Part III
Model Calibration

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

Adrian Galdran, RYC Research Fellow
Computer Vision Center, UAB, Barcelona, Spain

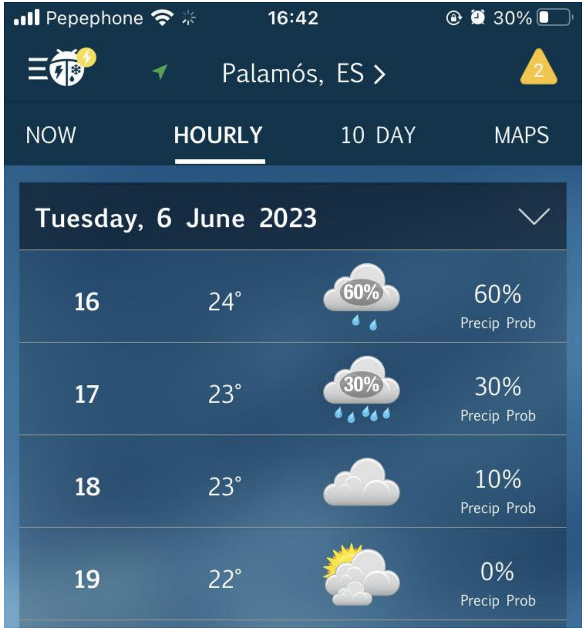
Contents

Contents

What is Calibration?

Calibration is the process of ensuring that a model's predicted probabilities match the actual outcomes.

Example: if a model predicts a 70% chance of an event, it should occur 70% of the time.



Why is Calibration important?

To avoid over and under confident predictions.

Example: a diagnostic of having or not having cancer has to be calibrated to be able to detect the ambiguous cases and act accordingly.

Understanding Calibration

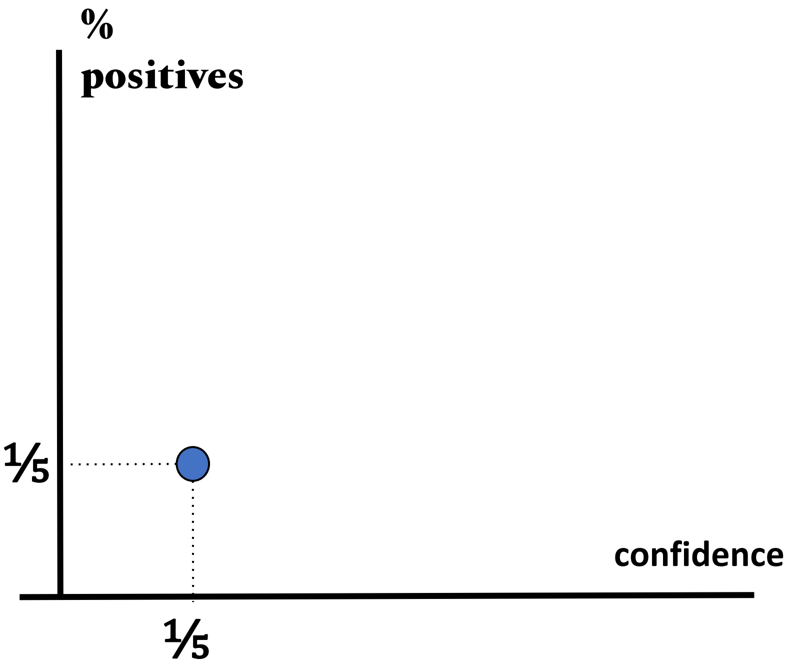
p	y
$\frac{1}{5}$	0
$\frac{1}{5}$	0
$\frac{1}{5}$	0
$\frac{1}{5}$	0
$\frac{1}{5}$	1

p	y
$\frac{1}{3}$	0
$\frac{1}{3}$	0
$\frac{1}{3}$	1
$\frac{1}{2}$	0
$\frac{1}{2}$	1

