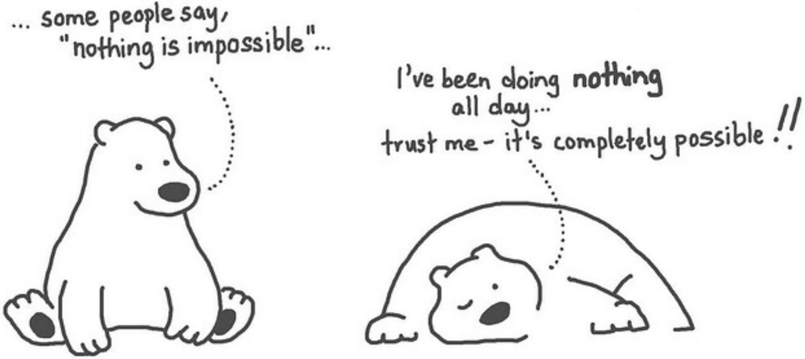
PROBABILITY: CERTAIN, POSSIBLE, IMPOSSIBLE



Uncertainty reasoning and machine learning

Introduction to notions of calibrated and valid predictions

Vu-Linh Nguyen

Chaire de Professeur Junior, Laboratoire Heudiasyc
Université de technologie de Compiègne

AOS4 master courses

Notions of Calibration with Examples (Exercise 1)

Basic setup (rephrased from an example in [3]):

- A dataset contains 40 instances
- A model ***h*** which partitions the input space into 4 regions:

# instances	Predicted probabilities	Class distributions
10	(0.3,0.3,0.4)	(4,2,4)
10	(0.4,0.3,0.3)	(3,4,3)
10	(0.4,0.6,0.0)	(5,5,0)
10	(0.3,0.6,0.1)	(2,7,1)

Notions of Calibration with Examples (Exercise 1)

Basic setup (rephrased from an example in [3]):

- A dataset contains 40 instances
- A model ***h*** which partitions the input space into 4 regions:

# instances	Predicted probabilities	Class distributions
10	(0.3,0.3,0.4)	(4,2,4)
10	(0.4,0.3,0.3)	(3,4,3)
10	(0.4,0.6,0.0)	(5,5,0)
10	(0.3,0.6,0.1)	(2,7,1)

Question: Check if the following statements are correct

- ***h*** is not confidence-calibrated
- ***h*** is classwise-calibrated
- ***h*** is not distribution-calibrated

Notions of Calibration with Examples (Solution 1.1)

Basic setup (rephrased from an example in [3]):

# instances	Predicted probabilities	Class distributions
10	(0.3,0.3,**0.4**)	(4,2,4)
10	(**0.4**,0.3,0.3)	(3,4,3)
10	(0.4,**0.6**,0.0)	(5,5,0)
10	(0.3,**0.6**,0.1)	(2,7,1)

Statement: ***h*** is not confidence-calibrated

$$P\left(y = \arg\max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} \text{ such that } \max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} = \beta\right) = \beta, \forall \beta \in [0,1]. \quad (1)$$

- $\beta = 0.4$: $P = {}^{(4+3)}/_{20} = {}^{7}/_{20} \neq 0.4$
- $\beta = 0.6$: $P = {}^{(5+7)}/_{20} = {}^{12}/_{20} = 0.6$

Notions of Calibration with Examples (Solution 1.2)

Basic setup (rephrased from an example in [3]):

# Instances	Predicted probabilities	Class distributions
10	(0.3,0.3,0.4)	(4,2,4)
10	(0.4,0.3,0.3)	(3,4,3)
10	(0.4,0.6,0.0)	(5,5,0)
10	(0.3,0.6,0.1)	(2,7,1)

Statement: ***h*** is classwise-calibrated

$$P(y \text{ such that } \theta_y | \boldsymbol{x} = \beta_y) = \beta_y, y \in \mathscr{Y}, \beta_y \in [0,1]. \quad (2)$$

- $y_1 \wedge \beta_1 = 0.3$: $P = (2+4)/20 = 0.3$, $y_1 \wedge \beta_1 = 0.4$: $P = (3+5)/20 = 0.4$
- $y_2 \wedge \beta_2 = 0.3$: $P = (2+4)/20 = 0.3$, $y_2 \wedge \beta_2 = 0.6$: $P = (5+7)/20 = 0.6$
- $y_3 \wedge \beta_3 = 0.4$: $P = 4/10 = 0.4$, $y_3 \wedge \beta_3 = 0.3$: $P = 3/10 = 0.3$
- $y_3 \wedge \beta_3 = 0.0$: $P = 0/10 = 0.0$, $y_3 \wedge \beta_3 = 0.1$: $P = 1/10 = 0.1$

