Part III
Model Calibration

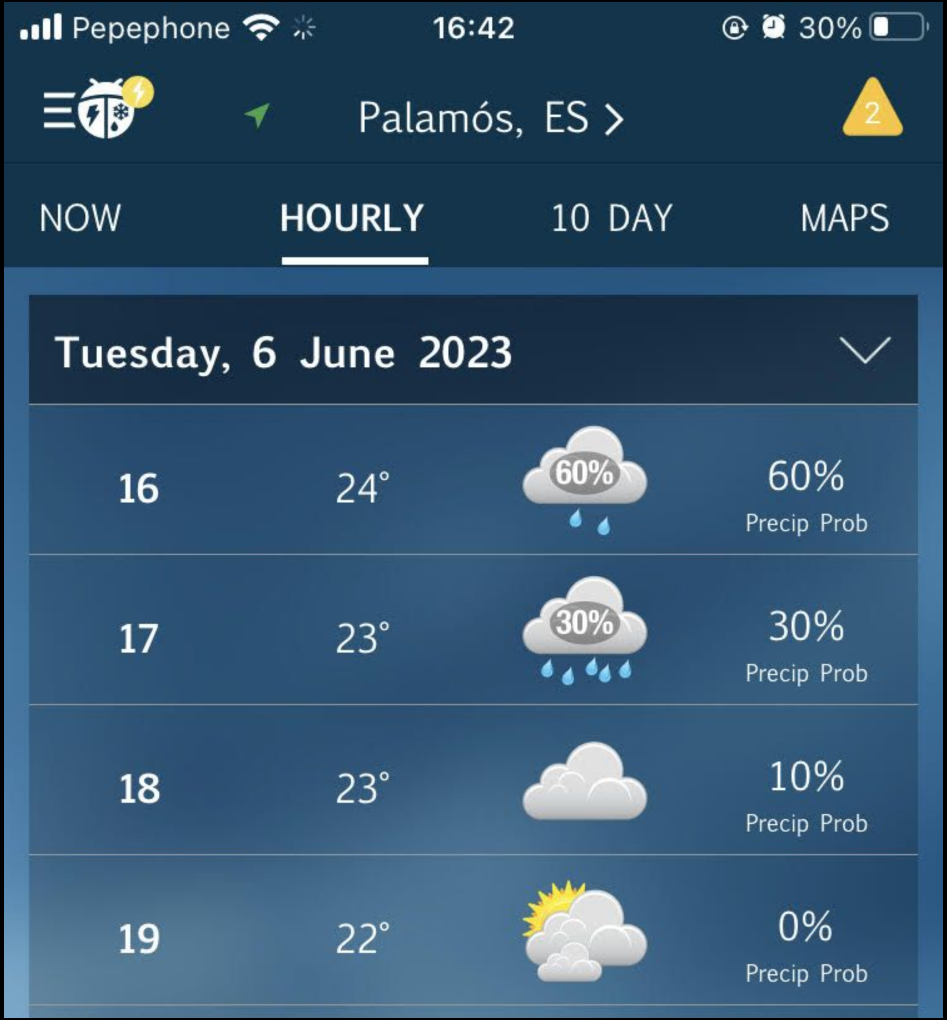
Adrian Galdran, MSC Research Fellow
Universitat Pompeu Fabra, Barcelona, Spain
University of Adelaide, Australia

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

MICCAI2023
Vancouver
CANADA

Contents

1. Calibration: what, when and why?
2. Visualizing & Measuring Calibration
3. Improving Calibration
4. Practical Hands-On Session



Palamós, ES >

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

1. Calibration: what, when and why?

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. Calibration: what, when and why?

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1

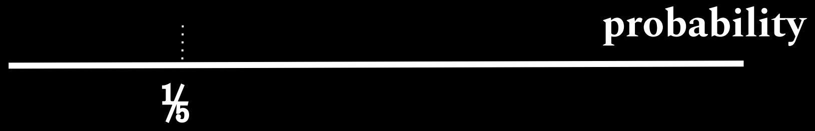
p	y
⅓	0
⅓	0
⅓	1
½	0
½	1

p	y
¾	0
¾	1
¾	1
¾	1
1	1

MICCAI 2023 Vancouver CANADA

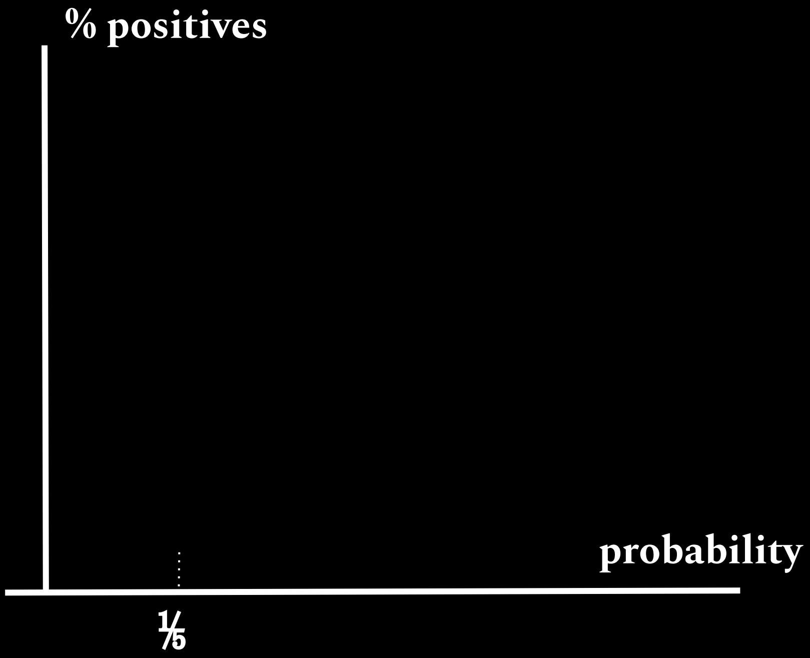
1. Calibration: what, when and why?

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



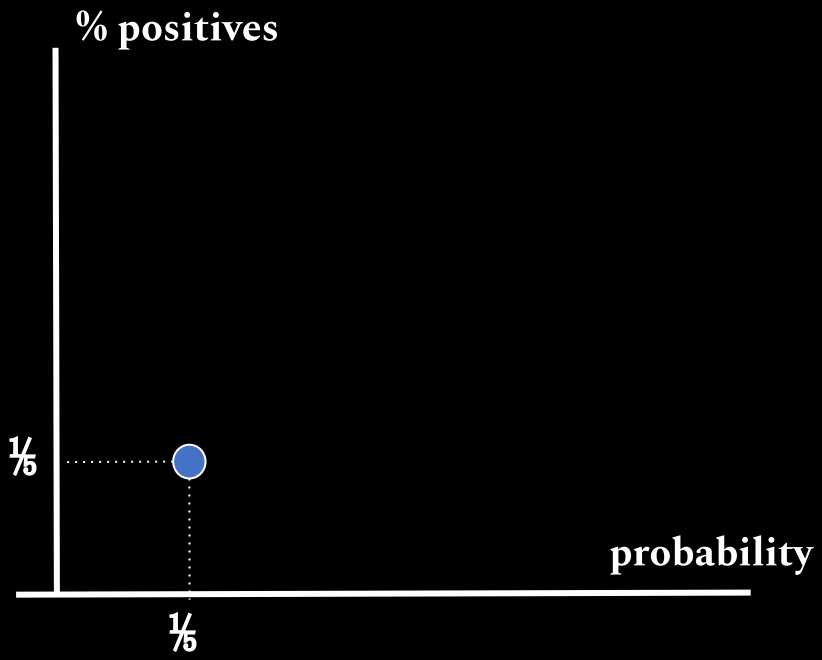
1. Calibration: what, when and why?

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



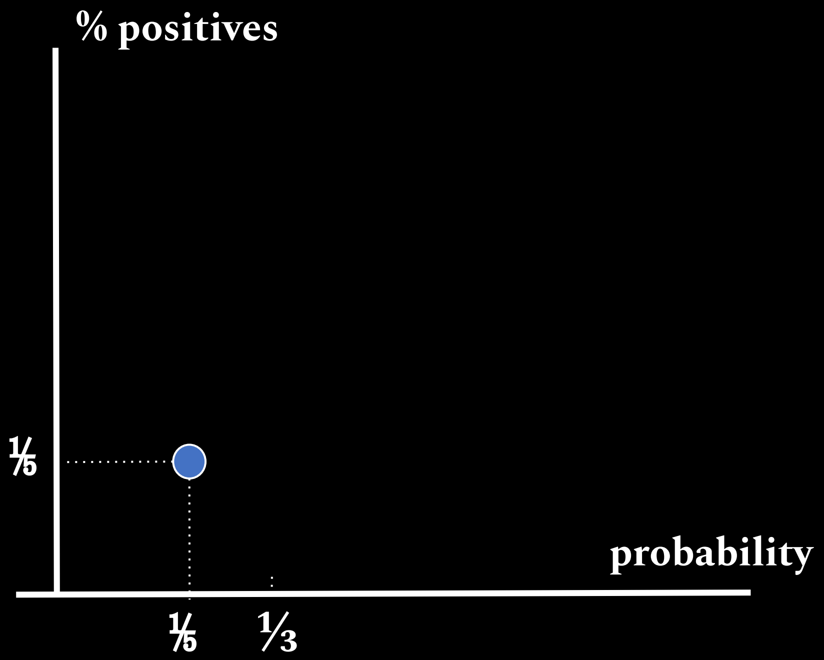
1. Calibration: what, when and why?

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



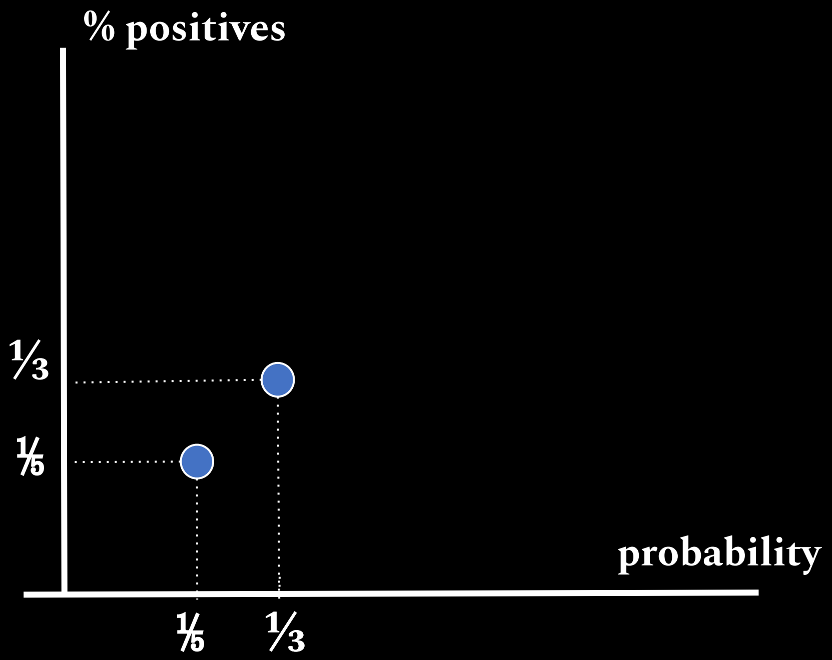
1. Calibration: what, when and why?

p	y
⅓	0
⅓	0
⅓	1



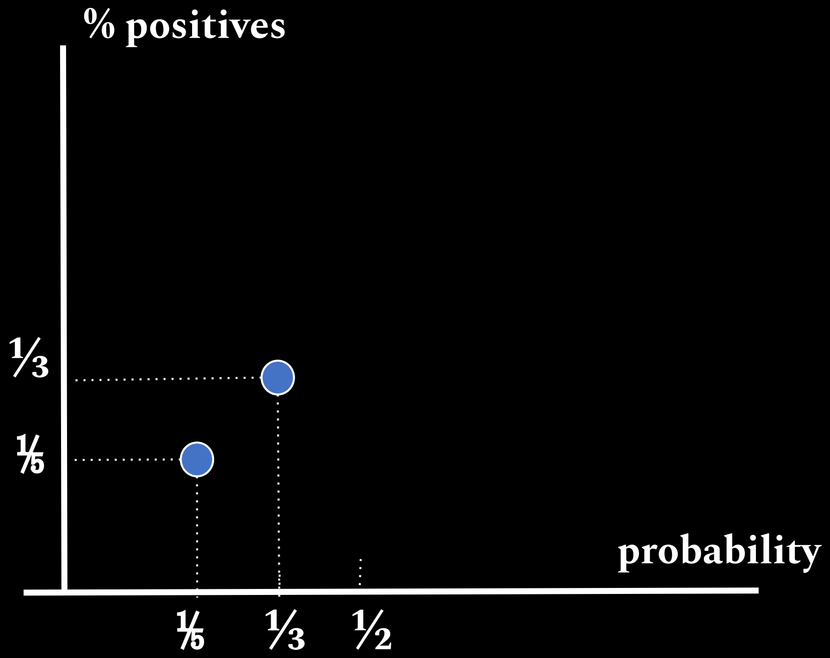
1. Calibration: what, when and why?

