Advanced topics and conclusion

Classifier Calibration Tutorial
ECML PKDD 2020

Prof. Peter Flach, Dr. Hao Song

First.Last@bristol.ac.uk

classifier-calibration.github.io/

Departamento de ESTATÍSTICA

University of BRISTOL

UNIVERSITY OF TARTU

Table of Contents

An alternative view of calibration

- Brier curves
- Calibrating for F-score

Calibration for regression

Concluding remarks

Table of Contents

An alternative view of calibration

Brier curves

Calibrating for F-score

Calibration for regression

Concluding remarks

An alternative view of calibration

So far we have framed calibration as aiming to get as close as possible to the Bayes-optimal classifier, predicting the ‘true’ $P(y|x)$.

An alternative view starts from the observation that the output of a calibrated classifier allows to calculate the cost ratio (or class prior) under which an instance is on the decision boundary.

This view leads to a useful visualisation of calibration by means of **Brier curves**. It also opens the way to calibrating for alternative classification performance measures, in particular **F-score**.

Brier curves I

As we have seen before, relative misclassification costs are given by the cost parameter $c = \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}} \in [0, 1]$.

Decision theory shows that the optimal decision threshold on well-calibrated probabilities is c. Let π denote the proportion of positives.

The normalised loss at threshold c is then

$$L(c) = 2c(1 - \pi) fpr(c) + 2(1 - c)\pi\,(1 - tpr(c))$$

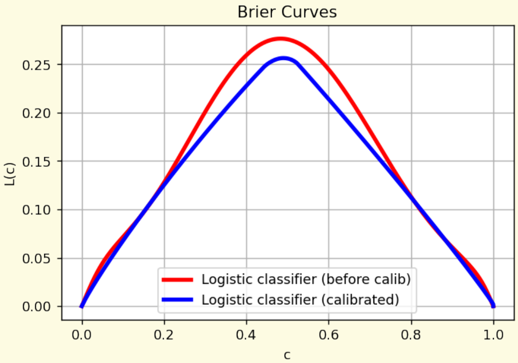
The **Brier curve** plots $L(c)$ against c (Hernández-Orallo et al., 2011). It is a particular kind of cost curve (Drummond and Holte, 2006).

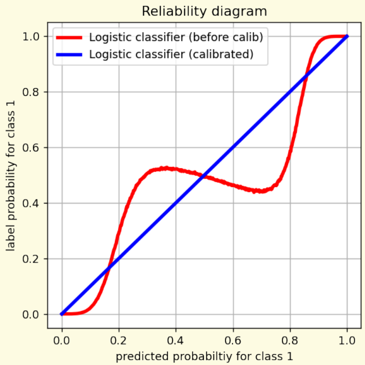
Brier curves II

Properties of Brier curves:

- The area under the Brier curve (i.e., the expected loss over uniform c) is equal to the Brier score.
- Among all possible score transformations, a perfectly calibrated classifier has the lowest possible Brier curve (lower envelope).
- This visualises the decomposition of Brier score into refinement loss and calibration loss.

Brier curves III





The area under the red curve is the uncalibrated model's Brier score; the area under the blue curve is the refinement loss; the area between the two curves is the calibration loss of the uncalibrated model.

Reinterpreting the output of a calibrated classifier

A calibrated score of $p = r/(r+1)$ tells us that this instance's predicted class wouldn't affect performance if negatives are $r = p/(1-p)$ times more important than positives.

So each calibrated score p has an associated **weighted accuracy measure** $acc_p = 2p \cdot (1-\pi)tnr + 2(1-p) \cdot \pi tpr$ for which instances with that score are on the decision boundary.

Q: What changes if we are interested in F-score rather than accuracy?

Calibrating for F-score

F_β is a weighted harmonic mean of precision and recall:

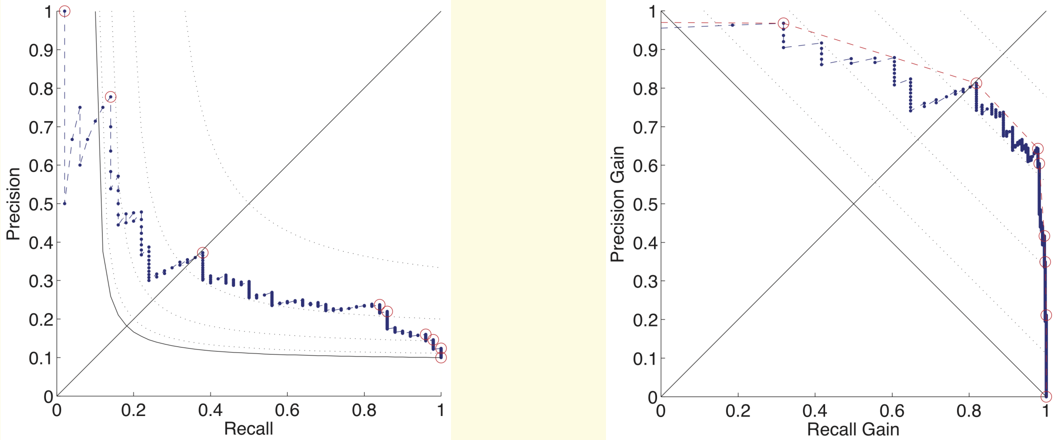
$$F_\beta \triangleq \frac{1}{\frac{1}{1+\beta^2}\Big/ prec + \frac{\beta^2}{1+\beta^2}\Big/ rec} = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + FP + \beta^2 FN}$$

It is more convenient to have a $[0,1]$ weight $d = 1/(1+\beta^2)$:

$$F_d \triangleq \frac{1}{d/prec + (1-d)/rec} = \frac{TP}{TP + d \cdot FP + (1-d) \cdot FN}$$

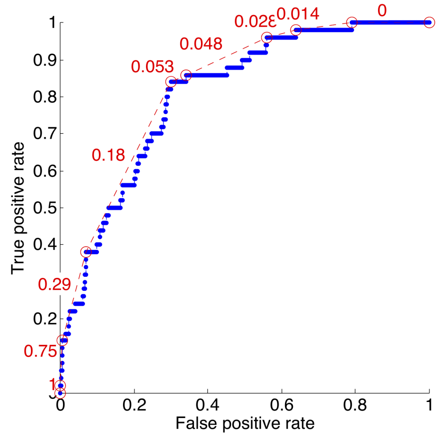
Q: Can we construct a calibration procedure that produces the value of d for which an instance is on the F_d decision boundary?

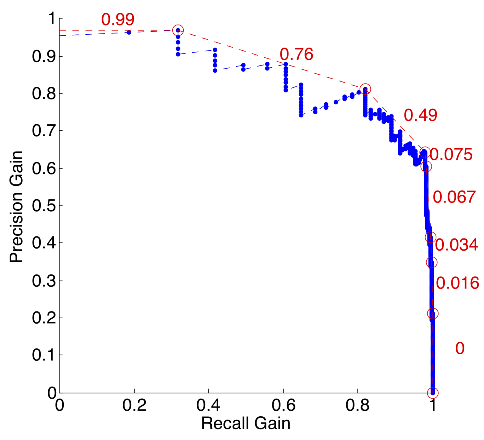
PRG curves to the rescue (Flach and Kull, 2015)



Left: Precision-recall curve. Right: linearised PRG curve.

Isotonic regression applied to PRG curve





Left: ROC curve with accuracy-calibrated scores. Right: PRG curve with F_d-calibrated scores.

Table of Contents

An alternative view of calibration

- Brier curves
- Calibrating for F-score

Calibration for regression

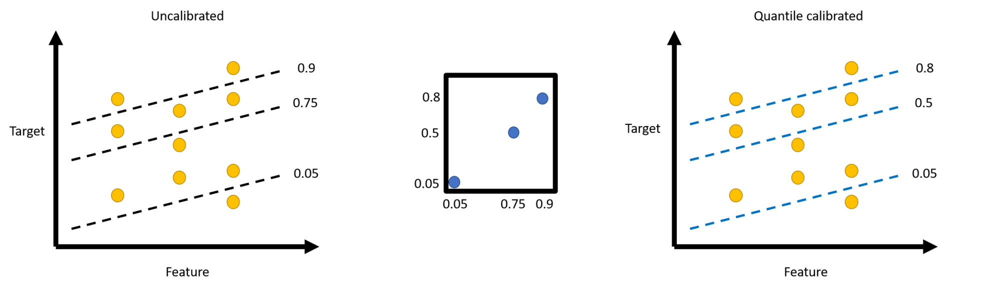
Concluding remarks

Quantile-Calibrated Regression

If we denote a quantile regression model as $g : \mathbb{X} \times [0, 1] \rightarrow \mathbb{R}$, then this model is defined as **quantile-calibrated**, if and only if

$$P\Big(Y \leq g(\mathbf{X}, \tau)\Big) = \tau \qquad \text{for } \forall \tau \in [0, 1]$$

Quantile-Calibrated Regression



Distribution-Calibrated Regression

If we denote a regression model (or conditional density estimator) as

$$f : \mathbb{X} \to \{s \mid s : \mathbb{R} \to [0, \infty), \int s(y)dy = 1\}$$

then this model is defined as **distribution-calibrated**, if and only if

$$P\Big(Y = y \mid f(\mathbf{X}) = s)\Big) = s(y) \qquad \text{for } \forall s : \mathbb{R} \to [0, \infty), \int s(y)dy = 1$$

Being distribution-calibrated is a sufficient but not necessary condition for being quantile-calibrated (Song et al., 2019, Theorem 1).

Distribution-Calibrated v.s. Quantile-Calibrated

Quantile-calibrated regression considers the global quantiles, and distribution-calibrated regression further models local distributions.

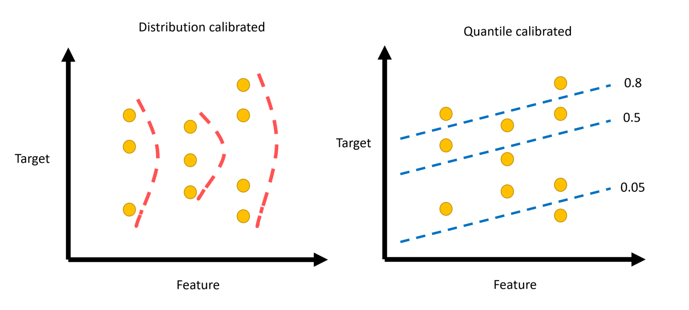


Table of Contents

An alternative view of calibration

- Brier curves
- Calibrating for F-score

Calibration for regression

Concluding remarks

The myth of the decision boundary

Resist thinking in terms of decision boundaries.

- That is like describing Mount Everest with a single contour line, halfway up.
- To faithfully characterise the mountain you need many contour lines at different elevations.

The **only case** in which the decision boundary is useful is when the class prior (and misclassification costs, if any) won’t change after training.

- In that case you don’t need to calibrate the entire probability range, only the decision threshold.

The long history of classifier calibration

Contrary what recent machine learning literature may lead you to believe, **calibration research predates machine learning** and has been studied for three-quarters of a century.

- Make sure you take full advantage of that history, and don't just follow the citations in the latest NeurIPS paper on DNN calibration.

Recent proposals such as confidence calibration, temperature scaling and expected calibration error have their value, but are far from the only options and may not be suitable for your particular application.

- Which leads me to. . .

Calibration is an art as well as a science

Many practical questions don't have *a priori* answers:

- Which calibration method?
- How many bins?
- What evaluation metric?
- Do I use a validation set? How large?
- ...

We hope this tutorial has given you the tools to navigate this space!

References

C. Drummond and R. C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

P. A. Flach and M. Kull. Precision-Recall-Gain Curves: PR Analysis Done Right. In *Advances in Neural Information Processing Systems (NIPS'15)*, pages 838–846, 2015. URL `http://people.cs.bris.ac.uk/~flach/PRGcurves/`.

J. Hernández-Orallo, P. Flach, and C. Ferri. Brier Curves: A New Cost-Based Visualisation of Classifier Performance. In *28th International Conference on Machine Learning (ICML'11)*, pages 585–592, 2011.

H. Song, T. Diethe, M. Kull, and P. Flach. Distribution calibration for regression. In *36th International Conference on Machine Learning*, pages 5897–5906, 2019.

Acknowledgements

- The work of MPN was supported by the SPHERE Next Steps Project funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R005273/1.
- The work of PF and HS was supported by The Alan Turing Institute under EPSRC Grant EP/N510129/1.
- The work of MK was supported by the Estonian Research Council under grant PUT1458.
- The background used in the title slide has been modified by MPN from an original picture by Ed Webster with license CC BY 2.0.

Advanced topics and conclusion

Classifier Calibration Tutorial
ECML PKDD 2020

Prof. Peter Flach, Dr. Hao Song

`First.Last@bristol.ac.uk`

`classifier-calibration.github.io/`

SAPIENTIA ÆDIFICAT Departamento de ESTATÍSTICA

University of BRISTOL

1632 UNIVERSITY OF TARTU