Notions of Calibration with Examples (Solution 1.3)

Basic setup (rephrased from an example in [3]):

# Instances	Predicted probabilities	Class distributions
10	(0.3,0.3,0.4)	(4,2,4)
10	(0.4,0.3,0.3)	(3,4,3)
10	(0.4,0.6,0.0)	(5,5,0)
10	(0.3,0.6,0.1)	(2,7,1)

Statement: $\boldsymbol{h}$ is not distribution-calibrated

$$P(y \text{ such that } \boldsymbol{\theta} | \boldsymbol{x} = \boldsymbol{q}) = \boldsymbol{q}, \forall \boldsymbol{q} \in \triangle^{|\mathscr{Y}|}, \tag{3}$$

- $\boldsymbol{q} = (0.3, 0.3, 0.4)$: $P = (4/10, 2/10, 4/10) = (0.4, 0.2, 0.4) \neq (0.3, 0.3, 0.4)$
- $\boldsymbol{q} = (0.4, 0.3, 0.3)$: $P = (3/10, 4/10, 3/10) = (0.3, 0.4, 0.3) \neq (0.4, 0.3, 0.3)$
- $\boldsymbol{q} = (0.4, 0.6, 0.0)$: $P = (5/10, 5/10, 0/10) = (0.5, 0.5, 0.0) \neq (0.4, 0.6, 0.0)$
- $\boldsymbol{q} = (0.3, 0.6, 0.1)$: $P = (2/10, 7/10, 1/10) = (0.2, 0.7, 0.1) \neq (0.3, 0.6, 0.1)$

A Note on Classifier Calibration (Exercise 2)

Consider three notions of classifier calibration:

- Confidence calibration [1]:

$$P(y = \arg\max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} \text{ such that } \max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} = \beta) = \beta, \forall \beta \in [0,1]. \tag{4}$$

- Classwise calibration [4]:

$$P(y \text{ such that } \theta_y | \boldsymbol{x} = \beta_y) = \beta_y, y \in \mathcal{Y}, \beta_y \in [0,1]. \tag{5}$$

- Distribution calibration [2]:

$$P(y \text{ such that } \boldsymbol{\theta} | \boldsymbol{x} = \boldsymbol{q}) = \boldsymbol{q}, \forall \boldsymbol{q} \in \triangle^{|\mathcal{Y}|}, \tag{6}$$

  where $\triangle^{|\mathcal{Y}|}$ is the $|\mathcal{Y}|$-dimensional simplex.

Prove that these notions are equivalent for binary classification?

A Note on Classifier Calibration (Exercise 3)

Consider three notions of classifier calibration:

- Confidence calibration [1]:

$$P(y = \arg\max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} \text{ such that } \max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} = \beta) = \beta, \forall \beta \in [0,1]. \quad (7)$$

- Classwise calibration [4]:

$$P(y \text{ such that } \theta_y | \boldsymbol{x} = \beta_y) = \beta_y, y \in \mathcal{Y}, \beta_y \in [0,1]. \quad (8)$$

- Distribution calibration [2]:

$$P(y \text{ such that } \boldsymbol{\theta} | \boldsymbol{x} = \boldsymbol{q}) = \boldsymbol{q}, \forall \boldsymbol{q} \in \triangle^{|\mathcal{Y}|}, \quad (9)$$

where $\triangle^{|\mathcal{Y}|}$ is the $|\mathcal{Y}|$-dimensional simplex.

A Note on Classifier Calibration (Exercise 3)

Consider three notions of classifier calibration:

- Confidence calibration [1]:

$$P(y = \arg\max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} \text{ such that } \max_{y \in \mathcal{Y}} \theta_y | \boldsymbol{x} = \beta) = \beta, \forall \beta \in [0,1]. \tag{7}$$

- Classwise calibration [4]:

$$P(y \text{ such that } \theta_y | \boldsymbol{x} = \beta_y) = \beta_y, y \in \mathcal{Y}, \beta_y \in [0,1]. \tag{8}$$

- Distribution calibration [2]:

$$P(y \text{ such that } \boldsymbol{\theta} | \boldsymbol{x} = \boldsymbol{q}) = \boldsymbol{q}, \forall \boldsymbol{q} \in \triangle^{|\mathcal{Y}|}, \tag{9}$$

  where $\triangle^{|\mathcal{Y}|}$ is the $|\mathcal{Y}|$-dimensional simplex.

Prove that $\boldsymbol{h}(\boldsymbol{x}) = P(\mathcal{Y})$, $\forall\ \boldsymbol{x}$, is perfectly calibrated?

Outline

- Calibration Errors
- Post-hoc Calibration

- Conformal Prediction

Calibration Error: The Binary Case (Exercise 4)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1\}$
- The proportion of instances with $y = 1$ is $0.5 + \epsilon$
- The decision rule is $0/1$ loss ℓ and the number of bins is 10

Questions:

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

Calibration Error: The Binary Case (Exercise 4)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1\}$
- The proportion of instances with $y = 1$ is $0.5 + \epsilon$
- The decision rule is $0/1$ loss ℓ and the number of bins is 10

Questions:

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.5 - \epsilon$$

Calibration Error: The Binary Case (Exercise 4)

Basic setup:

- A given data set $\mathbf{D} = \left\{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\right\}$ with $y \in \{0, 1\}$
- The proportion of instances with $y = 1$ is $0.5 + \epsilon$
- The decision rule is $0/1$ loss ℓ and the number of bins is 10

Questions:

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.5 - \epsilon$$

- Can we find worse perfectly calibrated classifiers?

Calibration Error: The Binary Case (Exercise 5)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1\}$
- The proportion of instances with $y = 1$ is $\alpha \neq 0.5$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

Calibration Error: The Binary Case (Exercise 5)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1\}$
- The proportion of instances with $y = 1$ is $\alpha \neq 0.5$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = \min(\alpha, 1 - \alpha)$$

Calibration Error: The Binary Case (Exercise 5)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1\}$
- The proportion of instances with $y = 1$ is $\alpha \neq 0.5$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{Binary-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = \min(\alpha, 1 - \alpha)$$

- Can we find worse perfectly calibrated classifiers?

Classwise Calibration Error (Exercise 6)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1, 2\}$
- The proportions of instances with $(y = 0, y = 1, y = 2)$ are $(\alpha_0, \alpha_1, \alpha_2)$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Can we find at least one classifier with

$$\text{classwise-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

Classwise Calibration Error (Exercise 6)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1, 2\}$
- The proportions of instances with $(y = 0, y = 1, y = 2)$ are $(\alpha_0, \alpha_1, \alpha_2)$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Can we find at least one classifier with

$$\text{classwise-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{classwise-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 1 - \max(\alpha_0, \alpha_1, \alpha_2)$$

Classwise Calibration Error (Exercise 6)

Basic setup:

- A given data set $\mathbf{D} = \left\{(\boldsymbol{x}_n, y_n) | n = 1, \dots, N\right\}$ with $y \in \{0, 1, 2\}$
- The proportions of instances with $(y = 0, y = 1, y = 2)$ are $(\alpha_0, \alpha_1, \alpha_2)$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Can we find at least one classifier with

$$\text{classwise-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{classwise-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 1 - \max(\alpha_0, \alpha_1, \alpha_2)$$

- Can we find worse perfectly calibrated classifiers?

Confidence Calibration Error (Exercise 7)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1, 2\}$
- The proportions of instances with $(y = 0, y = 1, y = 2)$ are $(\alpha_0, \alpha_1, \alpha_2)$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Show that there is at least one classifier with

$$\text{Confidence-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

Confidence Calibration Error (Exercise 7)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1, 2\}$
- The proportions of instances with $(y = 0, y = 1, y = 2)$ are $(\alpha_0, \alpha_1, \alpha_2)$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Show that there is at least one classifier with

$$\text{Confidence-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{Confidence-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N} \sum_{n=1}^{N} \ell(y_n^*, y_n) = 1 - \max(\alpha_0, \alpha_1, \alpha_2)$$

Confidence Calibration Error (Exercise 7)

Basic setup:

- A given data set $\mathbf{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \ldots, N\}$ with $y \in \{0, 1, 2\}$
- The proportions of instances with $(y = 0, y = 1, y = 2)$ are $(\alpha_0, \alpha_1, \alpha_2)$
- The decision rule is $0/1$ loss ℓ and the number of bins is M

Questions:

- Show that there is at least one classifier with

$$\text{Confidence-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 0.0$$

- Show that there is at least one classifier with

$$\text{Confidence-ECE}(\mathbf{D}) = 0.0 \text{ and } \frac{1}{N}\sum_{n=1}^{N} \ell(y_n^*, y_n) = 1 - \max(\alpha_0, \alpha_1, \alpha_2)$$

- Can we find worse perfectly calibrated classifiers?

Outline

- Calibration Errors
- Post-hoc Calibration

- Conformal Prediction

Empirical Binning (Exercise 8)

Basic Setup:

- