Resolving Power Online Resource 1

Colin Beam PhD

Abstract

This supplement provides additional examples and details of the resolving power approach using the diabetes readmissions data from the main text.

1 Linear approximation of signal curves

This section uses Appendix A's linear approximation method to estimate resolving power for the diabetes readmissions data and then compares the results to those from the main text. Previously, we formed a fine grid of many $(\text{AUROC}, \text{AUPRC})$ pairs both above and below the baseline model. Now we only find three points: The baseline, one point by nudging the class distributions slightly closer, and one point by nudging them slightly further apart. The approach from Section 6 can be used to determine the nudge size (the $\delta + \epsilon$) needed to achieve a specific AUROC increment. We use a .001 AUROC increment and then fit a line to the three points to approximate the signal curve.

Results are shown in Table 1 below. The AUROC and AUPRC resolving power estimates are nearly identical to those obtained using the more comprehensive grid approach (shown in Table 2 from the main text). In this example, then, a linear signal curve provides a good approximation while requiring much less computation. Moreover, recall that resolving power's central aim is to inform metric choice for model search. This task only requires approximate relative magnitudes; precise resolving power estimates are not necessary.

Table 1 Summary of linear approximation results.

metric	κ	resolving power
AUROC	0.0148	67.6
AUPRC	0.0250	40.1
H-measure	0.0190	52.7

Heretofore, the focus has exclusively been on comparisons between the AUROC and the AUPRC, but resolving power can easily be applied to other metrics. One criticism of the AUROC is that it implies that different classifiers apply different costs to the two types of misclassification.[1] But costs should depend on the relative severity of the outcomes, not on which classifier is used. The H-measure addresses this issue by using the same cost distribution to compare different classifiers (Hand, 2009; Hand and Anagnostopoulos, 2023).

Table 1 shows the resolving power estimate for the H-measure using the default beta prior cost distribution as described in Hand and Anagnostopoulos (2023). The H-measure has resolving power that is intermediate of the AUROC and the AUPRC. This is another example in which different evaluation metrics may be more suitable for different stages of model development. Despite its theoretical shortcomings, the AUROC may still be preferable for model search. Once search is complete, one could then switch to the more informative H-measure to communicate the results.

2 Model improvement assumption

Recall that the two general model improvement assumptions are that:

1. Model improvement is a mostly homogeneous process, and
2. We can approximate this process.

This section addresses the second assumption. The main text specifically assumes additive fixed increments where all scores shift by a constant amount. For many classifiers we can test this assumption by observing how risk scores evolve as the algorithm learns. For example, we can monitor risk scores as the number of iterations increase for gradient boosted trees or neural networks.

We again use the diabetes readmissions data and the XGBoost model. The model was trained over 25, 50, and 75 iterations with the same random seed, allowing us to observe risk scores evolve over their common trajectories. The cross-validation AUROC estimates are respectively 0.641, 0.651, and 0.653, indicating continued improvement as total iterations are increased.

Figure 1 shows that the positive and negative class risk scores become more dispersed as the model learns. We also see that the median risk score decreases for both classes. Our interest is in the mechanism that drives this change.

We can assess the improvement mechanism by exploring how risk scores change across different locations in the class distributions. If the additive improvement assumption holds, then all scores should change by a similar amount regardless of their location in the distributions. Figure 2 plots how each risk score changes from the 25 to 50 iteration model (first_change) and from the 50 to 75 iteration model (second_change).

The red lines show average change while the blue lines show a cubic spline fit. If the additive increment assumption was approximately true, risk score changes would be randomly distributed around the average change, with no trend across distribution location. What we see in actuality is larger negative changes in the lower tails of the

[1] As one reviewer observed in their comments, this property equally applies to the AUPRC.

