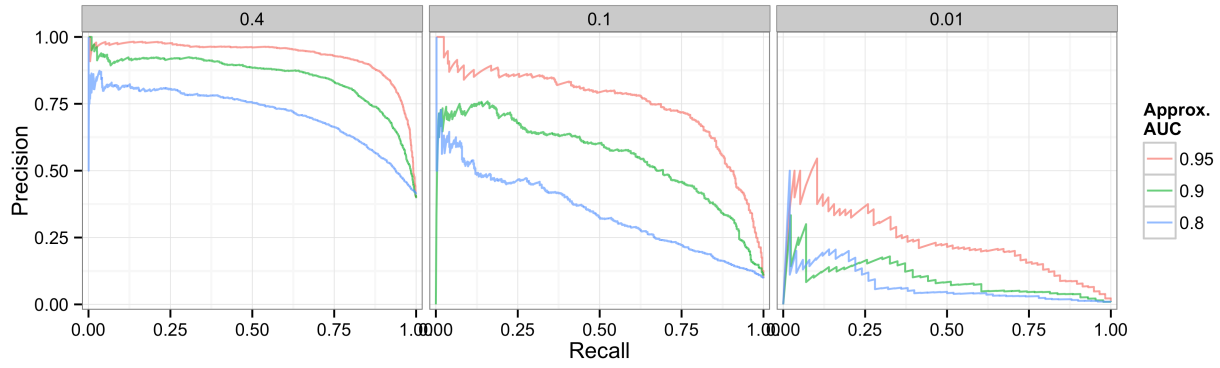

Figure 4: Precision-recall curves for the same 9 models of 3 increasingly sparse datasets. The loss of precision as the data become more sparse is apparent, even though all models have the same AUC.

hard limits here, at least with this kind of problem. With that in mind it seems strange to me, on the few occasions I've been exposed to this, that people commissioning these kinds of forecasts aim for recall, e.g. require that models reach 0.8 or some other threshold, rather than precision, which might make quantitative modeling more credible to non-technical audiences. But then, my cost for false negatives, which I would tend to have more of with this rationale, is probably also much lower.

What about Brier scores?

Brier scores, i.e. mean squared error for binary outcomes and predictions, are another fit statistic for binary models. They were proposed by Glenn W. Brier in 1950 (Brier, 1950), who did not initially name them after himself, and, like many things related to prediction, Brier scores had their origins in weather forecasting. They are calculated as the mean squared difference between predicted and observed values:

$$\text{Brier} = \sum_{i=1}^{N} (p_i - y_i)^2$$

How do they fit in ROC and PR curves? It helps here to go back to a more general view of binary fit, which decomposes the fit into two seperate dimensions, calibration and discrimination. From Cook (2007):

> […] Calibration is a measure of how well predicted probabilities agree with actual observed risk. When the average predicted risk within subgroups of a prospective cohort, for example, matches the proportion that actually develops disease, we say a model is well calibrated. […]
>
> Discrimination is a measure of how well the model can separate those who do and do not have the disease of interest. If the predicted values for cases are all higher than for non-cases, we say the model can discriminate perfectly, even if the predicted risk does not match the proportion with disease. […]

ROC and PR curves are both measures of discrimination only. We implicitly treat the model predictions are meaningless in and of themselves. All that matters is their

relative ranking, and I can in fact transform them in any way that doesn't alter this ranking. For the initial example above, I can cut all probabilities in half and I would still get the same ROC/PR information (recall is 0.5, precision is 0.5, FPR is 0.5):

Observed	Predicted	Predicted / 2
0	0.1	0.05
0	0.6	0.3
1	0.4	0.2
1	0.8	0.4
…	…	…

Although I more or less know how well the models I am working on do in terms of AUC-ROC and AUC-PR, I am not in fact sure to what extent one can take the probabilities generated by them at face value. For one, they tend to be very low, way below 0.1, which makes it a bit harder to assess forecast accuracy (Did I miss, even though I'm getting the baseline right?), but then, we are also dealing with very rare events.

There are some fit statistics that allow one to get at calibration, like calibration curves and the Hosmer-Lemeshow test. Both are based on cutting the data into groups, based on the ranked predicted probabilities, and comparing expected predicted frequencies of events against observed event frequencies. For example, we take the lowest 10% of predictions, calculate the expected number of events, and compare against the observed number of events for that group. Calibratrion curves plot the observed vs. expected frequencies, while the Hosmer-Lemeshow test uses similar binned data to calculate a more formal test of the null hypothesis that the predictions are not calibrated well. Since the binning is based on ranked probabilities, they should pick up a little bit on discrimination as well, but we don't actually know what the ranking within each bin is, so these are mostly measures of calibration.

There does not seem to be an easy way to fit Brier scores directly into a discussion in these terms, but at the extreme end of performance, Brier scores should be sensitive to both discrimination and calibration, since the only way to maximize the Brier score would also maximize both of those measures. But that wouldn't mean that low Brier scores indicate either more of a discrimination or calibration problem, so in the end one would probably have to go back to one of the more specific measures for either dimension to diagnose further.

Conclusion

If you are doing (conflict) research with sparse binary data and are interested for whatever reason in model fit, (1) your models don't do as well as ROC might lead one to believe, and (2) consider precision-recall curves as an addition or alternative.

References

Beger, A., C. L. Dorff, and M. D. Ward (2016). Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models. *International Journal of Forecasting 32*(1), 98–111.

Beger, A., B. Radford, and M. D. Ward (2015). Small-scale prediction of IEDs in Afghanistan using split-population duration regression. Paper Presented at ISA 2015.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*(1), 1–3.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation 115*(7), 928–935.

Fearon, J. D. and D. D. Laitin (2003, February). Ethnicity, Insurgency, and Civil War. *American Political Science Review 97*(01), 75–90.