Part III
Model Calibration

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

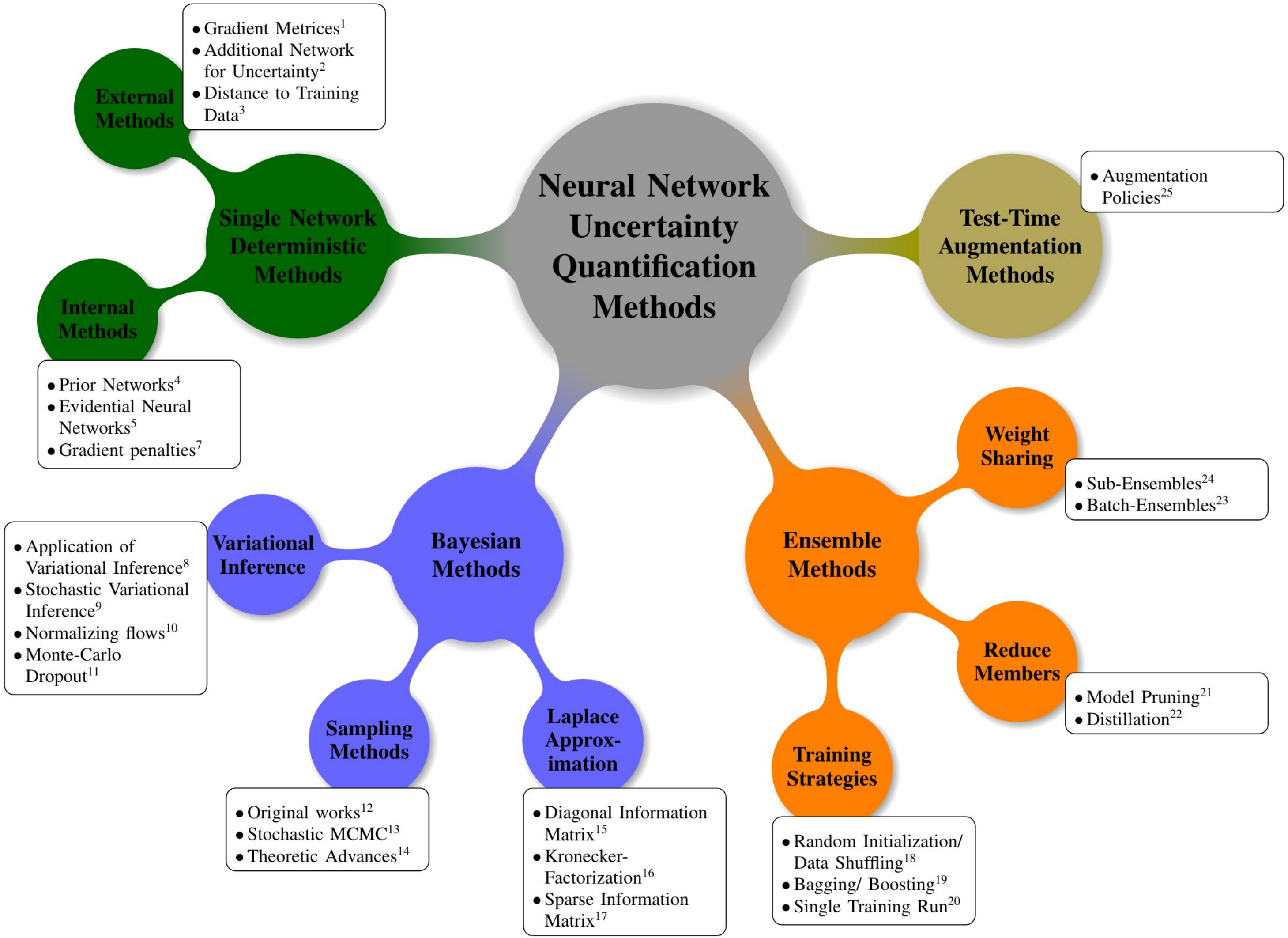
Meritxell Riera i Marin, Researcher
Sycai Medical, Barcelona, Spain
Universitat Pompeu Fabra, Barcelona, Spain

MICCAI2023 Vancouver CANADA

Contents

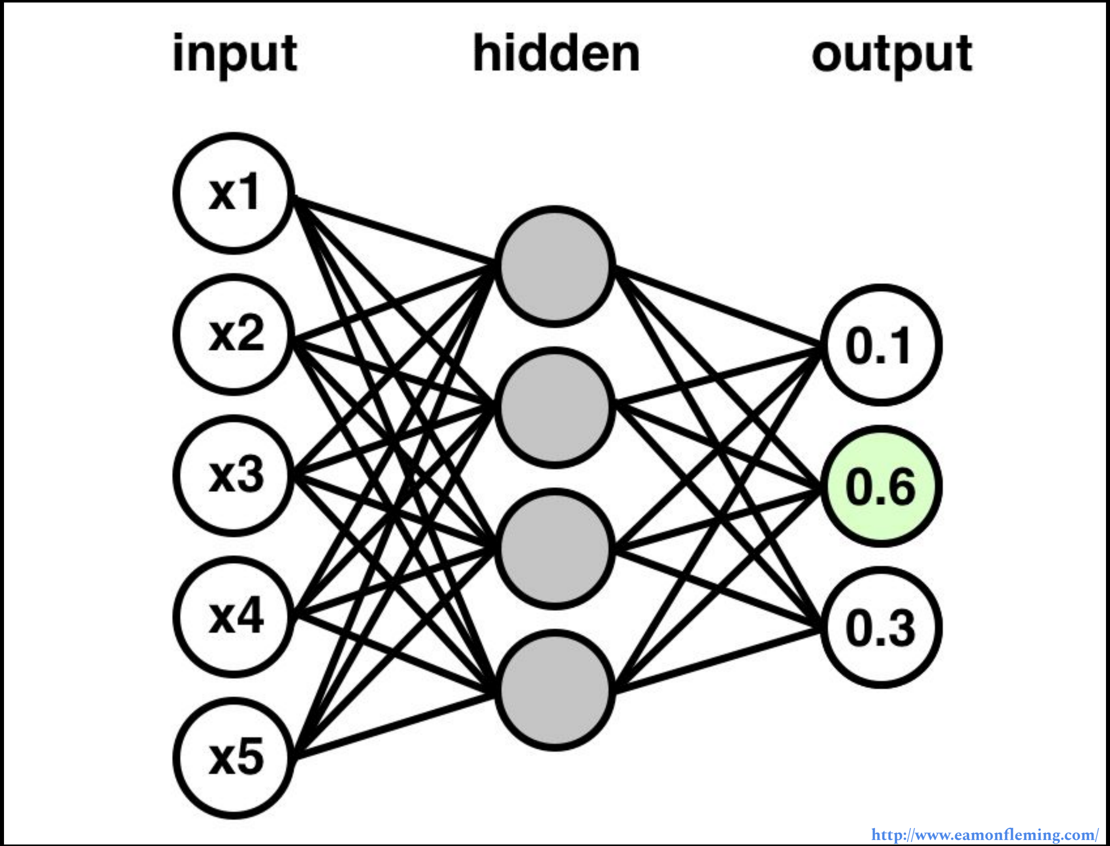
1. Understanding Calibration
2. Measuring Calibration
3. Improving Calibration
4. Practical Hands-On Session

MICCAI 2023 Vancouver CANADA



MICCAI 2023 Vancouver CANADA

But, don't we already have probabilities? (probability= confidence = uncertainty)?



MICCAI 2023 Vancouver CANADA


Pepephone 4G
19:12
65%
Bilbao, ES >
1
NOW
HOURLY
10 DAY
MAPS
September 12 - September 21, 2023
Tue
80%
24° | 18°
80% Chance of Storms
Wed
30%
22° | 17°
30% Chance of Rain
Thu
25° | 18°
Partly Cloudy
Fri
30%
30° | 20°
30% Chance of Rain
Sat
50%
30° | 20°
50% Chance of Rain
Sun
40%
25° | 17°
40% Chance of Rain
Mon
25° | 17°
Unknown