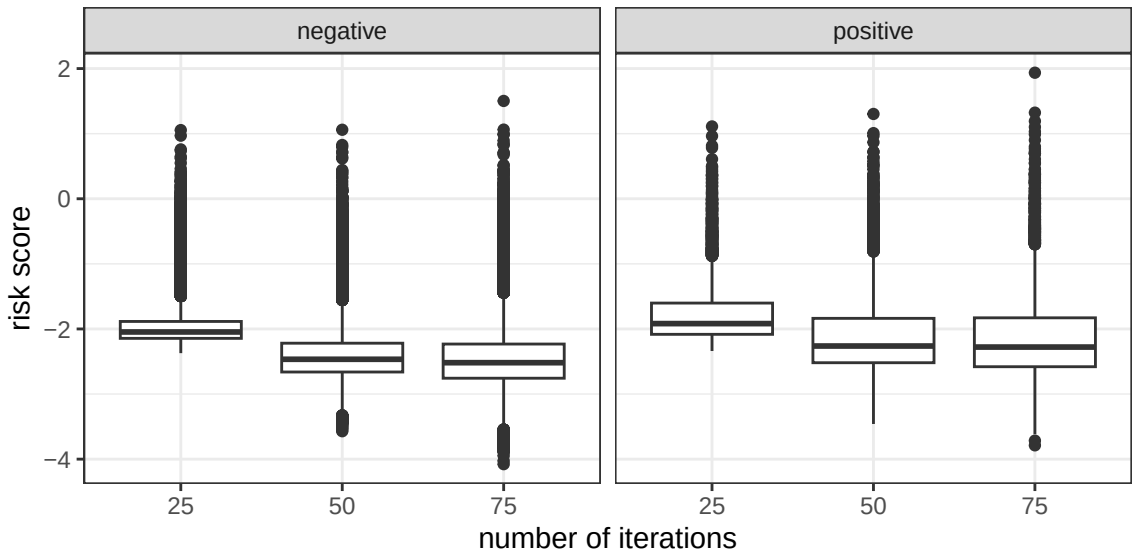

Fig. 1 Boxplots of negative and positive class risk scores over increasing number of boosting iterations

distributions. It appears as if model improvement largely occurs by predicting the negative class with more confidence. Thus, there is some evidence that the additive improvement assumption is violated. From this type of data we could also formulate various statistical tests of the additive improvement assumption. If we are especially concerned about the veracity of the improvement mechanism, we could just use the observed changes in risk scores to model classifier improvement. This empirical path could then be used to estimate the signal curve.

Finally, we might extend this type of analysis to also test the "homogeneous improvement" assumption. We can observe how improvement proceeds under various hyperparamter settings to see if different model types change the risk score distributions in reliably different ways.

3 Practical considerations for estimating resolving power

Section in progress. Important point: Effects estimates should match the size of the validation set. So for the three-way-holdout resolving power should be estimated on only the validation set. For cross-validation resolving power should be estimated from the average of the folds.

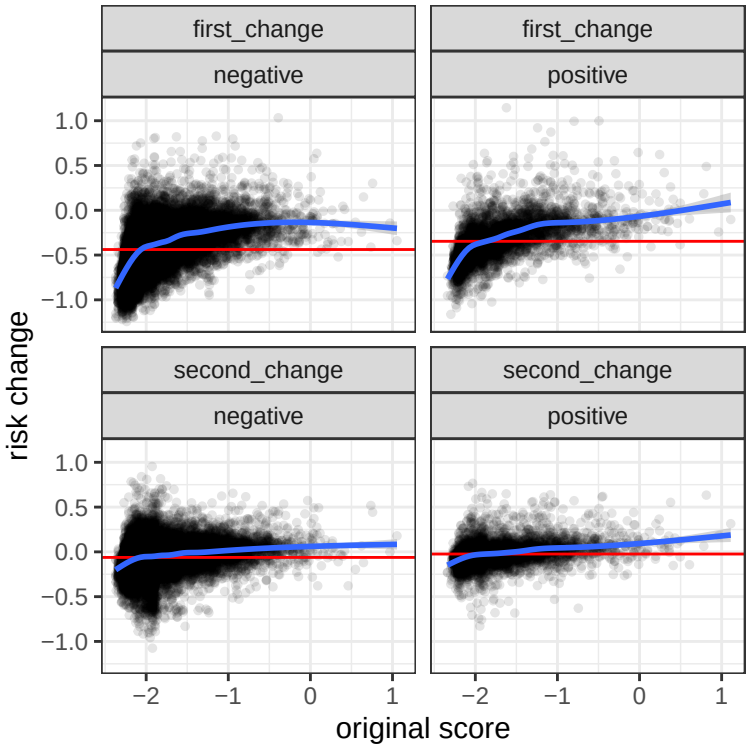

Fig. 2 Risk change versus original score for the positive and negative classes. Average change is shown by the horizontal red lines while blue lines are fit by a cubic spline.

References

Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the roc curve. Machine learning 77(1):103–123

Hand DJ, Anagnostopoulos C (2023) Notes on the h-measure of classifier performance. Advances in Data Analysis and Classification 17(1):109–124