Part III
Model Calibration

Adrian Galdran, MSC Research Fellow
Universitat Pompeu Fabra, Barcelona, Spain
University of Adelaide, Australia

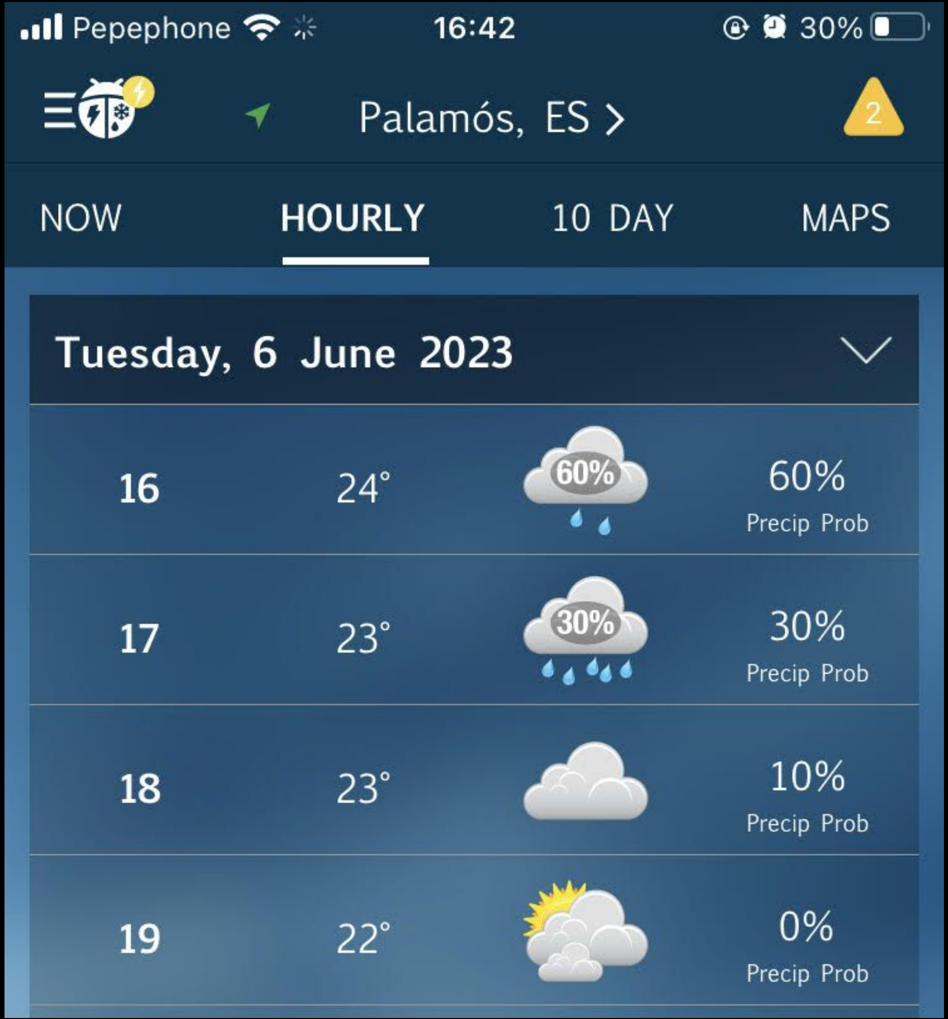
Meritxell Riera i Marin, Researcher
Sycai Technologies, Barcelona, Spain
Universitat Pompeu Fabra, Barcelona, Spain

MICCAI 2023 Vancouver CANADA

Contents

1. Understanding Calibration
2. Measuring Calibration
3. Improving Calibration
4. Practical Hands-On Session

MICCAI 2023 Vancouver CANADA


Pepephone
16:42
30%
Palamós, ES >
2
NOW
HOURLY
10 DAY
MAPS
Tuesday, 6 June 2023
16
24°
60%
60%
Precip Prob
17
23°
30%
30%
Precip Prob
18
23°
10%
Precip Prob
19
22°
0%
Precip Prob

One of the most important tests of a forecast — I would argue that it is the single most important one — is called calibration. Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated. If it wound up raining just 20 percent of the time instead, or 60 percent of the time, they weren't.

Nate Silver, The Signal and the Noise: Why So Many Predictions Fail – but Some Don't

MICCAI 2023 Vancouver CANADA

1. Understanding Calibration

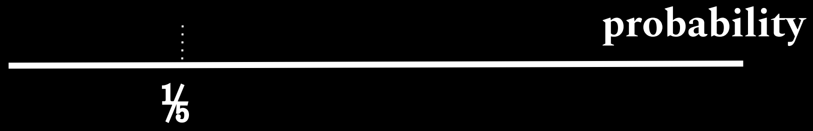
p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1

p	y
¾	0
¾	1
¾	1
¾	1
1	1

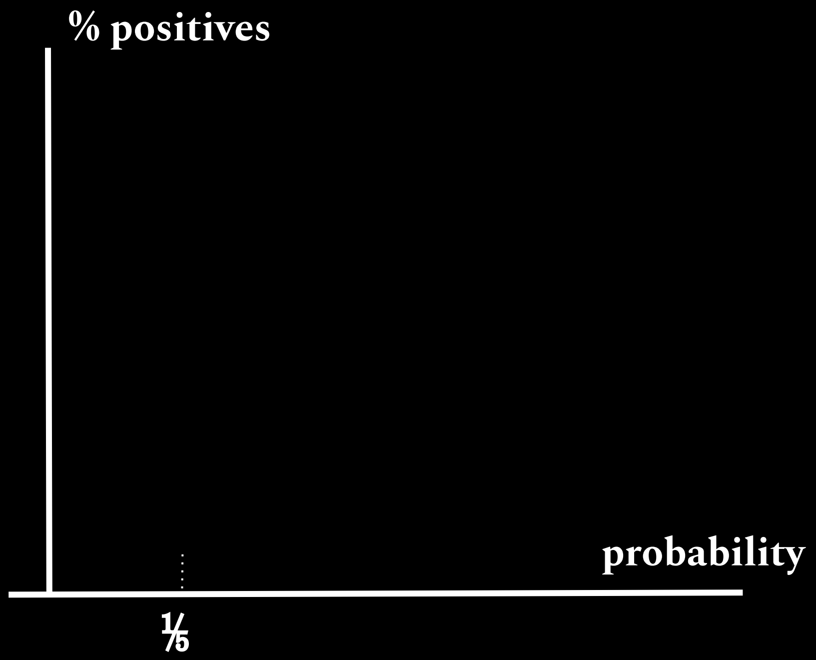
1. Understanding Calibration

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



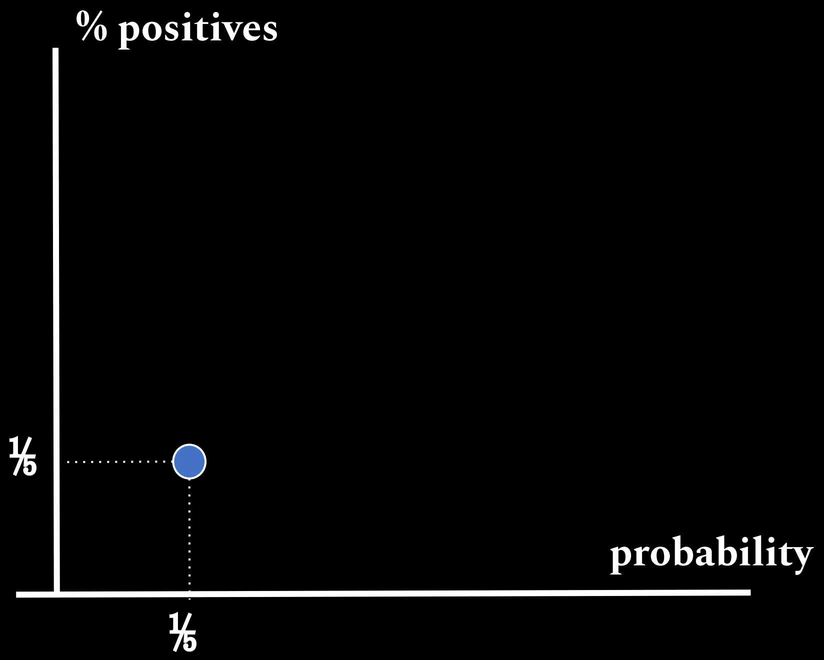
1. Understanding Calibration

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



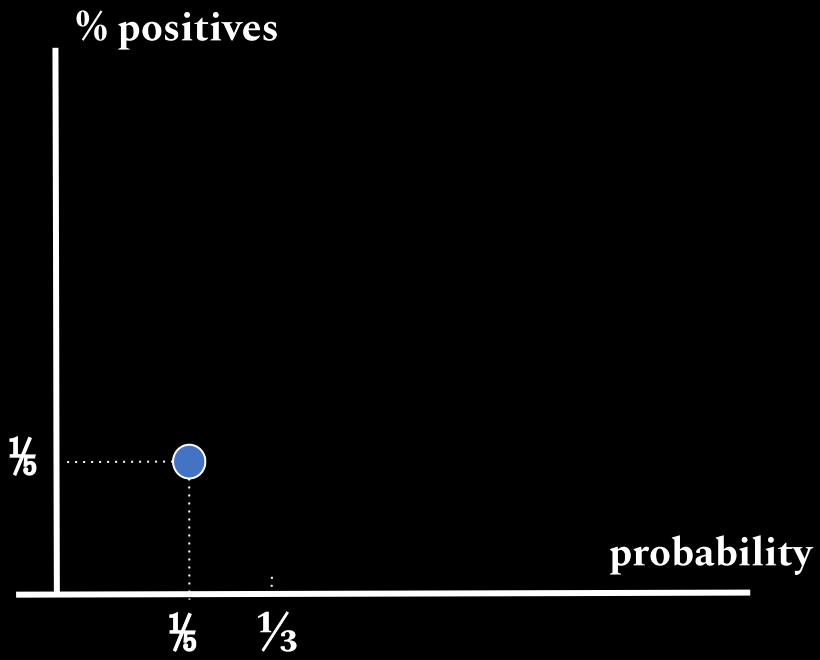
1. Understanding Calibration

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



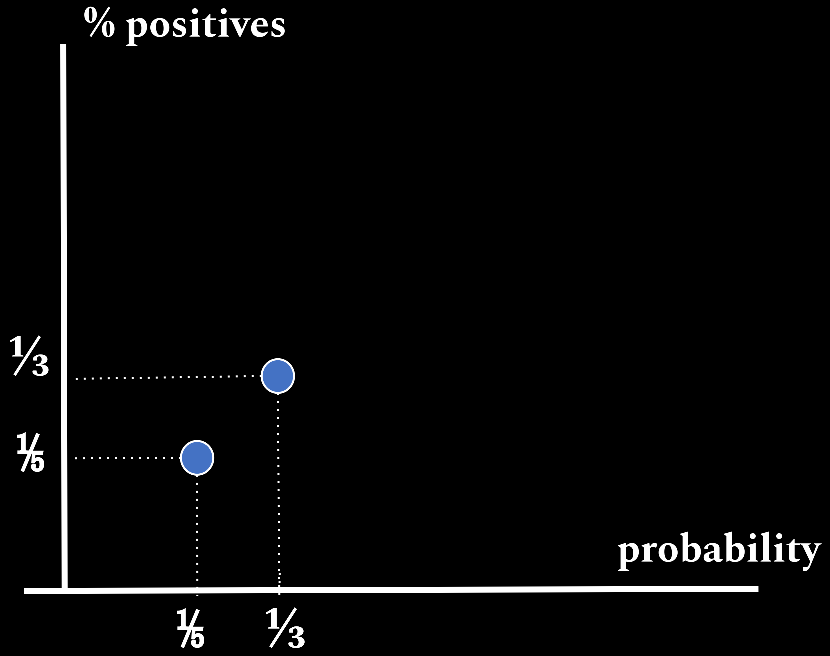
1. Understanding Calibration

p	y
⅓	0
⅓	0
⅓	1



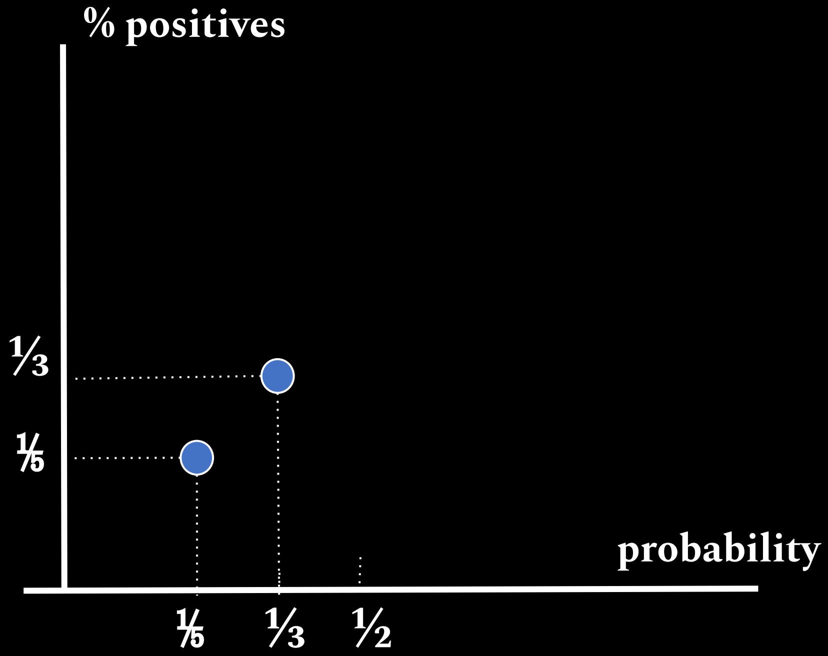
1. Understanding Calibration