p	y
$\frac{3}{4}$	0
$\frac{3}{4}$	1
$\frac{3}{4}$	1
$\frac{3}{4}$	1
1	1

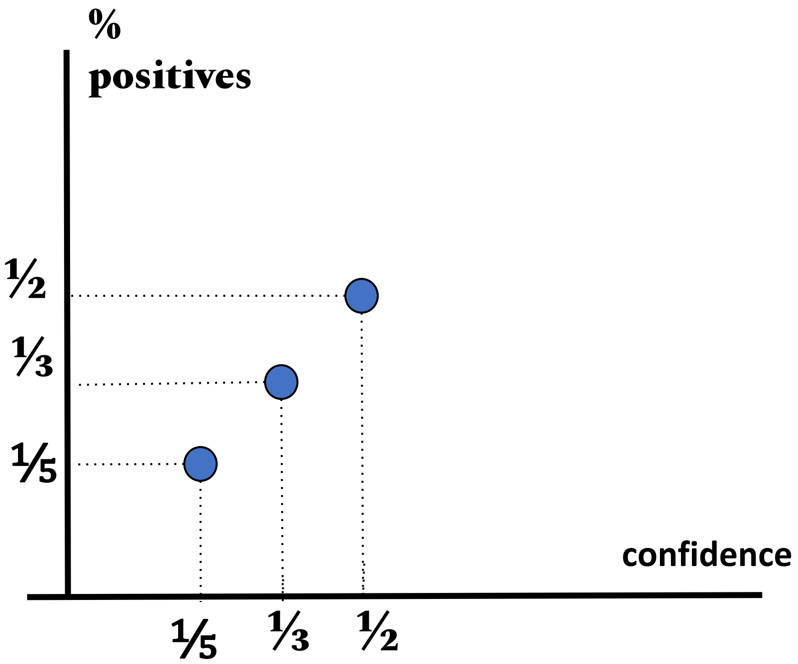
Understanding Calibration

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1



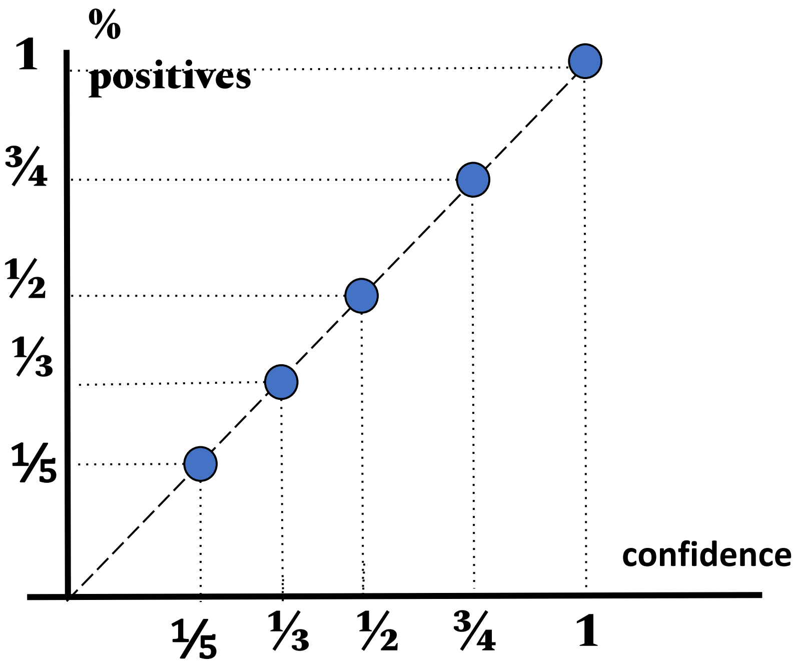
Understanding Calibration

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1



Understanding Calibration