p	y
⅓	0
⅓	0
⅓	1



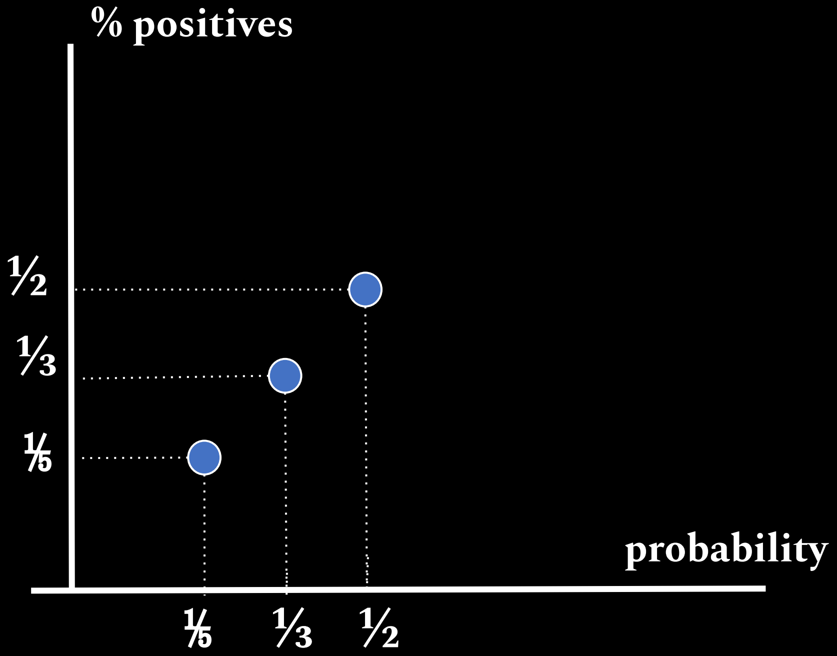
1. Calibration: what, when and why?

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1



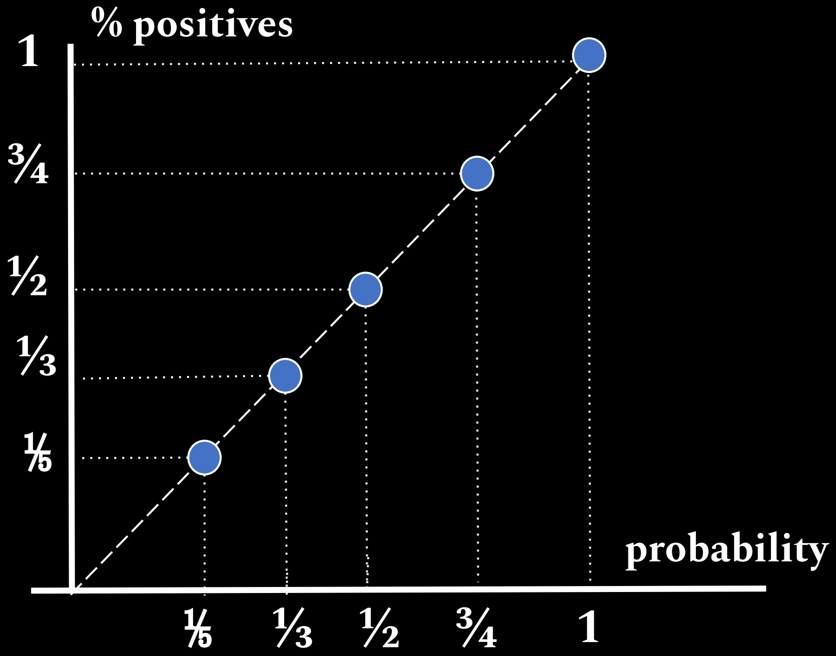
1. Calibration: what, when and why?

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1



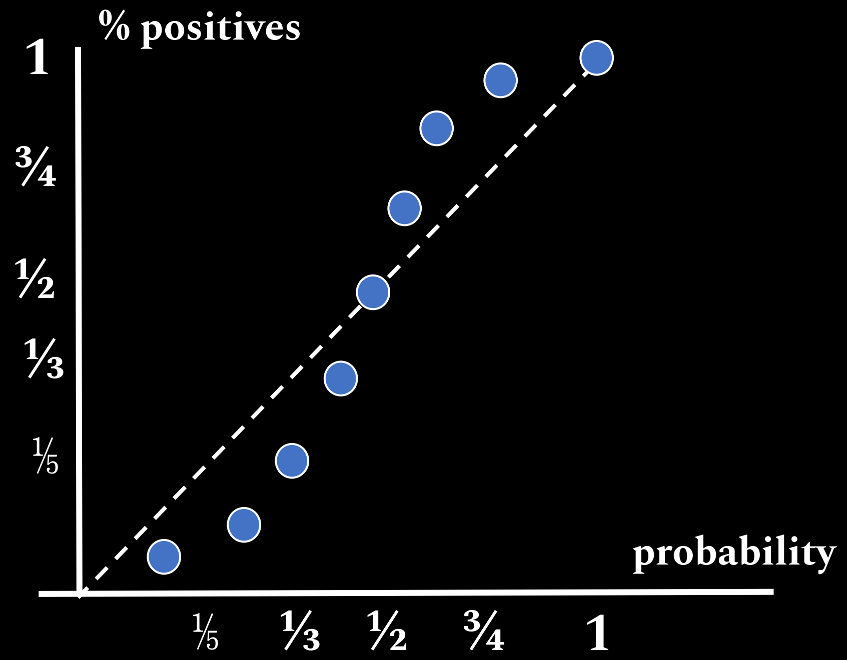
1. Calibration: what, when and why?

p	y
¾	0
¾	1
¾	1
¾	1
1	1



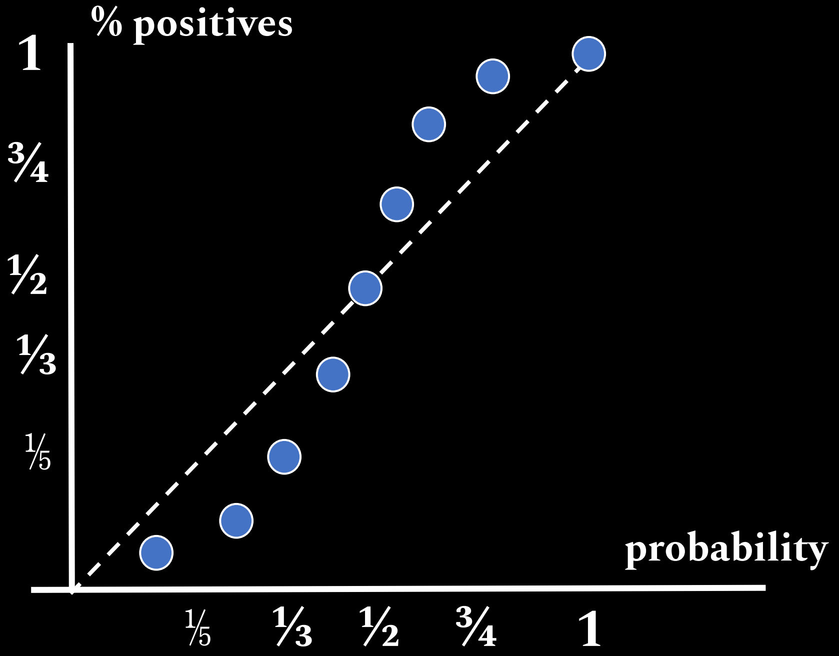
1. Calibration: what, when and why?

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



1. Calibration: what, when and why?

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



2. Visualizing & 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:

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

2. Visualizing & 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. Visualizing & 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. Visualizing & 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

2. Visualizing & 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. Visualizing & 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. Visualizing & 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

2. Visualizing & 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. Visualizing & 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	(ŷ, c)	y
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	2

p	(ŷ, c)	y
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	3

p	(ŷ, c)	y
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	1

2. Visualizing & 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	(ŷ, c)	y
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	2

p	(ŷ, c)	y
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	3

p	(ŷ, c)	y
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	2

2. Visualizing & Measuring Calibration

- **Expected Calibration Error (binary)**

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

- **Expected Class-Wise Calibration Error (multi-class)**

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

- **Expected Confidence–Calibration Error (multi-class)**

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

2. Visualizing & 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. Visualizing & 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 conf-calibrated, but useless.

2. Visualizing & 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

2. Visualizing & 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 conf-calibrated, but it has more discrimination ability than random guessing.

2. Visualizing & 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. Visualizing & 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 totally calibrated and perfectly discriminative.

PSRs are a tool for measuring calibration & discrimination jointly.

MICCAI 2023 Vancouver CANADA

2. Visualizing & Measuring Calibration

- **Proper Scoring Rules**

Measure discrimination+calibration at individual item level

Most popular: Brier Score, Negative Log-Likelihood

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

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

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

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

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

4. Hands-On

https://shorturl.at/hyFO3

1. Calibration: what, when and why?

Connections with previous parts of the tutorial:

- Sources of miscalibration.
- Uncertainty vs Confidence.
- When can misalibration happen?
- Why is it important to address miscalibration?
- Does it only make sense for classification?

- **Why is it important to address miscalibration?**

  Similar to probabilistic weather predictions, decisions derived from high-confidence wrong predictions can result in bad outcomes. Meaningful probabilities can increase trust in automatic systems.

1. Calibration: what, when and why?

- **Uncertainty vs Confidence:**

  While there are many mechanisms to quantify uncertainty, in calibration we focus on the simplest one: interpreting sigmoid/softmax output of a neural net as a probability.

  We refer to this as confidence in this tutorial.

- **Does it only make sense for classification?**

  95% of existing work on calibration has to do with classification, but it makes sense to consider other problems, e.g. segmentation. This is less explored terrain: ideas do not always translate trivially, there is no consensus on evaluation metrics, etc. RecSys!

2. Visualizing & Measuring Calibration

- **Reliability Plots**

These are just the plot we saw before. Careful! We normally won’t have enough items with a given confidence to estimate population statistics decently.

“When the model predicts with p=0.2, it’s 20% accurate”

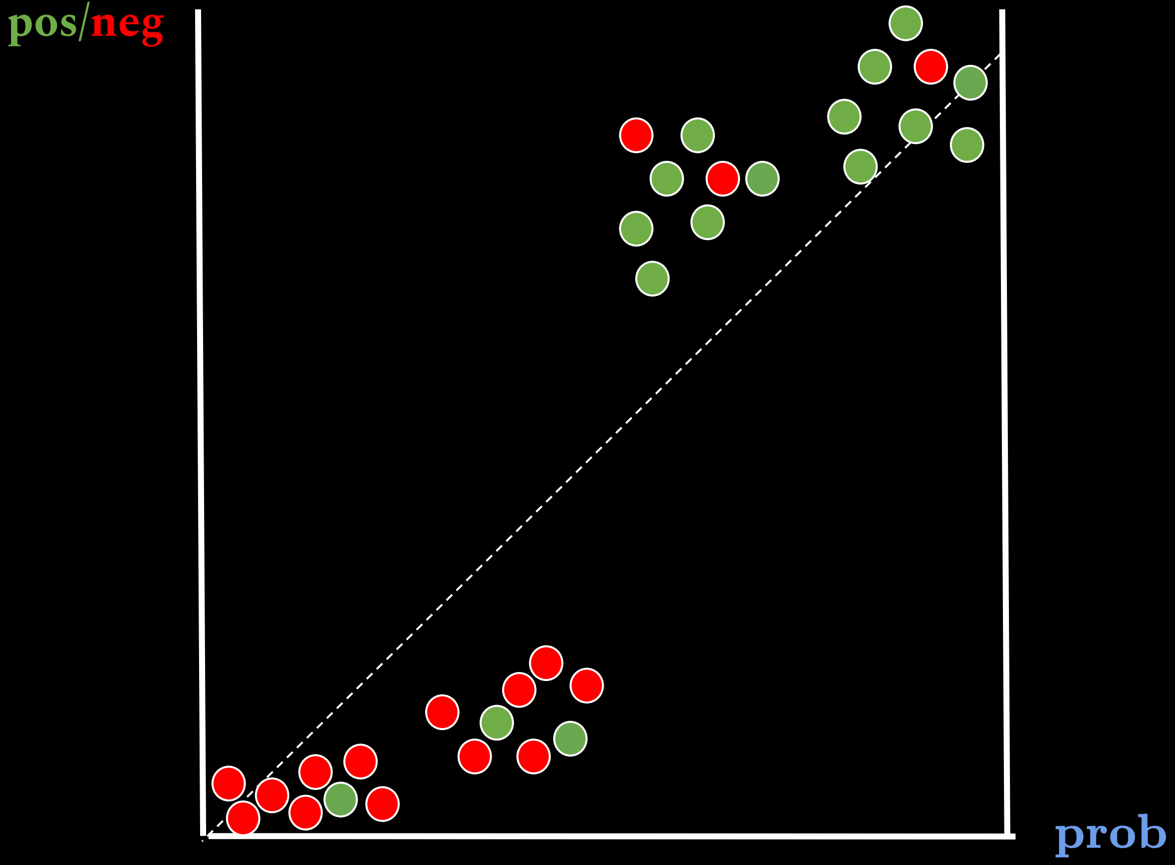
What if you only have 2 items predicted with p=0.2? We need to group predictions, and binning is critical.

- **Expected Calibration Error**

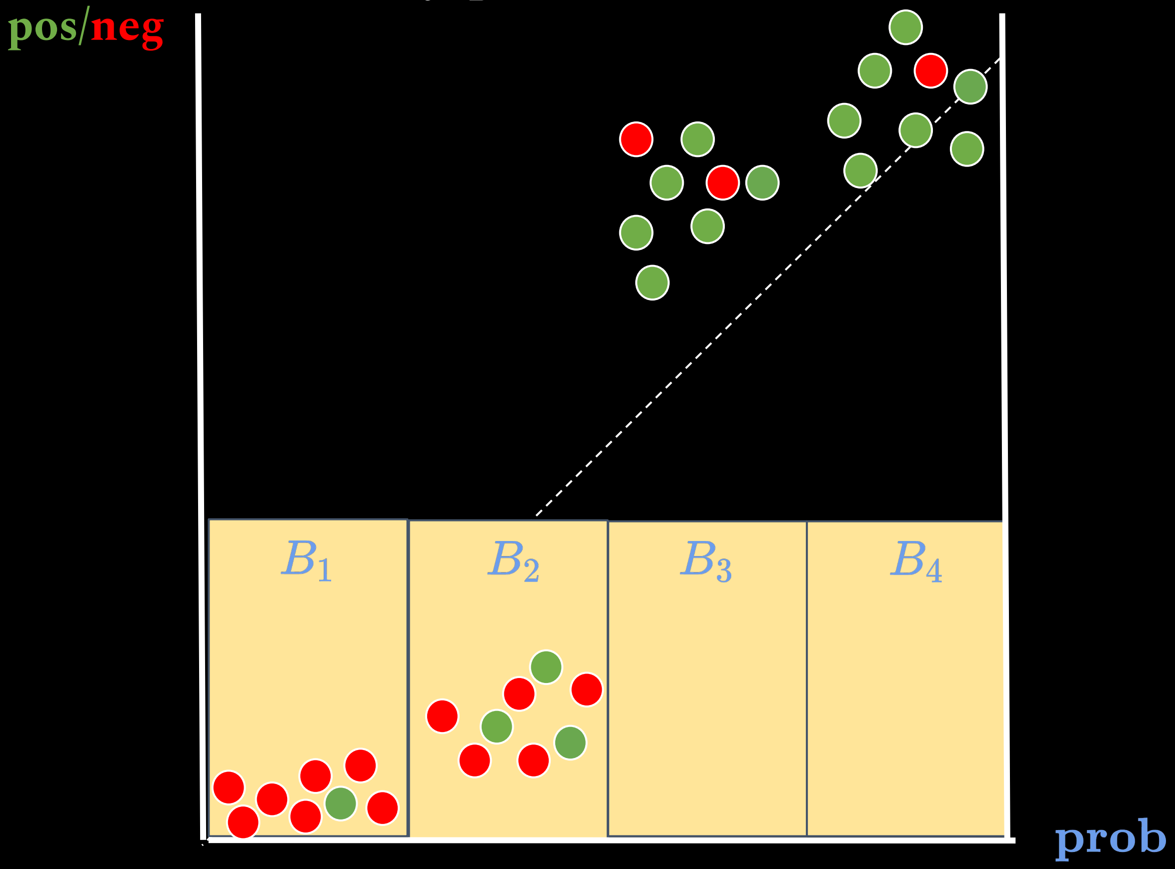
The average gaps across bins, weighted by bin population:

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

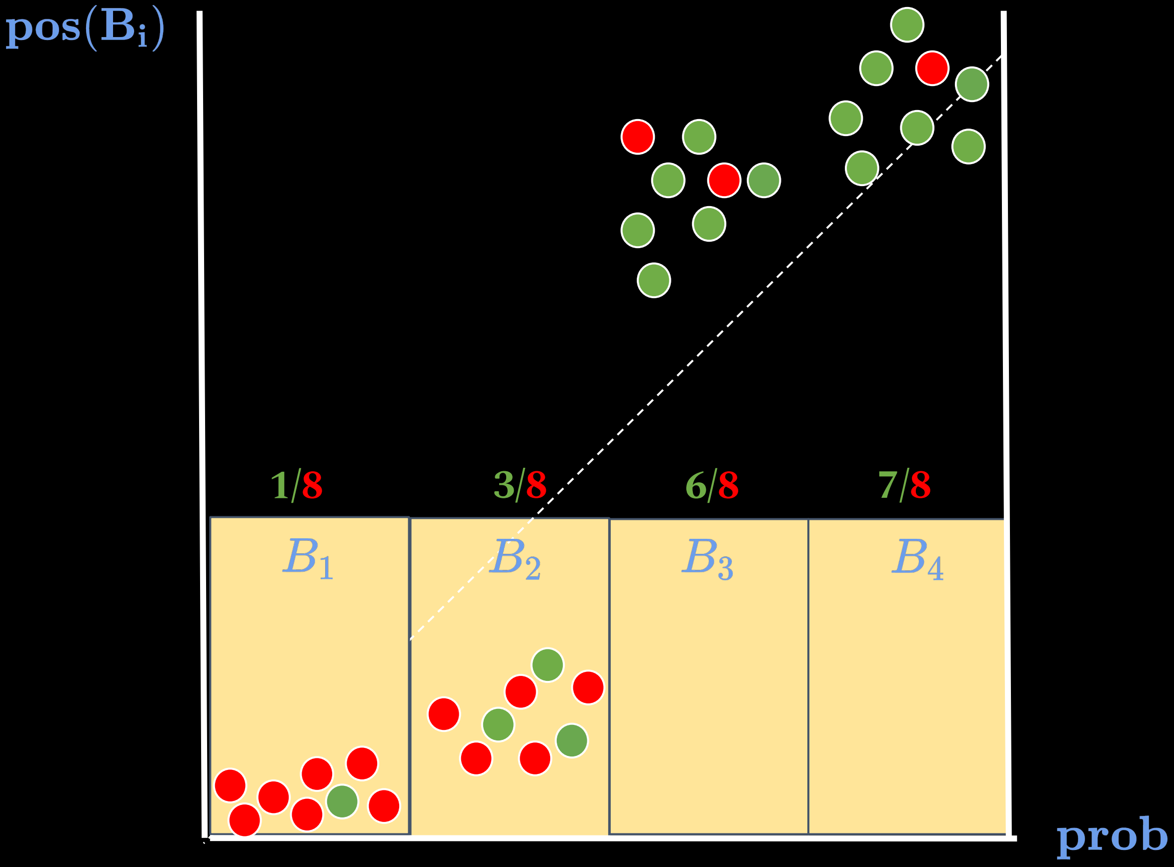
$$\mathrm{ECE} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|B_i|}\left|prob(B_i) - pos(B_i)\right|$$



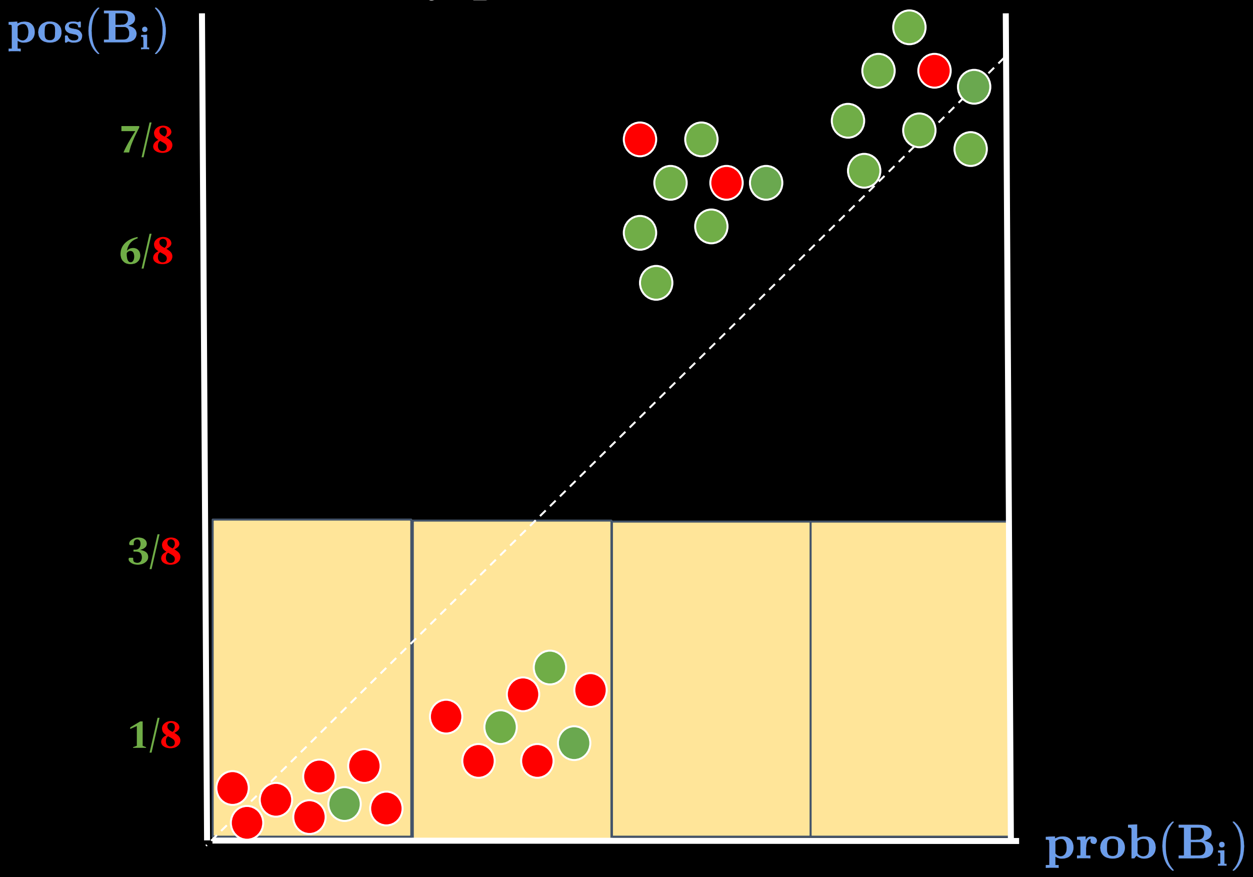
$$\mathrm{ECE} = \frac{1}{4}\sum_{i=1}^{4}\frac{1}{|B_i|}\left|prob(B_i) - pos(B_i)\right|$$



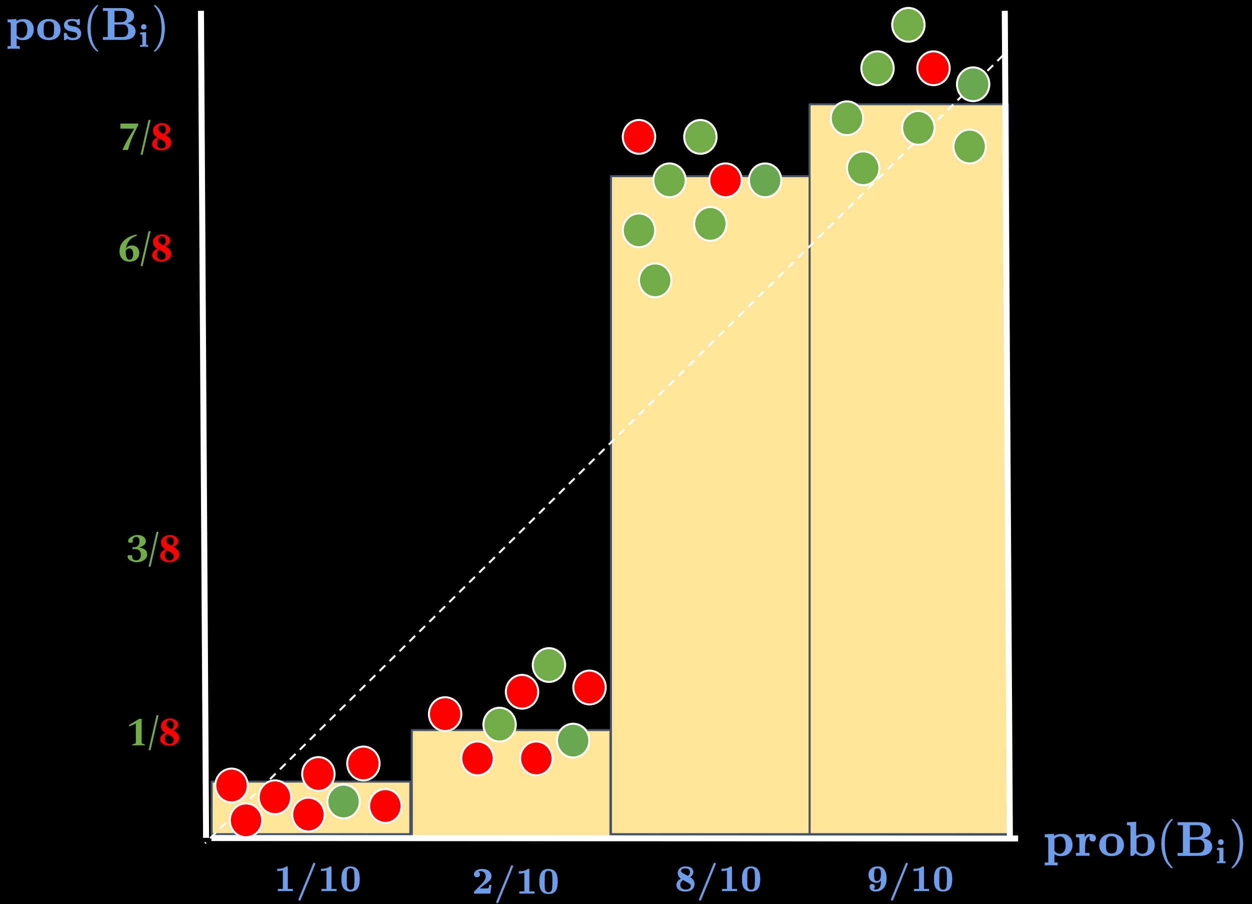
$$\mathrm{ECE} = \frac{1}{4}\sum_{i=1}^{4}\frac{1}{|B_i|}\left|prob(B_i) - pos(B_i)\right|$$



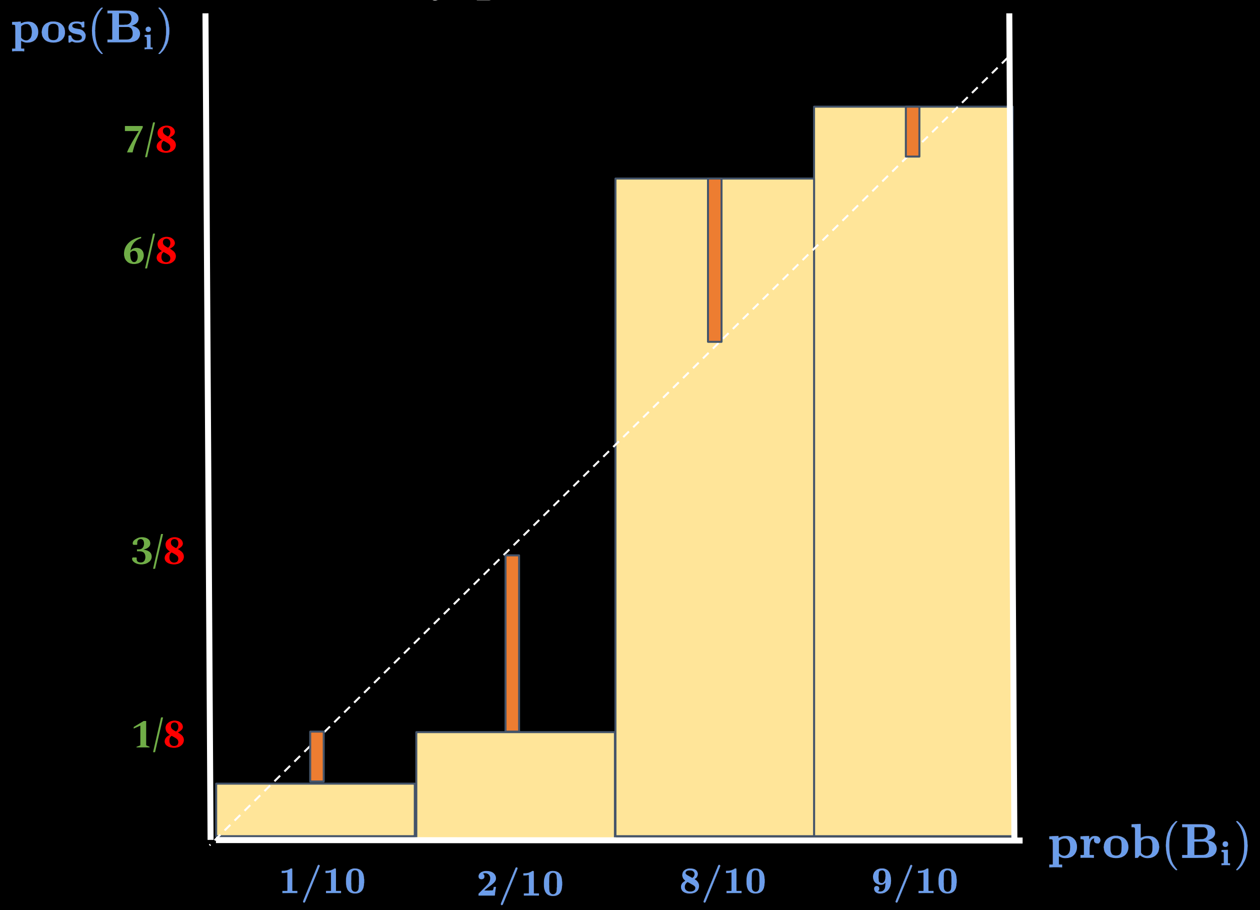
$$\text{ECE} = \frac{1}{4} \sum_{i=1}^{4} \frac{1}{|B_i|} \left|prob(B_i) - pos(B_i)\right|$$



$$\mathrm{ECE} = \frac{1}{4}\sum_{i=1}^{4}\frac{1}{|B_i|}\left|prob(B_i) - pos(B_i)\right|$$



$$\mathrm{ECE} = \frac{1}{4}\sum_{i=1}^{4}\frac{1}{|B_i|}\left|prob(B_i) - pos(B_i)\right|$$



2. Visualizing & Measuring Calibration

- **From Binary to Multi-Class**

How do we generalize the previous ideas? Several ways.

Most obvious: Take maximum softmax probability (confidence), compute accuracy per bin. Compute ECE and draw diagrams in a similar way.

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

Question: Diagram range in x-axis is now [1/K, 1]. Why?

Other: Class-Wise (one vs rest) calibration, Full calibration

We will see examples in the hands-on

2. Visualizing & Measuring Calibration

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

p	y
(3/6, 2/6, 1/6)	1
(3/6, 2/6, 1/6)	1
(3/6, 1/6, 2/6)	1
(1/6, 2/6, 3/6)	1
(1/6, 3/6, 2/6)	1
(1/6, 2/6, 3/6)	1

p	y
(1/6, 4/6, 1/6)	2
(1/6, 4/6, 1/6)	2
(0, 4/6, 2/6)	2
(2/6, 4/6, 0)	2
(4/6, 1/6, 1/6)	2
(0, 2/6, 4/6)	2

p	y
(1/6, 0, 5/6)	3
(1/6, 0, 5/6)	3
(1/6, 0, 5/6)	3
(0, 1/6, 5/6)	3
(0, 1/6, 5/6)	3
(5/6, 1/6, 0)	3

2. Visualizing & Measuring Calibration

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

conf.	y
3/6	1
3/6	1
3/6	1
3/6	1
3/6	1
3/6	1

conf.	y
4/6	2
4/6	2
4/6	2
4/6	2
4/6	2
4/6	2

conf.	y
5/6	3
5/6	3
5/6	3
5/6	3
5/6	3
5/6	3

MICCAI 2023 Vancouver CANADA

2. Visualizing & Measuring Calibration

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

conf.	correct
3/6	1
3/6	1
3/6	1
3/6	0
3/6	0
3/6	0

conf.	correct
4/6	1
4/6	1
4/6	1
4/6	1
4/6	0
4/6	0

conf.	correct
5/6	1
5/6	1
5/6	1
5/6	1
5/6	1
5/6	0

2. Visualizing & Measuring Calibration

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

conf.	acc.
3/6	3/6

conf.	correct
4/6	4/6

conf.	correct
5/6	5/6

So ECE=0 – Note: This classifier is Class-wise calibrated, not fully.

How could an overconfident classifier look like?

conf.	acc.
3/6	1/6

conf.	correct
4/6	2/6

conf.	correct
5/6	3/6

2. Visualizing & Measuring Calibration

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

p	y
$(\frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon)$	1
$(\frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon)$	1
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	2

p	y
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	2
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	3
$(\frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon)$	3

2. Visualizing & 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-ε, ⅓+2ε)	2
(⅓-ε, ⅓-ε, ⅓+2ε)	3
(⅓-ε, ⅓+2ε, ⅓-ε)	3

2. Visualizing & Measuring Calibration

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

confidence	correct
⅓+2ε	1
⅓+2ε	0
⅓+2ε	0

confidence	correct
⅓+2ε	0
⅓+2ε	1
⅓+2ε	0

So this classifier is predicting a random class with full uncertainty.
It always has a confidence of ⅓, and it has an accuracy of ⅓.
Therefore it is perfectly calibrated, but useless.

MICCAI 2023 Vancouver CANADA

2. Visualizing & Measuring Calibration

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

p	y
$(\frac{1}{3}-\varepsilon, 0, \frac{2}{3}+\varepsilon)$	1
$(\frac{2}{3}+\varepsilon, 0, \frac{1}{3}-\varepsilon)$	1
$(0, \frac{2}{3}+\varepsilon, \frac{1}{3}-\varepsilon)$	2

p	y
$(\frac{2}{3}+\varepsilon, 0, \frac{1}{3}-\varepsilon)$	2
$(\frac{1}{3}-\varepsilon, 0,\frac{2}{3}+\varepsilon)$	3
$(0, \frac{1}{3}-\varepsilon, \frac{2}{3}+\varepsilon)$	3

2. Visualizing & Measuring Calibration

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

p	y
$(\frac{1}{3}-\varepsilon, 0, \frac{2}{3}+\varepsilon)$	1
$(\frac{2}{3}+\varepsilon, 0, \frac{1}{3}-\varepsilon)$	1
$(0, \frac{2}{3}+\varepsilon, \frac{1}{3}-\varepsilon)$	2

p	y
$(\frac{2}{3}+\varepsilon, 0, \frac{1}{3}-\varepsilon)$	2
$(\frac{1}{3}-\varepsilon, 0,\frac{2}{3}+\varepsilon)$	3
$(0, \frac{1}{3}-\varepsilon, \frac{2}{3}+\varepsilon)$	3

2. Visualizing & Measuring Calibration

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

confidence	correct
⅔+ε	0
⅔+ε	1
⅔+ε	1

confidence	correct
⅔+ε	0
⅓+ε	1
⅔+ε	1

So this classifier is predicting the correct class with ~⅔ confidence. Additionally, it has an accuracy of ⅔. It is also perfectly calibrated, but it has more discrimination ability than random guessing.

2. Visualizing & 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. Visualizing & Measuring Calibration

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

confidence	correct
1	1
1	1
1	1

confidence	correct
1	1
1	1
1	1

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

PSRs can measure calibration and discrimination jointly.

2. Visualizing & 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: $y = 3,\ \mathbf{y} = (0, 0, 1, 0),\ \mathbf{p} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right),\ \mathbf{q} = \left(0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$

$$\mathrm{Brier}(\mathbf{p},\mathbf{y}) = (1/4)^2 + (1/4)^2 + (3/4)^2 + (1/4)^2 = 3/4$$

$$\mathrm{Brier}(\mathbf{q},\mathbf{y}) = (1/4)^2 + (1/2)^2 + (1/4)^2 = 1/2$$

$$\mathrm{Brier}(\mathbf{y},\mathbf{y}) = 0$$

2. Visualizing & 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: $y = 3,\ \mathbf{y} = (0, 0, 1, 0),\ \mathbf{p} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right),\ \mathbf{q} = \left(0, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$

$$\mathrm{Brier}(\mathbf{p},\mathbf{y}) = 3/4 \qquad \mathrm{NLL}(\mathbf{p},\mathbf{y}) = \text{-}\log(\mathbf{p}_3) = \text{-}\log(1/4) \approx 0.602$$

$$\mathrm{Brier}(\mathbf{q},\mathbf{y}) = 1/2 \qquad \mathrm{NLL}(\mathbf{q},\mathbf{y}) = \text{-}\log(\mathbf{q}_3) = \text{-}\log(1/2) \approx 0.301$$

$$\mathrm{Brier}(\mathbf{y},\mathbf{y}) = 0 \qquad \mathrm{NLL}(\mathbf{y},\mathbf{y}) = \text{-}\log(\mathbf{y}_3) = \text{-}\log(1) = 0$$

Note that fully uncertain predictions do not score well.

3. Improving Calibration

- **Post-Training Calibration**

Classic methods: Platt Scaling & Isotonic Regression:

- Platt is parametric: Fits a logistic regression model on predictions using validation set, corrects “sigmoidal preds.”
- Isotonic is non-parametric: Fits a monotonic piecewise constant mapping, optimizing bins to maximize calibration

These are designed for binary classification scenarios.

Temperature Scaling: designed specifically for NNs. Use a validation set to learn a scalar T dividing logits before applying softmax and tempers their value:

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

We will see examples in the hands-on

3. Improving Calibration

- **Model Ensembling**

  It's been shown that the simple approach of ensembling several diverse models can improve calibration, surpassing most other techniques. Of course it comes with a computational overhead.

- **Training Time Calibration**

  As NNs are over-parametrized, they can keep on learning the training set until they are fully confident, by minimizing NLL indefinitely. We can avoid this by regularizing and trying to promoted balanced confidence.

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

We will see examples in the hands-on

4. Hands-On

- Train a miscalibrated classifier on med-MNIST so that it is fast. Or use blur to create Aleatoric Unc
- Underfit by training very shortly, LS -> underconfident
- Overfit by training on a training subset -> overconfident

- Draw Reliability plots, measure Expected Calibration Error, PSR, other metrics.
- Recalibrate a model with Temperature Scaling, plot all.
- Re-Train with regularization/ensemble, plot all.

- Ideally run some experiment on segmentation?