p	y
⅓	0
⅓	0
⅓	1



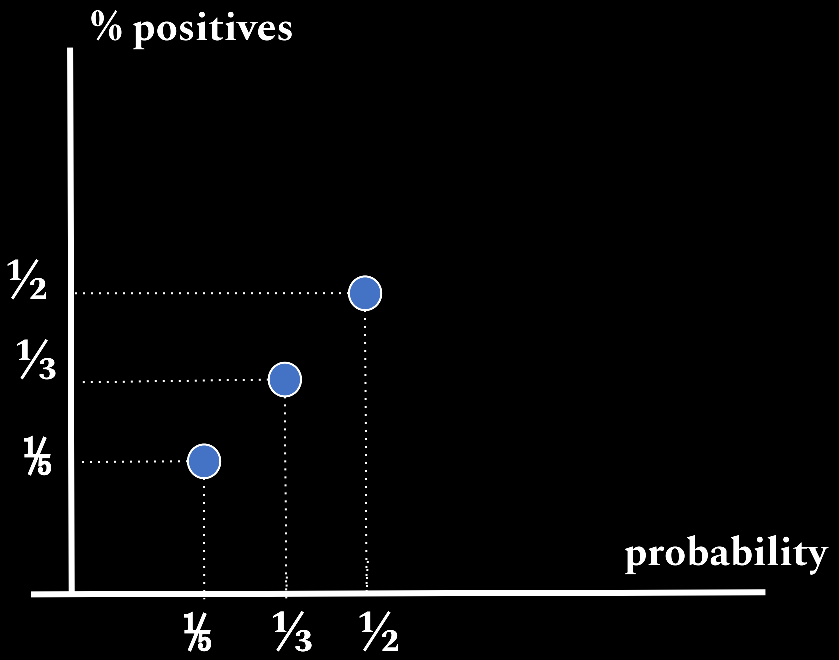
1. Understanding Calibration

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1



1. Understanding Calibration

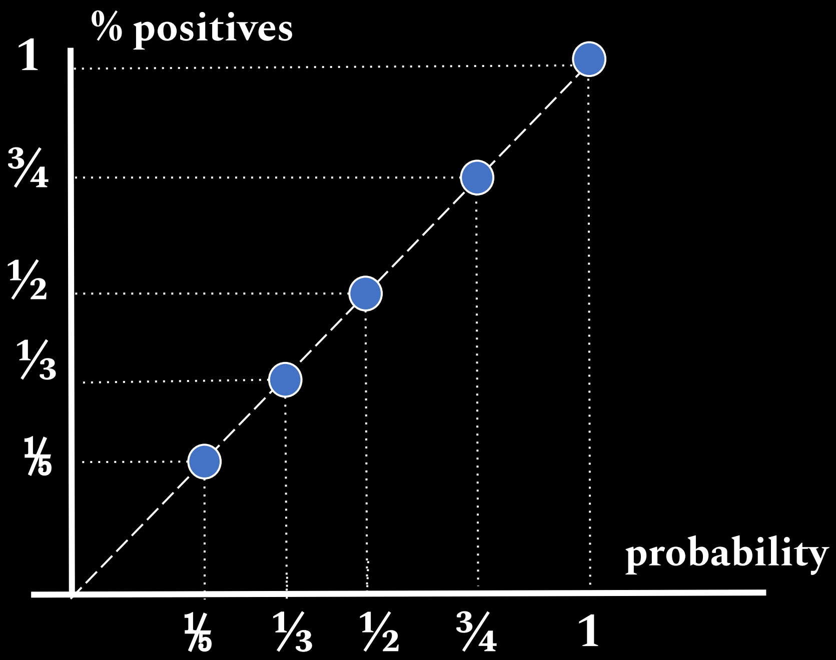
p	y
⅓	0
⅓	0
⅓	1
½	0
½	1



MICCAI 2023 Vancouver CANADA

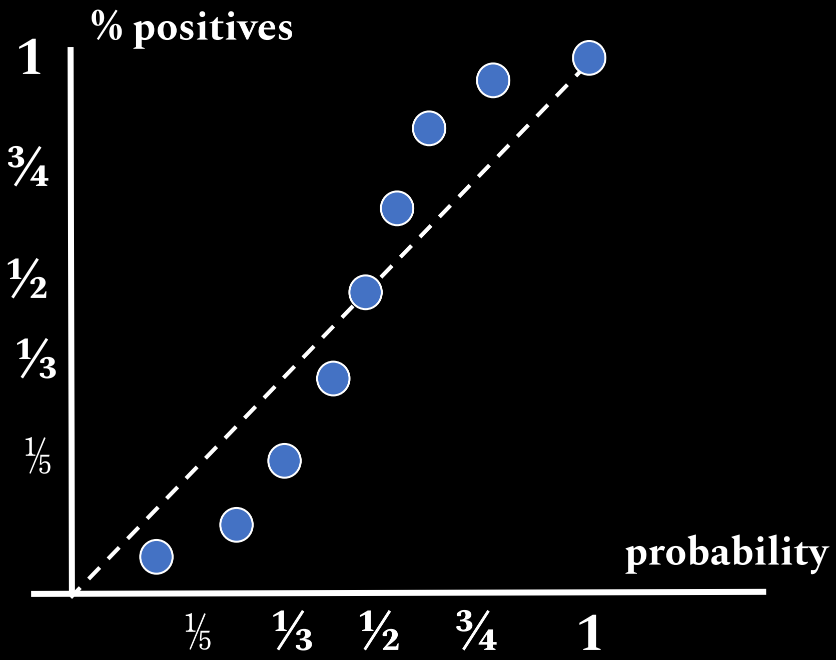
1. Understanding Calibration

p	y
¾	0
¾	1
¾	1
¾	1
1	1



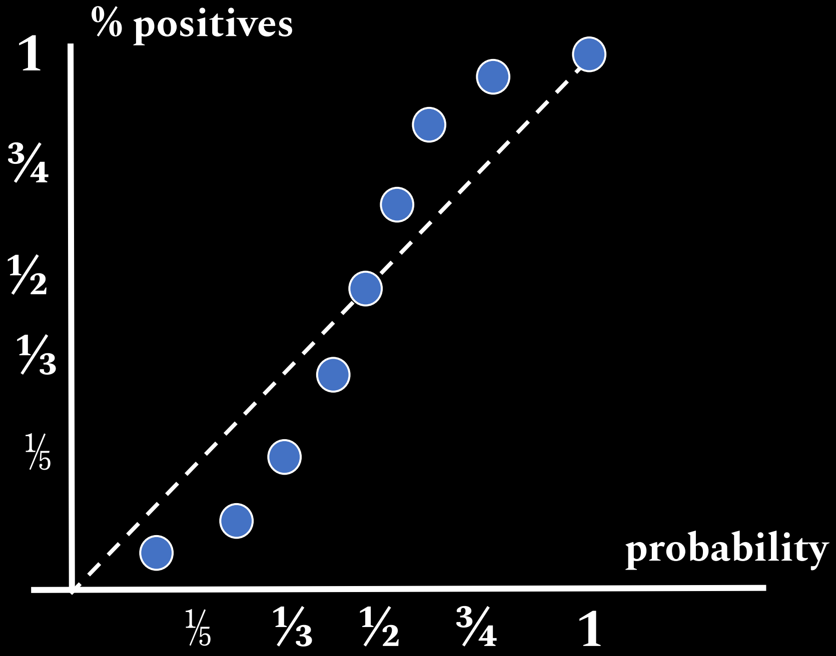
1. Understanding Calibration

QUESTION:
Are these predictions
<u>under</u>-confident
or
<u>over</u>-confident?



1. Understanding Calibration

QUESTION:
Are these predictions
under-confident
or
over-confident?



2. Measuring Calibration

- **Reliability Plots**

  Not enough items with a given confidence to estimate population statistics decently:

  model predicts with p=0.2 → 20% positives"

  What if you only have 2 items predicted with p=0.2?

  We can group predictions in bins, and plot them against y=x.

- **Expected Calibration Error**

  The average of gaps across bins, weighted by bin population:

$$\mathrm{ECE} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|B_i|}\left|prob(B_i) - pos(B_i)\right|$$

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	2

MICCAI 2023 Vancouver CANADA

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	2

MICCAI 2023 Vancouver CANADA

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	$(\hat{y}, c)$	y
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	2

p	$(\hat{y}, c)$	y
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	3

p	$(\hat{y}, c)$	y
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	1

2. Measuring Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Full-calibration: consider the whole probability vector.
Class-wise calibration: only consider marginal probabilities.
Confidence calibration: only consider highest probability.

p	$(\hat{y}, c)$	y
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	2

p	$(\hat{y}, c)$	y
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	3

p	$(\hat{y}, c)$	y
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	2

2. Measuring Calibration

- **Expected Full Calibration Error**

$$\text{full-ECE} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathbb{B}_i|}\,\|\,prob(\mathbb{B}_i) - true(\mathbb{B}_i)\,\|$$

- **Expected Class-Wise Calibration Error**

$$\text{cw-ECE} = \frac{1}{K}\sum_{k=1}^{K}\text{bin-ECE}_k \quad [\text{one-vs-rest}]$$

- **Expected Confidence–Calibration Error**

$$\text{conf-ECE} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|B_i|}\,|conf(B_i) - acc(B_i)|$$

MICCAI 2023 Vancouver CANADA

2. Measuring Calibration

- **Alternative Calibration Measures: Proper Scoring Rules**

p	y
(⅓+2ε, ⅓-ε, ⅓-ε)	1
(⅓-ε, ⅓+2ε, ⅓-ε)	1
(⅓-ε, ⅓-ε, ⅓+2ε)	2

p	y
(⅓-ε, ⅓+2ε, ⅓-ε)	2
(⅓-ε, ⅓+2ε, ⅓-ε)	3
(⅓-ε, ⅓-ε, ⅓+2ε)	3

2. Measuring Calibration

- **Alternative Calibration Measures: Proper Scoring Rules**

p	y
(⅓+2ε, ⅓-ε, ⅓-ε)	1
(⅓-ε, ⅓+2ε, ⅓-ε)	1
(⅓-ε, ⅓-ε, ⅓+2ε)	2

p	y
(⅓-ε, ⅓+2ε, ⅓-ε)	2
(⅓-ε, ⅓+2ε, ⅓-ε)	3
(⅓-ε, ⅓-ε, ⅓+2ε)	3

**This classifier predicts a random class with full uncertainty.
It always has a confidence of ~⅓, and it has an accuracy of ⅓.
Therefore it is perfectly confidence-calibrated, but useless.**