1. Understanding Calibration

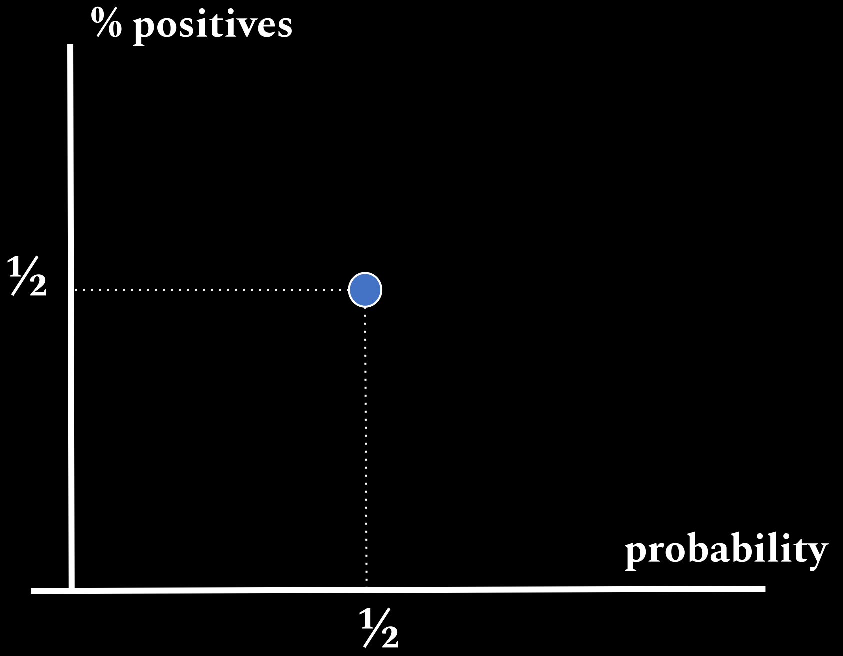
p	y
½	0
½	0
½	0
½	1
½	1
½	1

p	y
¾	0
¾	1
¾	1
¾	1
1	1
1	1

1. Understanding Calibration

p	y
½	0
½	0
½	0
½	1
½	1
½	1

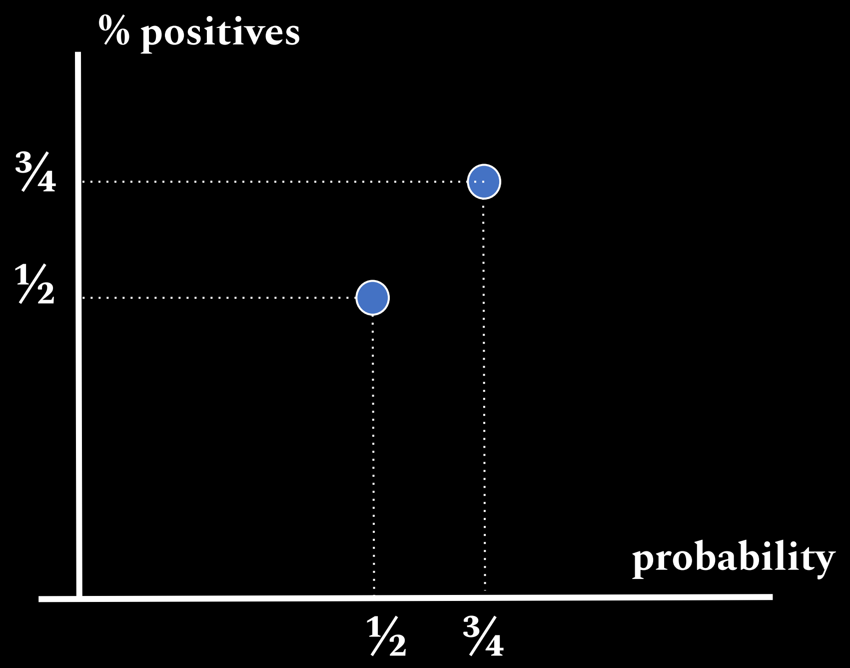
p	y
¾	0
¾	1
¾	1
¾	1
1	1
1	1



1. Understanding Calibration

p	y
½	0
½	0
½	0
½	1
½	1
½	1

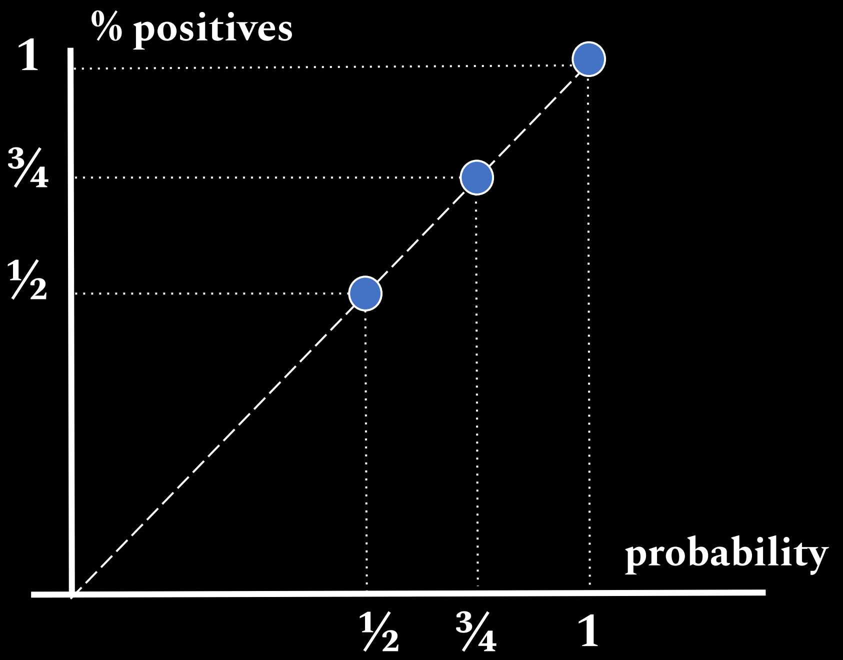
p	y
¾	0
¾	1
¾	1
¾	1
1	1
1	1



1. Understanding Calibration

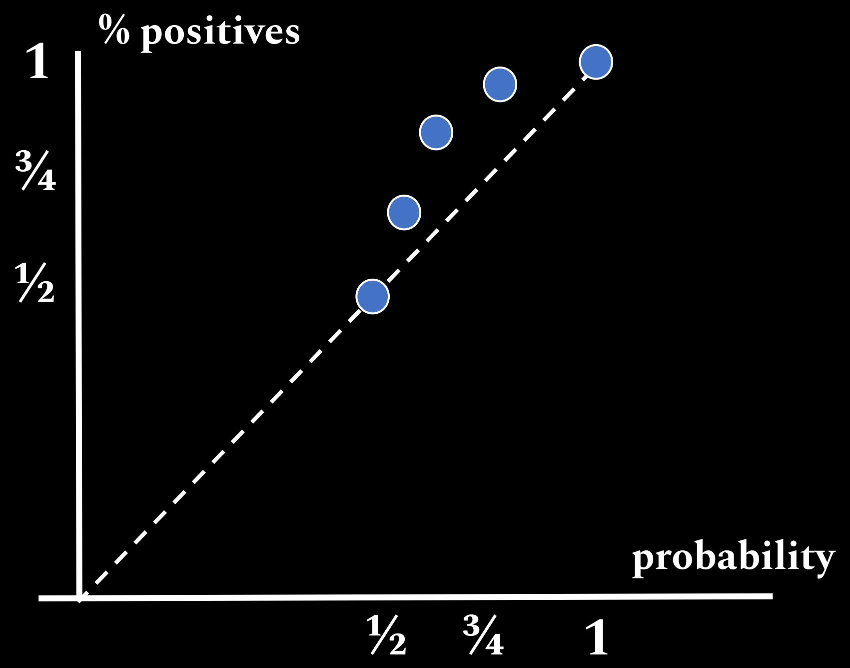
p	y
½	0
½	0
½	0
½	1
½	1
½	1

p	y
¾	0
¾	1
¾	1
¾	1
1	1
1	1



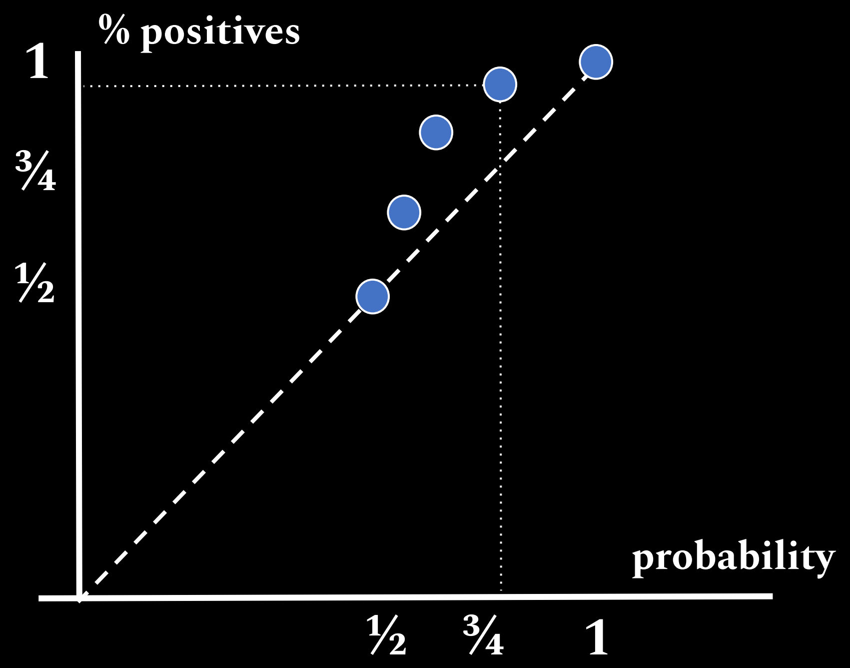
1. Understanding Calibration

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



1. Understanding Calibration

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

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

MICCAI 2023 Vancouver CANADA

2. Measuring Calibration

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

Full-calibration: consider the whole probability vector.

p	y
⅔	1
⅔	1
⅔	2

p	y
⅔	2
⅔	2
⅔	3

p	y
⅔	3
⅔	3
⅔	1

2. Measuring Calibration

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

Full-calibration: consider the whole probability vector.
Confidence calibration: only consider highest probability.

p	y
⅔	1
⅔	1
⅔	2

p	y
⅔	2
⅔	2
⅔	3

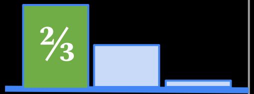
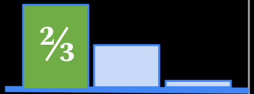
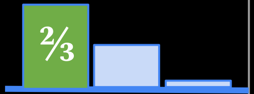
p	y
⅔	3
⅔	3
⅔	1

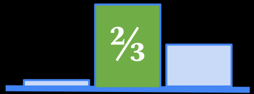
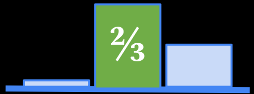
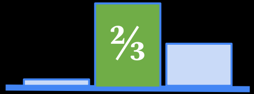
MICCAI 2023 Vancouver CANADA

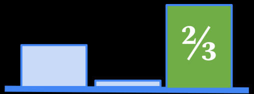
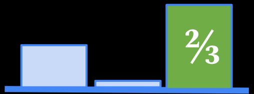
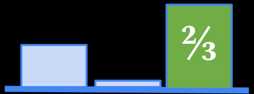
2. Measuring Calibration

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

Full-calibration: consider the whole probability vector.
Confidence calibration: only consider highest probability.

p	$(\hat{y}, c)$	y
 ⅔	(1,⅔)	1
 ⅔	(1,⅔)	1
 ⅔	(1,⅔)	2

p	$(\hat{y}, c)$	y
 ⅔	(2,⅔)	2
 ⅔	(2,⅔)	2
 ⅔	(2,⅔)	3

p	$(\hat{y}, c)$	y
 ⅔	(3,⅔)	3
 ⅔	(3,⅔)	3
 ⅔	(3,⅔)	1

*Also **Class-wise calibration**: consider marginal probabilities, 1vsRest.

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 Confidence–Calibration Error**

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

- **Expected Class-Wise Calibration Error**

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

MICCAI 2023 Vancouver CANADA

2. Measuring Calibration

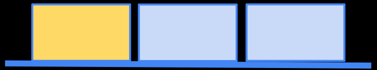
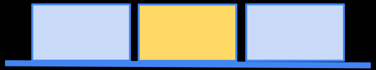
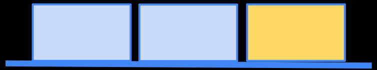
- Alternative Calibration Measures: Proper Scoring Rules

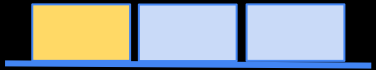
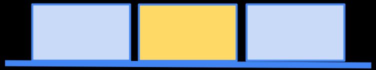
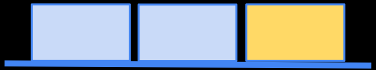
p	ŷ	y
		1
		1
		2

p	ŷ	y
		2
		3
		3

2. Measuring Calibration

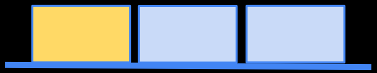
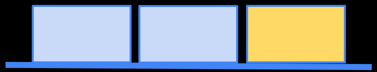
- **Alternative Calibration Measures: Proper Scoring Rules**

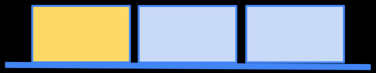
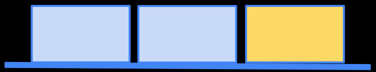
p	$\hat{y}$	y
		1
		1
		2

p	$\hat{y}$	y
		2
		3
		3

2. Measuring Calibration

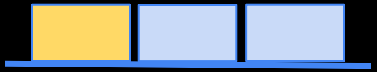
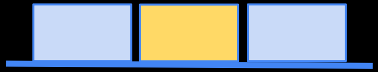
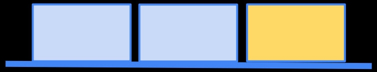
- **Alternative Calibration Measures: Proper Scoring Rules**

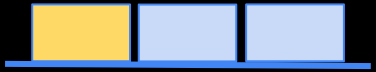
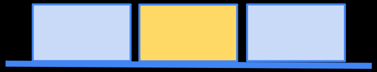
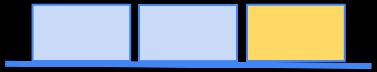
p	ŷ	y
	1	1
	2	1
	3	2

p	ŷ	y
	1	2
	2	3
	3	3

2. Measuring Calibration

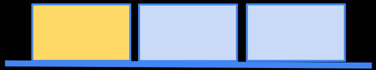
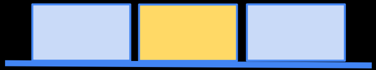
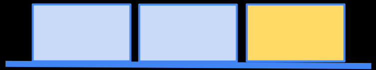
- **Alternative Calibration Measures: Proper Scoring Rules**

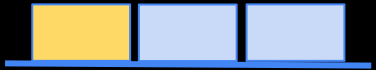
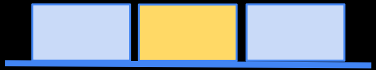
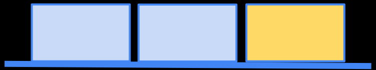
p	ŷ	y
	1	1
	2	1
	3	2

p	ŷ	y
	1	2
	2	3
	3	3

2. Measuring Calibration

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

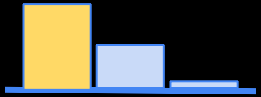
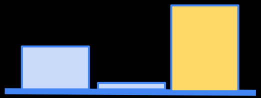
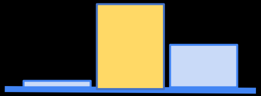
p	ŷ	y
	1	1
	2	1
	3	2

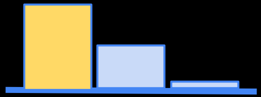
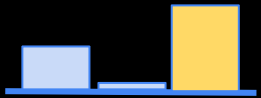
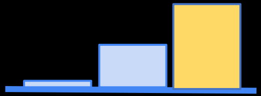
p	ŷ	y
	1	2
	2	3
	3	3

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

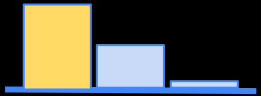
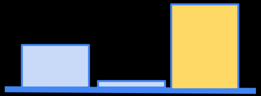
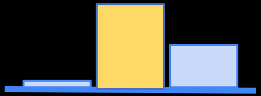
- **Alternative Calibration Measures: Proper Scoring Rules**