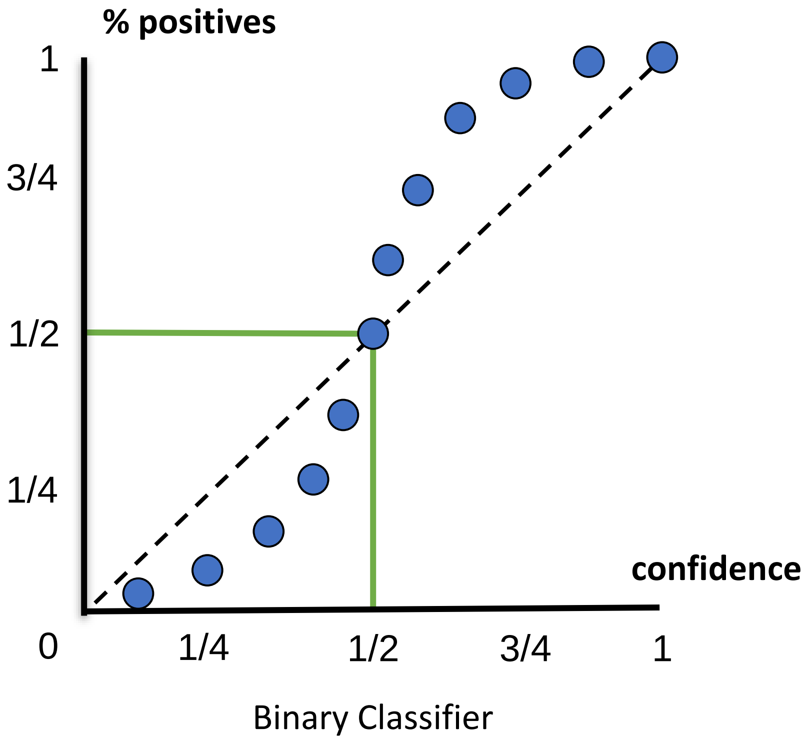
p	y
¾	0
¾	1
¾	1
¾	1
1	1



Contents

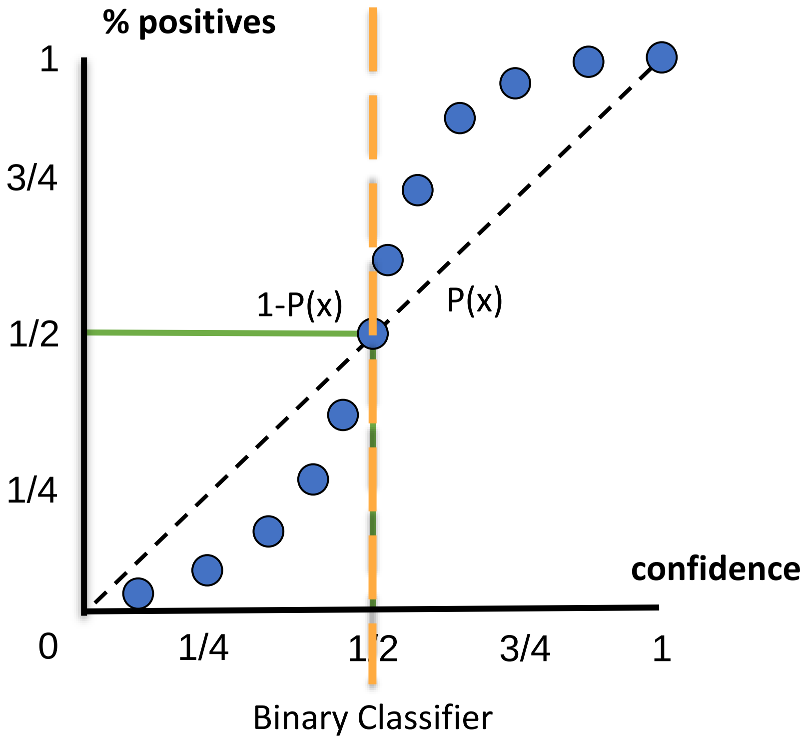
Measuring calibration: Reliability plots

X-Axis: Predicted probability bins
(e.g., 0.1, 0.2, ..., 0.9, 1.0).
Y-Axis: Observed frequency of positive outcomes within each bin.
Ideal Calibration: Points lie on the diagonal line (y = x), indicating perfect calibration.



Measuring calibration: Reliability plots

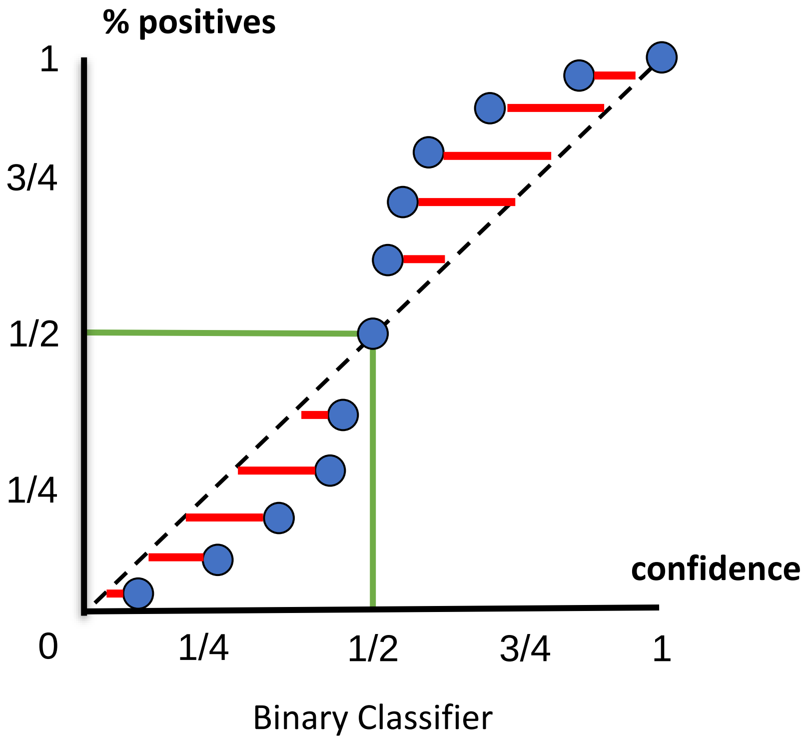
QUESTION:
Are these predictions
under-confident
or
over-confident?



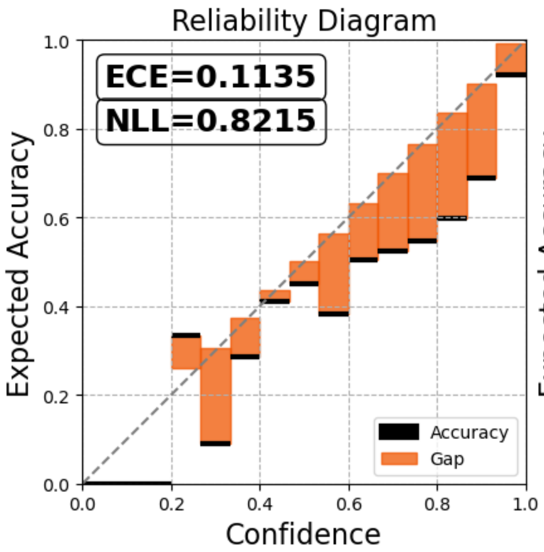
Measuring calibration: Reliability plots

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

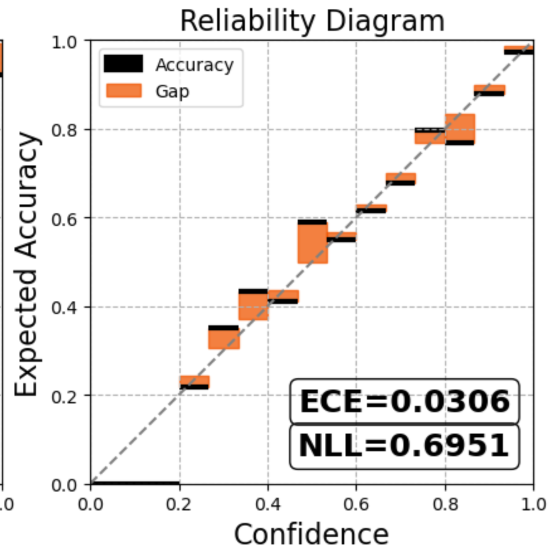
ANSWER:
<u>under</u>-confident



Measuring calibration: Reliability plots (1)



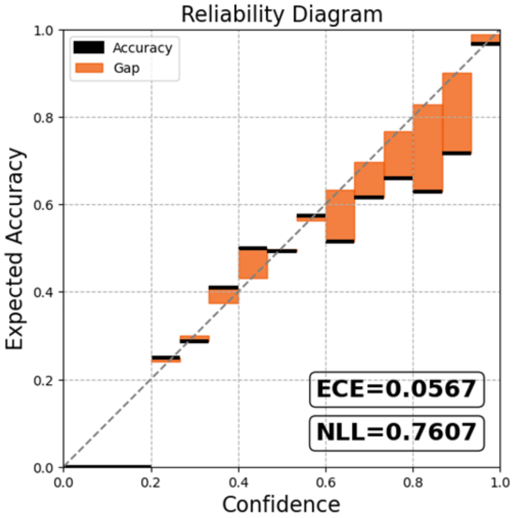
Classification model



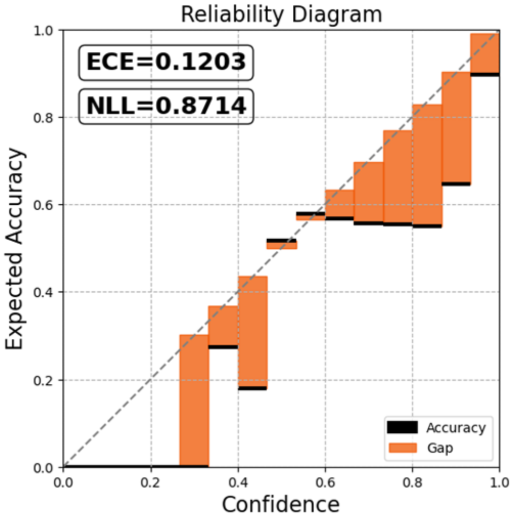
Classification model with temperature scaling (more calibrated)

X-Axis: Predicted probability bins (e.g., 0.1, 0.2, ..., 0.9, 1.0).
Y-Axis: Observed frequency of positive outcomes within each bin.
Ideal Calibration: Points lie on the diagonal line (y = x), indicating perfect calibration.

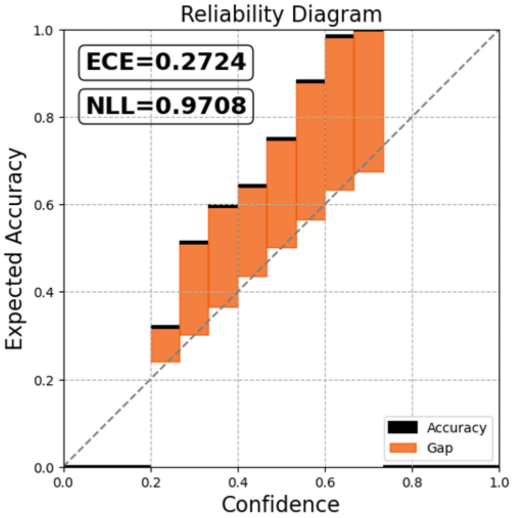
Measuring calibration: Reliability plots (2)



Classification model



Overfit Classification model



Underfit Classification model

Measuring calibration: Expected Calibration Error

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

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

Full Calibration	Class-wise Calibration	Confidence Calibration
Consider the whole probability vector	Only consider marginal probabilities	Only consider highest probability

Measuring Calibration

- Expected Full Calibration Error $$fullECE = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathbb{B}_i|}||prob(\mathbb{B}_i) - true(\mathbb{B}_i)||$$

- Expected Class-Wise Calibration Error $$cwECE = \frac{1}{K}\sum_{k=1}^{K} binECE_k$$

- Expected Confidence Calibration Error $$confECE = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|B_i|}|conf(B_i) - acc(B_i)|$$

Expected Calibration Error Example

- **Full ECE**: 0
- **Class-wise ECE**:
- Class 1: 0
- Class 2: 0
- Class 3: 0
- **Confidence ECE**: 0

[class1, class2, class3]

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

Full Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

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

Consider the whole probability vector

Bin 1 (confidence ~ [⅔, ⅓, 0]):
- Prob: [⅔, ⅓, 0], Acc: [⅔, ⅓, 0]

Bin 2 (confidence ~ [0, ⅔, ⅓]):
- Prob: [0, ⅔, ⅓], Acc: [0, ⅔, ⅓]

Bin 3 (confidence ~ [⅓, 0, ⅔]):
- Prob: [⅓, 0, ⅔], Acc: [⅓, 0, ⅔]

$$\mathrm{fullECE} = \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[\tfrac{2}{3}, \tfrac{1}{3}, 0] - [\tfrac{2}{3}, \tfrac{1}{3}, 0]\right|\right|\right) + \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[0, \tfrac{2}{3}, \tfrac{1}{3}] - [0, \tfrac{2}{3}, \tfrac{1}{3}]\right|\right|\right) + \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[\tfrac{1}{3}, 0, \tfrac{2}{3}] - [\tfrac{1}{3}, 0, \tfrac{2}{3}]\right|\right|\right) = \frac{1}{9}\times 0 + \frac{1}{9}\times 0 + \frac{1}{9}\times 0 = 0$$

Full Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

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

Consider the whole probability vector

Bin 1 (confidence ~ [⅔, ⅓, 0]):
- Prob: [⅔, ⅓, 0], Acc: [⅔, ⅓, 0]

Bin 2 (confidence ~ [0, ⅔, ⅓]):
- Prob: [0, ⅔, ⅓], Acc: [0, ⅔, ⅓]

Bin 3 (confidence ~ [⅓, 0, ⅔]):
- Prob: [⅓, 0, ⅔], Acc: [⅓, 0, ⅔]

$$\text{fullECE} = \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[\tfrac{2}{3}, \tfrac{1}{3}, 0] - [\tfrac{2}{3}, \tfrac{1}{3}, 0]\right|\right|\right) + \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[0, \tfrac{2}{3}, \tfrac{1}{3}] - [0, \tfrac{2}{3}, \tfrac{1}{3}]\right|\right|\right) + \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[\tfrac{1}{3}, 0, \tfrac{2}{3}] - [\tfrac{1}{3}, 0, \tfrac{2}{3}]\right|\right|\right) = \frac{1}{9}\times 0 + \frac{1}{9}\times 0 + \frac{1}{9}\times 0 = 0$$

fullECE = 0→ perfectly calibrated

Full Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	2

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

Consider the whole probability vector

Bin 1 (confidence ~ [⅔, ⅓, 0]):

- Prob: [⅔, ⅓, 0], Acc: [⅔, ⅓, 0]

Bin 2 (confidence ~ [0, ⅔, ⅓]):

- Prob: [0, ⅔, ⅓], Acc: [0, ⅔, ⅓]

Bin 3 (confidence ~ [⅓, 0, ⅔]):

