Calibrators

Classifier Calibration Tutorial
ECML PKDD 2020

Hao Song
hao.song@bristol.ac.uk
classifier-calibration.github.io/

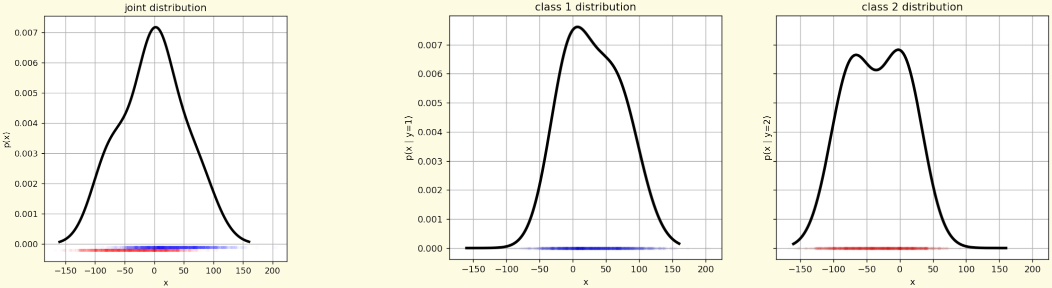
Departamento de ESTATÍSTICA
University of BRISTOL
UNIVERSITY OF TARTU

Table of Contents

Table of Contents

The toy dataset (feature space, generative view)

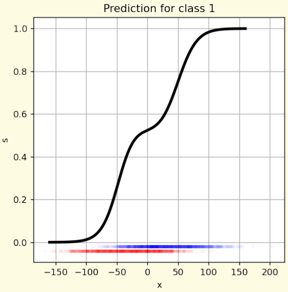
For better illustration, we adopt a toy dataset and a set of visualisations.



The toy dataset (feature space, discriminative view)

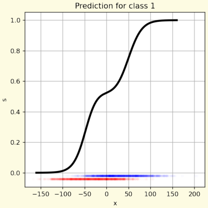
As we know the data distribution, we can calculate the Bayes optimal scoring model, that is:

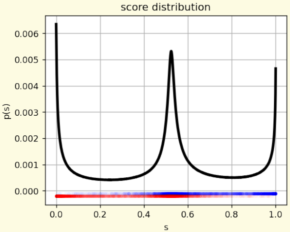
$$s(x) = P(Y = 1 | X = x)$$

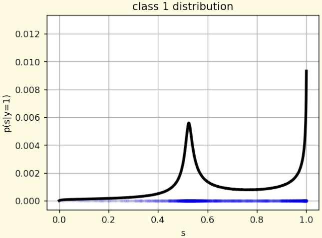


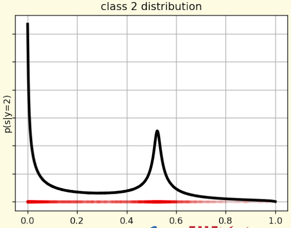
The distribution of predictions

We can further calculate the distribution of $s(x)$.



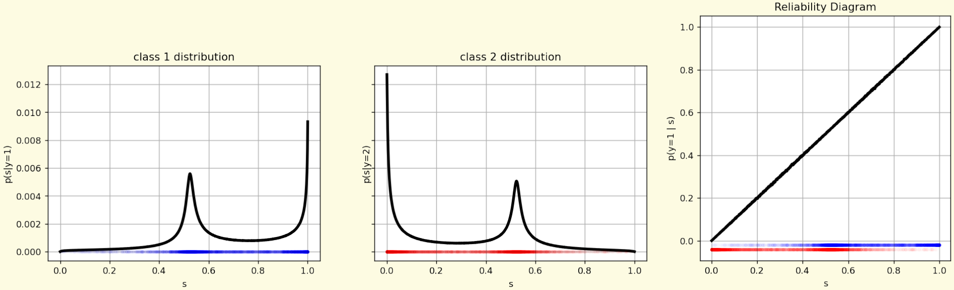






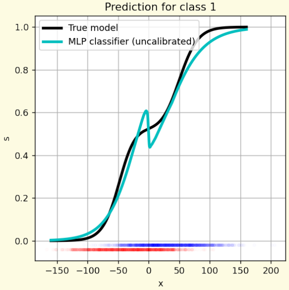
Reliability diagram

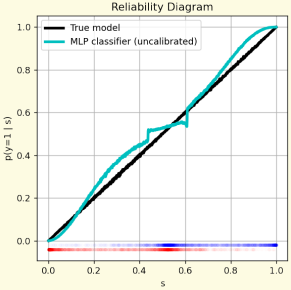
As well as $P\left(Y = 1 | s(x)\right)$ (reliability diagram). Since the model is Bayes optimal, we have a perfect reliability diagram.



With a classifier

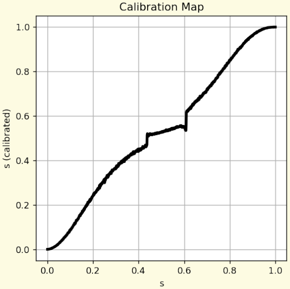
Now if we train a neural network with some samples.





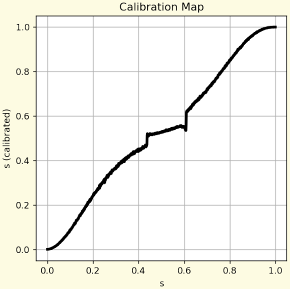
Post-hoc calibrators

This is where we need post-hoc calibrators.



Post-hoc calibrators (scaling view)

- It is also common to re-scale a real vector output into calibrated probability vector space. (e.g. SVM margins, final layer of a neural network)
- Since a real vector and a probability vector can be transformed into each other through link function and inverse link function (e.g. soft-max and logit transform), for later slides we will use the probability vector view by default.



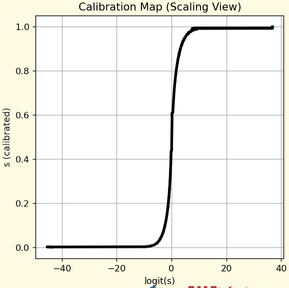


Table of Contents

Calibrators

Binary approaches:

- Empirical Binning
- Isotonic Regression
- Platt Scaling
- Beta calibration

Multi-class approaches:

- Temperature Scaling
- Vector Scaling
- Matrix Scaling
- Dirichlet calibration

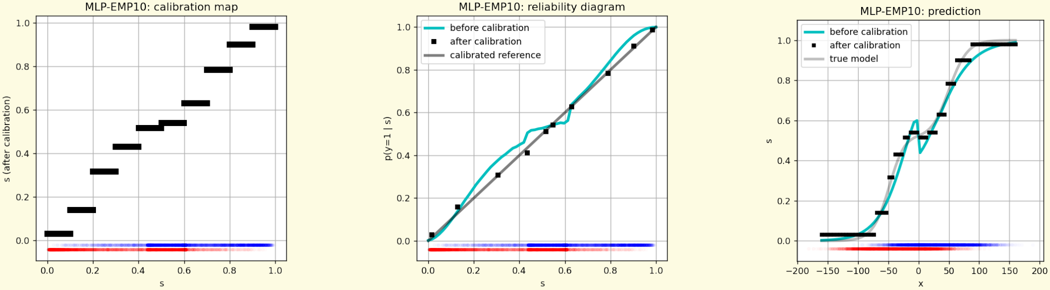
Notable mentions

Binary approaches

- M. P. Naeini and G. F. Cooper. Binary Classifier Calibration Using an Ensemble of Near Isotonic Regression Models.
  In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 360–369. Institute of Electrical and Electronics Engineers (IEEE), feb 2016
- J. Platt. Probabilities for SV Machines.
  In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large-Margin Classifiers*, pages 61—-74. MIT Press, 2000
- M. Kull, T. M. Silva Filho, and P. Flach. Beyond Sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration.
  Electronic Journal of Statistics, 11(2):5052–5080, 2017

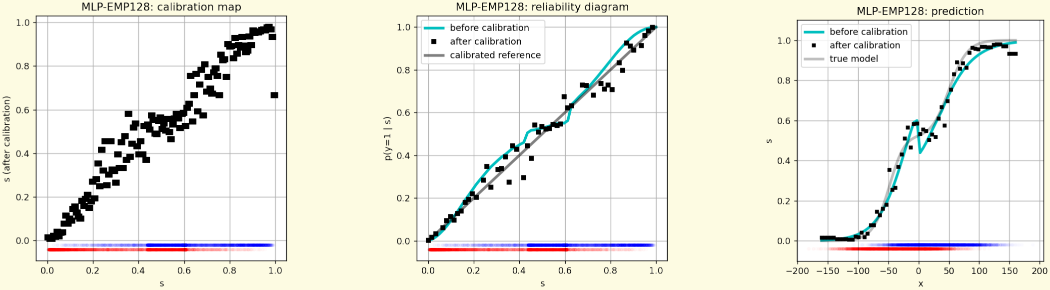
Empirical Binning

While being simple and effective, binning approaches can only give discrete outputs.



Empirical Binning

A suitable number of bins / binning algorithm is important to get good results.



Empirical Binning

Parameters:

Bins: $\{\mathbb{B}_1, \ldots, \mathbb{B}_M\}, \mathbb{B}_j \subset [0, 1]$

Bin averages: $\mathbf{a} = (a_1, \ldots, a_M)$, $a_j \in \{0, 1\}$

Predictive Function:

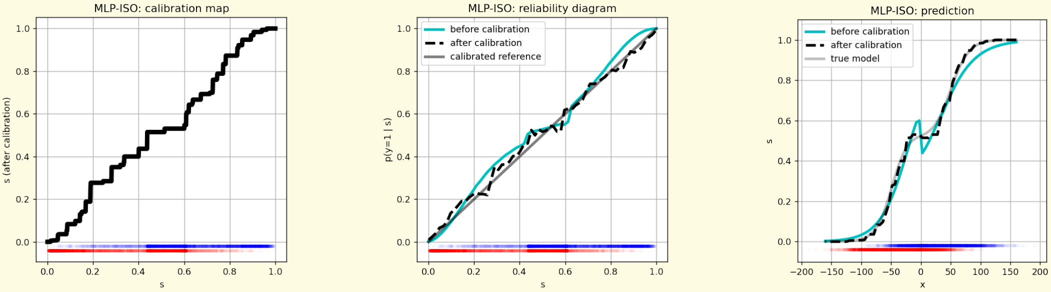
$c(p; \mathbb{B}_1, \ldots, \mathbb{B}_M, \mathbf{a}) = \sum_{j=1}^{M} \mathbb{I}(p \in \mathbb{B}_j) \cdot a_j$

Objective Function:

$\mathbb{L}(\mathbf{a}) = \sum_{j=1}^{M} \left| \frac{\sum_{i=1}^{N} \mathbb{I}(p_i \in B_j) \cdot y_i}{|\mathbb{B}_j|} - a_j \right|$

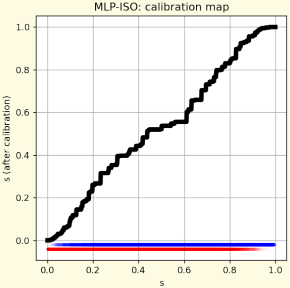
Isotonic Regression

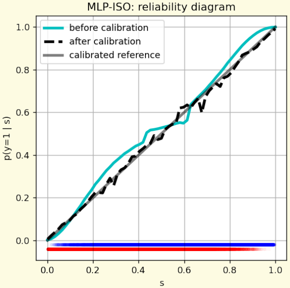
With the ROC-convex hull method, isotonic regression can give good calibration performance with automatic binning and interpolation.

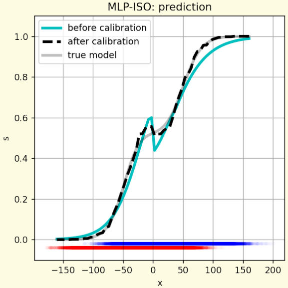


Isotonic Regression

More data points are beneficial for isotonic regression, but the monotonicity assumption might not be suitable for certain datasets / base models.







Isotonic Regression

Parameters:

Bin edges: $\mathbf{b} = (b_1, \ldots, b_M)$, $b_j \in \{0, 1\}$, $b_j < b_{j+1}$

Edge values: $\mathbf{v} = (v_1, \ldots, v_M)$, $v_j \in [0, 1]$, $v_j \leq v_{j+1}$

Predictive Function:

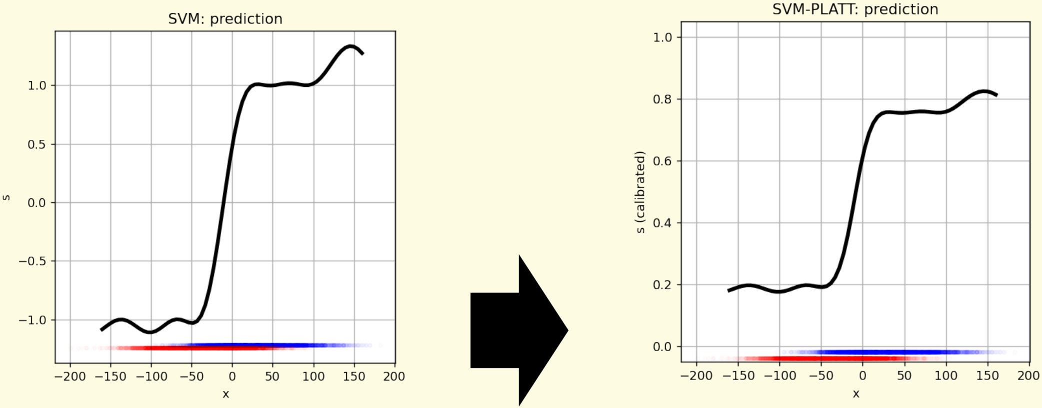
$$c(p; \mathbf{b}, \mathbf{v}) = \frac{1}{\sum_{j=1}^{M-1} \mathbb{I}(p \geq b_j) \cdot \mathbb{I}(p \leq b_{j+1})} \sum_{j=1}^{M-1} \mathbb{I}(p \geq b_j) \cdot \mathbb{I}(p < b_{j+1}) \cdot \left(v_j + \frac{p - b_j}{b_{j+1} - b_j} (v_{j+1} - v_j) \right)$$

Objective Function:

$$\mathbb{L}(\mathbf{b}, \mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} \left(c(p_i; \mathbf{b}, \mathbf{v}) - y_i \right)^2$$

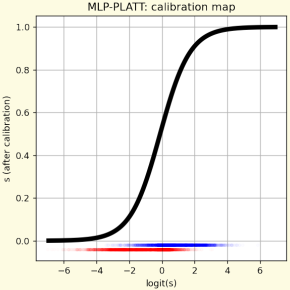
Platt Scaling

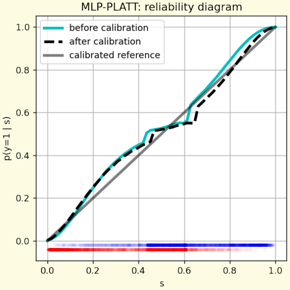
The ONE that allows SVMs to output probabilities.

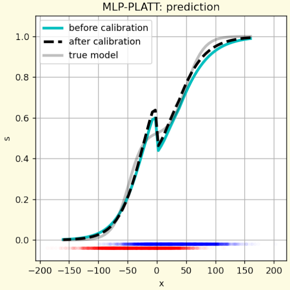


Platt Scaling

We can also use it to calibrate from the final layer of the MLP classifier.







Platt Scaling

Parameters:

Slope: $w \in \mathbb{R}$

Intercept: $b \in \mathbb{R}$

Predictive Function:

$$c(p; w, b) = \frac{1}{1+\exp(-w \cdot p - b)}$$

Objective Function:

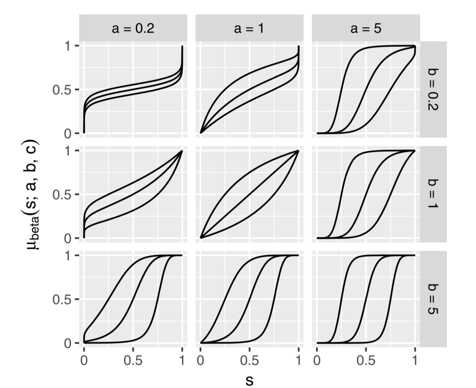
$$\mathbb{L}(w, b) = \frac{1}{N} \sum_{i=1}^{N} \ln\Big(\sum_{j=1}^{2} -\mathbb{I}(y_i = j) \cdot c(p; w, b) \Big)$$

Beta Calibration

While Platt scaling can be derived with conditional Gaussian distribution with shared variance, probabilities are on a finite support hence Gaussian is less suitable.

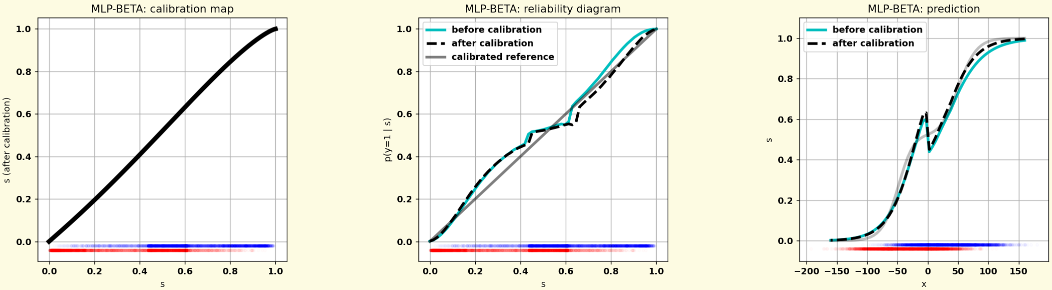
Beta Calibration

With a Beta assumption, we can get calibration maps that get beyond sigmoid.



Beta Calibration

As well as a identity map if the original model is already calibrated, or no better calibration map can be modelled.



Beta Calibration

Parameters:

Slope 1: $a \in \mathbb{R}$

Slope 2: $b \in \mathbb{R}$

Intercept: $c \in \mathbb{R}$

Predictive Function:

$$c(p; a, b, c) = \frac{1}{1+\exp\left(-a \cdot \ln p - b \cdot \ln(1-p) - c\right)}$$

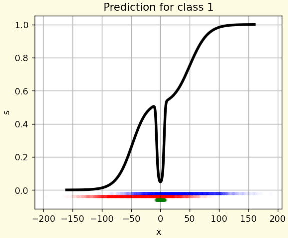
Objective Function:

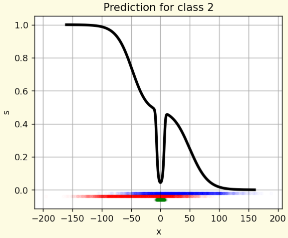
$$\mathbb{L}(a, b, c) = \frac{1}{N} \sum_{i=1}^{N} \ln\left(\sum_{j=1}^{2} -\mathbb{I}(y_i = j) \cdot c(p; a, b, c)\right)$$

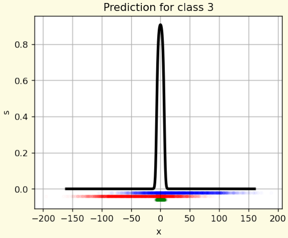
Multi-class approaches

- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks.
  In *34th International Conference on Machine Learning*, pages 1321–1330, Sydney, Australia, 2017
- M. Kull, M. Perello-Nieto, M. Kängsepp, T. S. Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration.
  In *Advances in Neural Information Processing Systems (NIPS'19)*, pages 12316–12326, 2019

Start with adding one more class







Temperature Scaling

Parameters:

Temperature: $t \in \mathbb{R}$

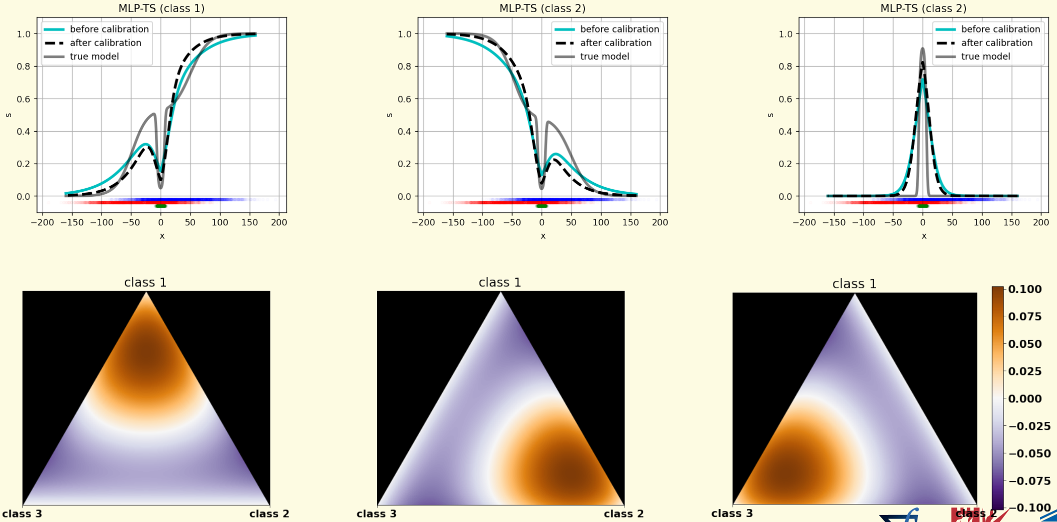
Predictive Function:

$$c_j(\mathbf{p}; t) = \frac{\exp\Big(-t \cdot \text{logit}_j(p)\Big)}{\sum_{j=1}^{K} \exp\Big(-t \cdot \text{logit}_j(p)\Big)}$$

Objective Function:

$$\mathbb{L}(t) = \frac{1}{N} \sum_{i=1}^{N} \ln\Bigg(\sum_{j=1}^{K} -\mathbb{I}(y_i = j) \cdot c_j(\mathbf{p}_i; t)\Bigg)$$

Temperature Scaling



Vector Scaling

Parameters:

Vector: $\mathbf{w} \in \mathbb{R}^K$

Intercept: $\mathbf{b} \in \mathbb{R}^K$

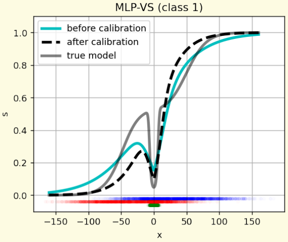
Predictive Function:

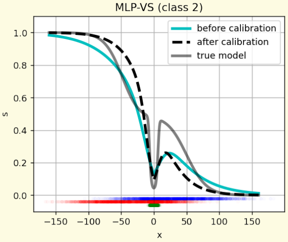
$$c_j(\mathbf{p}; \mathbf{w}) = \frac{\exp\left(-w_j \cdot \text{logit}_j(p) - b_j\right)}{\sum_{j=1}^{K} \exp\left(-w_j \cdot \text{logit}_j(p) - b_j\right)}$$

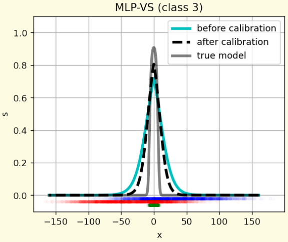
Objective Function:

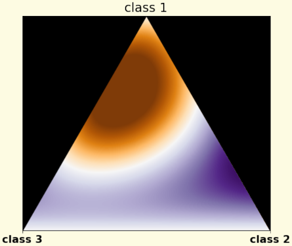
$$\mathbb{L}(\mathbf{w}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^{N} \ln\left(\sum_{j=1}^{K} -\mathbb{I}(y_i = j) \cdot c_j(\mathbf{p}_i; \mathbf{w}, \mathbf{b})\right)$$

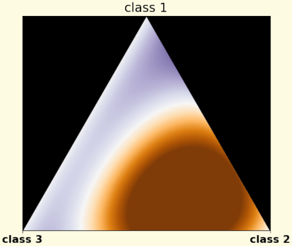
Vector Scaling

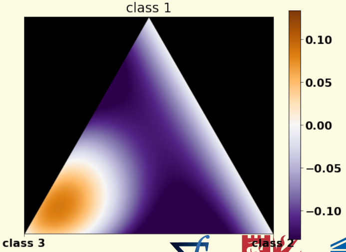












Matrix Scaling

Parameters:

Matrix: $(\mathbf{w}_1, \cdots, \mathbf{w}_K), \mathbf{w}_j \in \mathbb{R}^K$

Intercept: $\mathbf{b} \in \mathbb{R}^K$

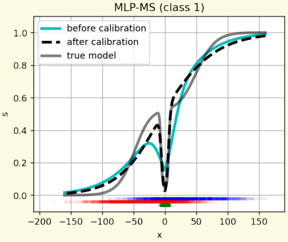
Predictive Function:

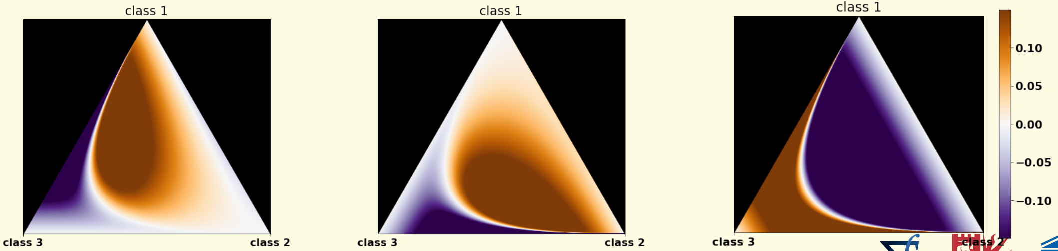
$$c_j(\mathbf{p}; \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{b}) = \frac{\exp\left(-\mathbf{w}_j^T \text{logit}(\mathbf{p}) - b_j\right)}{\sum_{j=1}^K \exp\left(-\mathbf{w}_j^T \text{logit}(\mathbf{p}) - b_j\right)}$$

Objective Function:

$$\mathbb{L}(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^N \ln\left(\sum_{j=1}^K -\mathbb{I}(y_i = j) \cdot c_j(\mathbf{p}_i; \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{b})\right)$$

Matrix Scaling





Dirichlet Calibration

Parameters:

Coefficients: $(\mathbf{w}_1, \cdots, \mathbf{w}_K), \mathbf{w}_j \in \mathbb{R}^K$

Intercept: $\mathbf{b} \in \mathbb{R}^K$

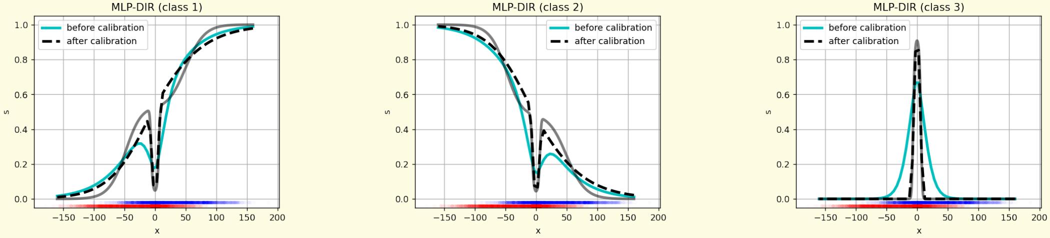
Predictive Function:

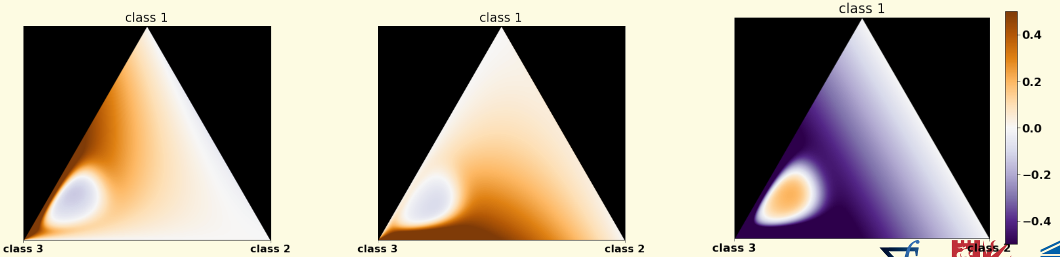
$$c_j(\mathbf{p}; \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{b}) = \frac{\exp(-b_j - \mathbf{w}_j^T \ln \mathbf{p})}{\sum_{j=1}^K \exp(-b_j - \mathbf{w}_j^T \ln \mathbf{p})}$$

Objective Function:

$$\mathbb{L}(\mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^N \ln\Big(\sum_{j=1}^K -\mathbb{I}(y_i = j) \cdot c_j(\mathbf{p}_i; \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{b}) \Big)$$

Dirichlet Calibration





Notable Mentions

- P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning.
  In *29th AAAI Conference on Artificial Intelligence*, feb 2015
- D. Milios, P. Michiardi, L. Rosasco, and M. Filippone. Dirichlet-based Gaussian Processes for Large-scale Calibrated Classification.
  In *Advances in Neural Information Processing Systems (NIPS'18)*, pages 6005–6015, 2018

Notable Mentions

- M.-L. Allikivi and M. Kull. Non-parametric Bayesian Isotonic Calibration: Fighting Over-confidence in Binary Classification.
  In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'19)*, pages 68–85, 2019
- A. Kumar, S. Sarawagi, and U. Jain. Trainable Calibration Measures For Neural Networks From Kernel Mean Embeddings.
  In *35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research*, pages 2805–2814, 2018
- J. Wenger, H. Kjellström, and R. Triebel. Non-parametric calibration for classification.
  In *International Conference on Artificial Intelligence and Statistics*, pages 178–190. PMLR, 2020

Notable Mentions

- V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression.
  In *35th International Conference on Machine Learning, ICML 2018*, volume 6, pages 4369–4377. International Machine Learning Society (IMLS), 2018
- H. Song, T. Diethe, M. Kull, and P. Flach. Distribution calibration for regression.
  volume 97 of *Proceedings of Machine Learning Research*, pages 5897–5906, Long Beach, California, USA, 09–15 Jun 2019. PMLR
- D. Widmann, F. Lindsten, and D. Zachariah. Calibration tests in multi-class classification: A unifying framework.
  In *Advances in Neural Information Processing Systems*, pages 12257–12267, 2019

Table of Contents

Regularisation and Bayesian Treatments

- Typical approaches are generally easy to adopt: L_0, L_1, L_2, lasso, ridge, as well as common Bayesian inference with certain priors on the parameters.
- In Dirichlet calibration, the authors proposed an off-diagonal L2 regularisation approach for Dirichlet calibration, which improves the generalisation of calibration for deep nets by limiting the pair-wise interaction among different classes.

Table of Contents

Implementation

There are also some common practices when implementing a calibration approach:

- Have multiple inner folds to train the base model and calibrators separately
- Approaches including Beta, Dirichlet, and matrix scaling can be easily trained with existing logistic regression implementations.
- For calibrators with a convex loss, when the number of data points and classes is manageable, explicit Newton approaches is generally better than stochastic optimisation.

Table of Contents

Lessons learned

To select a suitable calibrator, consider the following:

- Do you care about the entire probability vector or just about a single class? (the latter $\rightarrow$ binary approaches)
- Do you have a large calibration set? (yes $\rightarrow$ non-parametric approaches)
- Do you have a small calibration set? (yes $\rightarrow$ consider regularisation)
- Are you only interested in certain probability values? (yes $\rightarrow$ binning approaches)
- Any other questions?

What happens next

16.50 - Miquel Perello-Nieto: Hands-on session

17.30 - Peter Flach, Hao Song: Advanced topics and conclusion

Cost curves; calibrating for F-score; regressor calibration

All times in CEST.

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.

Calibrators

Classifier Calibration Tutorial
ECML PKDD 2020

Hao Song
hao.song@bristol.ac.uk
classifier-calibration.github.io/

Departamento de ESTATÍSTICA

University of BRISTOL

UNIVERSITY OF TARTU