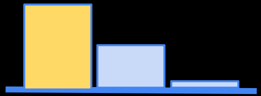
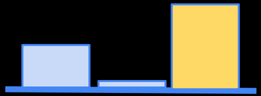
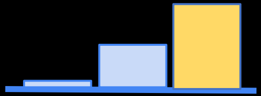
p	ŷ	y
		1
		1
		2

p	ŷ	y
		2
		3
		3

2. Measuring Calibration

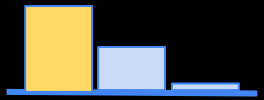
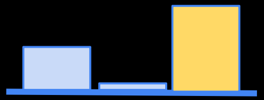
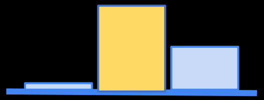
- **Alternative Calibration Measures: Proper Scoring Rules**

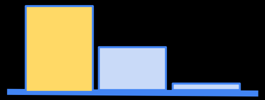
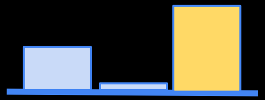
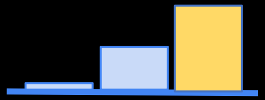
p	$\hat{y}$	y
	1	1
	3	1
	2	2

p	$\hat{y}$	y
	1	2
	3	3
	3	3

2. Measuring Calibration

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

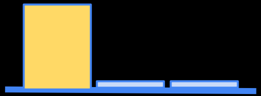
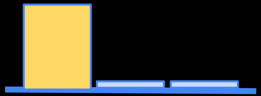
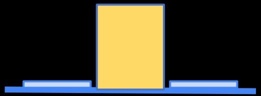
p	ŷ	y
	1	1
	3	1
	2	2

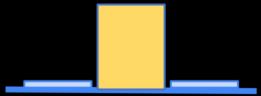
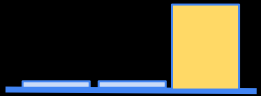
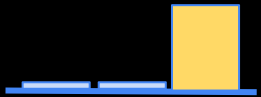
p	ŷ	y
	1	2
	3	3
	3	3

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

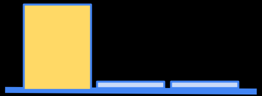
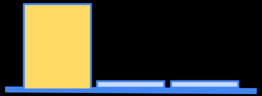
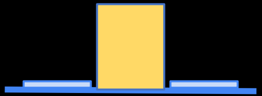
- **Alternative Calibration Measures: Proper Scoring Rules**

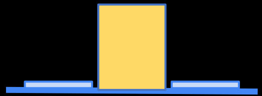
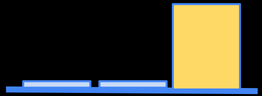
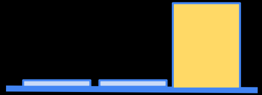
p	ŷ	y
	1	1
	1	1
	2	2

p	ŷ	y
	2	2
	3	3
	3	3

2. Measuring Calibration

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

p	ŷ	y
	1	1
	1	1
	2	2

p	ŷ	y
	2	2
	3	3
	3	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, Logarithmic Score (aka Cross-Entropy)

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

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

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

$$\mathrm{CE}(\mathbf{p}_{\mathrm{bad}}, \mathbf{y}) \approx 0.477 \qquad \mathrm{CE}(\mathbf{p}_{\mathrm{better}}, \mathbf{y}) \approx 0.176 \qquad \mathrm{CE}(\mathbf{y}, \mathbf{y}) = 0$$

Note that a fully uncertain prediction $\mathbf{p}_{\mathrm{bad}}$ does not score well.

3. Improving Calibration

- **Model Ensembling**

Ensembling several diverse models can reduce over-confidence.

- **Training Time Calibration**

Over-parametrized NNs can keep on learning the training set until they are fully confident, minimizing NLL indefinitely.

We can regularize to disencourage confidence : Label Smoothing, MixUp, Focal Loss... Careful of underfitting! Report also PSRs.

- **Post-Training Calibration**

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

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

MICCAI 2023 Vancouver CANADA

4. Hands-On