- Prob: [⅓, 0, ⅔], Acc: [0, ⅓, ⅔]

$$\mathrm{fullECE} = \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[⅔, ⅓, 0] - [⅔, ⅓, 0]\right|\right|\right) + \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[0, ⅔, ⅓] - [0, ⅔, ⅓]\right|\right|\right) + \frac{1}{3}\times\frac{1}{3}\times\left(\left|\left|[⅓, 0, ⅔] - [0, ⅓, ⅔]\right|\right|\right) == \frac{1}{9}\times 0 + \frac{1}{9}\times 0 + \frac{1}{9}\times \left|\left|[⅓, -⅓, 0]\right|\right| = \frac{1}{9}\times(\|⅓\|+\|-⅓\|+\|0\|) = \frac{1}{9}\times(⅔) = \frac{2}{27}$$

fullECE != 0→ not perfectly calibrated

Class-wise Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

$$cwECE = \frac{1}{K}\sum_{k=1}^{K} binECE_k$$

Only consider marginal probabilities

Class 1:

- Prob: $\frac{2/3+2/3+2/3+0+0+0+1/3+1/3+1/3}{3} = \frac{7}{9} \approx 0.7778$
- Acc: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{2+5}{2+5+1+1} = \frac{7}{9} \approx 0.7778$

Class 2 and Class 3:

- Same results as in class 1

$$\text{cwECE} = \frac{1}{3} \times (\frac{1}{3} \times (0.7778 - 0.7778) + \frac{1}{3} \times (0.7778 - 0.7778) + \frac{1}{3} \times (0.7778 - 0.7778)) = \frac{1}{9} \times 0 + \frac{1}{9} \times 0 + \frac{1}{9} \times 0 = 0$$

cwECE = 0→ perfectly calibrated

Class-wise Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	2

$$cwECE = \frac{1}{K}\sum_{k=1}^{K} binECE_k$$

Only consider marginal probabilities

Class 1:

- Acc: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{2+6}{2+6+1+0} = \frac{8}{9} \approx 0.8889$

Class 2:

- Acc: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{2+4}{2+4+1+2} = \frac{6}{9} \approx 0.6667$

Class 3 remains the same

$$\text{cwECE} = \frac{1}{3} \times (\frac{1}{3} \times (0.7778 - 0.8889) + \frac{1}{3} \times (0.7778 - 0.6667) + \frac{1}{3} \times (0.7778 - 0.7778)) = \frac{1}{9} \times |-\frac{1}{9}| + \frac{1}{9} \times |\frac{1}{9}| + \frac{1}{9} \times 0 = \frac{2}{27}$$

cwECE != 0→ not perfectly calibrated

Confidence Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

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

Only consider highest probability

There is just one bin, since all the max probabilities in this example are 2/3.

- Prob: $\frac{2/3+2/3+2/3+2/3+2/3+2/3+2/3+2/3+2/3}{3} = \frac{6}{9} \approx 0.6667$
- Acc: $\frac{6}{9} \approx 0.6667$

$$\text{confECE} = \frac{1}{3} \times |(0.6667 - 0.6667)| = 0$$

confECE = 0→ perfectly calibrated

Confidence Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	2

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

Only consider highest probability

There is just one bin, since all the max probabilities in this example are 2/3.

- Prob: $\frac{2/3+2/3+2/3+2/3+2/3+2/3+2/3+2/3+2/3}{3} = \frac{6}{9} \approx 0.6667$
- Acc: $\frac{6}{9} \approx 0.6667$

$$\text{confECE} = \frac{1}{3} \times |(0.6667 - 0.6667)| = 0$$

confECE = 0→ perfectly calibrated

Confidence Calibration

[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3

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

Only consider highest probability

There is just one bin, since all the max probabilities in this example are 2/3.

