Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration
Meelis Kull, Miquel Perello Nieto, Markus Kängsep, Telmo Silva Filho, Hao Song, Peter Flach

Class probabilities predicted by most multiclass classifiers are uncalibrated, often tending towards over-confidence. With neural networks, calibration can be improved by temperature scaling, a method to learn a single corrective multiplicative factor for inputs to the last softmax layer. On non-neural models the existing methods apply binary calibration in a pairwise or one-vs-rest fashion. We propose a natively multiclass calibration method applicable to classifiers from any model class, derived from Dirichlet distributions and generalising the beta calibration method from binary classification. It is easily implemented with neural nets since it is equivalent to log-transforming the uncalibrated probabilities, followed by one linear layer and softmax.

Contributions
➢ Dirichlet calibration:
   • Parametric multiclass calibration method
   • General-purpose (acts in class probability space)
   • Easy to implement as a neural layer or as multinomial logistic regression on log-transformed class probabilities

➢ ODIR (Off-Diagonal and Intercept Regularisation):
   • A new regularization method for Dirichlet calibration and matrix scaling

➢ Clarifications in calibration evaluation of multiclass classifiers.

Is my multiclass classifier calibrated?

Multiclass classifier:
\[ p(Y = c | X) = (p_1(X), \ldots, p_n(X)) \in \Delta_n \subset \mathbb{R}^n \]

Actual class:
\[ Y \in \{c_1, \ldots, c_d\} \]

Multiclass-calibrated:
\[ P(Y = c | X) = p(Y = c) \]

Classwise-calibrated:
\[ P(Y = c | X) = \logit^{-1}(p(Y = c)) \]

Confidence-calibrated:
\[ P(Y = c | X) = \logit^{-1}(p(Y = c)) \text{ for } c \in \{c_1, \ldots, c_d\} \]

How often are classifiers classwise-calibrated?

Example on a neural network

How to calibrate a multiclass classifier:
1. Choose logit-space or class probability space
2. Choose a calibration map family
   • Matrix Scaling
   • Temperature Scaling
   • Diagonal Dirichlet Calibration
3. Fit the calibration map
   by minimising cross-entropy on the validation data and optionally regularise (L2 or ODIR)

Derivations of calibration maps
\[ \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{| \Delta_a |} \sum_{c \in \Delta_a} \exp(\sum_{i} \log p_i(x_i) - s_i) \]

Interpretation of Dirichlet calibration maps

Non-neural experiment

21 UCI datasets and 11 sklearn classifiers
- 231 settings

Deep neural networks experiment

14 CNNs for CIFAR-10, CIFAR-100, SVHN 14
Calibration maps trained on 5000 validation instances with 5-fold-cross-validation

Conclusion:
➢ Dirichlet calibration:
   • New parametric general-purpose multiclass calibration method
   • Natural extension of binary Beta calibration
   • Easy to implement as a neural layer or as multinomial logistic regression on log-transformed class probabilities
   • Best or tied best average rank across 21 datasets x 11 classifiers
➢ ODIR regularisation:
   • Matrix scaling with ODIR is tied best in log-loss
   • Dirichlet with ODIR is tied best in error rate