2. Measuring Calibration

- **Alternative Calibration Measures: Proper Scoring Rules**

p	y
(⅔, 0, ⅓)	1
(0, ⅓, ⅔)	1
(⅓, ⅔, 0)	2

p	y
(0, ⅓, ⅔)	2
(⅓, 0, ⅔)	3
(0, ⅓, ⅔)	3

MICCAI 2023 Vancouver CANADA

2. Measuring Calibration

- **Alternative Calibration Measures: Proper Scoring Rules**

p	y
(⅔, 0, ⅓)	1
(0, ⅓, ⅔)	1
(⅓, ⅔, 0)	2

p	y
(0, ⅓, ⅔)	2
(⅓, 0, ⅔)	3
(0, ⅓, ⅔)	3

This classifier always predicts with ~⅔ confidence. Also, it has an accuracy of ⅔. It is perfectly confidence-calibrated, but it has more discrimination ability than random guessing.

2. Measuring Calibration

- **Alternative Calibration Measures: Proper Scoring Rules**

p	y
(1, 0, 0)	1
(1, 0, 0)	1
(0, 1, 0)	2

p	y
(0, 1, 0)	2
(0, 0, 1)	3
(0, 0, 1)	3

2. Measuring Calibration

- **Alternative Calibration Measures: Proper Scoring Rules**

p	y
(1, 0, 0)	1
(1, 0, 0)	1
(0, 1, 0)	2

p	y
(0, 1, 0)	2
(0, 0, 1)	3
(0, 0, 1)	3

This is a god-like classifier. It is always 100% confident, and always right. It is full-calibrated and perfectly discriminative.

PSRs are a tool for measuring calibration & discrimination jointly.

2. Measuring Calibration

- **Proper Scoring Rules**

Measure discrimination+calibration at individual item level

Most popular: Brier Score, Negative Log-Likelihood

$$\mathrm{Brier}(\mathbf{p},\mathbf{y}) = \|\mathbf{p} - \mathbf{y}\|_2^2 \qquad \mathrm{NLL}(\mathbf{p}, \mathbf{y}) = -\log(\mathbf{p}_y)$$

Example: $y = \mathbf{3},\ \mathbf{y} = (0,\ 0,\ 1),\ \mathbf{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right),\ \mathbf{q} = \left(0, \frac{1}{3}, \frac{2}{3}\right)$

$$\mathrm{Brier}(\mathbf{p}, \mathbf{y}) = 2/3 \qquad \mathrm{Brier}(\mathbf{q}, \mathbf{y}) = 2/9 \qquad \mathrm{Brier}(\mathbf{y}, \mathbf{y}) = 0$$

$$\mathrm{NLL}(\mathbf{p}, \mathbf{y}) \approx 0.477 \qquad \mathrm{NLL}(\mathbf{q}, \mathbf{y}) \approx 0.176 \qquad \mathrm{NLL}(\mathbf{y}, \mathbf{y}) = 0$$

Note that a fully uncertain prediction (p) does not score well.

3. Improving Calibration

- **Post-Training Calibration**

Classic methods: Platt Scaling & Isotonic Regression:

- Platt: Fits a logistic regression model using validation set.
- Isotonic: Fits a monotonic piecewise constant mapping, optimizing bins to maximize calibration.

Both designed for binary models

Temperature Scaling: Uses a validation set to learn a scalar T dividing logits before applying softmax and tempers their value:

$$p_{\mathrm{j}} = \frac{\mathrm{e}^{\mathrm{z_j}}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{e}^{\mathrm{z_k}}} \longmapsto p_j = \frac{\mathrm{e}^{(\mathrm{z_j}/\mathrm{T})}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{e}^{(\mathrm{z_k}/\mathrm{T})}}$$

We will see examples in the hands-on

3. Improving Calibration

- **Model Ensembling**

Ensembling several diverse models can improve calibration. Of course it comes with a computational overhead.

- **Training Time Calibration**

Over-parametrized NNs can keep on learning the training set until they are fully confident, minimizing NLL indefinitely. We can avoid this by regularizing so as to disencourage confidence.

Label Smoothing, MixUp, Focal Loss... Careful of underfitting! Always report also a PSR, not only ECE.

We will see examples in the hands-on

MICCAI 2023 Vancouver CANADA

4. Hands-On

Github repository:
https://github.com/agaldran/calibration_tutorial/

nb_viewer: https://shorturl.at/hyFO3

MICCAI 2023 Vancouver CANADA