- Prob: $\frac{2/3+2/3+2/3+2/3+2/3+2/3+2/3+2/3+2/3}{3} = \frac{6}{9} \approx 0.6667$
- Acc: $\frac{7}{9} \approx 0.7778$

$$\text{confECE} = \frac{1}{3} \times |(0.6667 - 0.7778)| = \frac{1}{27}$$

confECE != 0→ not perfectly calibrated

Alternative Calibration Measures: Proper Scoring Rules

- Evaluate the quality of of probabilistic predictions by ensuring that the best score is achieved when the predicted probabilities match the true probabilities.
- Tool for measuring calibration and discrimination jointly.
- Most popular: Brier Score, Negative Log-Likelihood

$$Brier(p, y) = ||p - y||_2^2 \qquad NLL(p, y) = -\log p_y$$

Example: $y = 3, y = (0{,}0{,}1), p_{bad} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), p_{better} = (0, \frac{1}{3}, \frac{2}{3})$

$Brier(p_{bad}, y) = \frac{2}{3}$ $\quad$ $Brier(p_{better}, y) = \frac{2}{9}$ $\quad$ $Brier(y, y) = 0$

$NLL(p_{bad}, y) \approx 0.477$ $\quad$ $NLL(p_{better}, y) \approx 0.176$ $\quad$ $NLL(y, y) = 0$

Note that a fully uncertain prediction p_{bad} does not score well.

Alternative Calibration Measures: Proper Scoring Rules

$\mathbf{p}$	$\hat{\mathbf{y}}$	$\mathbf{y}$
(⅓+2ε, ⅓-ε, ⅓-ε)	1	1
(⅓-ε, ⅓+2ε, ⅓-ε)	2	1
(⅓-ε, ⅓-ε, ⅓+2ε)	3	2

$\mathbf{p}$	$\hat{\mathbf{y}}$	$\mathbf{y}$
(⅓+2ε, ⅓-ε, ⅓-ε)	1	2
(⅓-ε, ⅓+2ε, ⅓-ε)	2	3
(⅓-ε, ⅓-ε, ⅓+2ε)	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**.

Alternative Calibration Measures: Proper Scoring Rules

p	ŷ	y
(⅔, 0, ⅓)	1	1
(0, ⅓, ⅔)	3	1
(⅓, ⅔, 0)	2	2

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

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

Alternative Calibration Measures: Proper Scoring Rules

$\mathbf{p}$	$\hat{\mathbf{y}}$	$\mathbf{y}$
$(1, 0, 0)$	1	1
$(1, 0, 0)$	1	1
$(0, 1, 0)$	2	2

$\mathbf{p}$	$\hat{\mathbf{y}}$	$\mathbf{y}$
$(0, 1, 0)$	2	2
$(0, 0, 1)$	3	3
$(0, 0, 1)$	3	3

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

Contents

Improving Calibration: Training Calibration

- **Model Ensembling:**
    - Ensembling several diverse models can improve calibration.
    - It comes with a computational overhead.

- **Training Time Calibration:**
    - Over-parameterized 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 discourage confidence.
        - Label Smoothing, MixUp, Focal Loss... Careful of underfitting! Always report also a PSR, not only ECE.

Improving Calibration: Post-Training Calibration

For binary classifiers:

- **Platt Scaling:**
  - Fits a logistic regression model to the classifier's scores, mapping the to well calibrated probabilities, using a validation set.
  - Used in ML to improve the reliability of probability estimates.
- **Isotonic Regression:**
  - Non-parametric technique
  - Fits a monotonically increasing function to the scores, optimizing bins to maximize calibration.
  - Especially useful when the relationship between predicted scores and actual probabilities is non-linear.

Improving Calibration: Post-Training Calibration

Temperature Scaling:

- Platt scaling tailored for neural networks, particularly in multi-class classification
- Uses a validation set to learn a scalar parameter called the *temperature* (T)
- It divides the logits (pre-softmax scores) by this T before applying the softmax function and tempers their value.
- This method helps to adjust the confidence of predictions without changing their order, improving the reliability of the predicted probabilities.

$$p_j = \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}} \rightarrow p_j = \frac{e^{\left(\frac{z_j}{T}\right)}}{\sum_{k=1}^{N} e^{\left(\frac{z_k}{T}\right)}}$$

Contents

Hands-on Session & Complementary Materials

- GitHub repository: uqinmia-miccai-2024
- Video recording of different sessions on YouTube.
- More information about the event on our website.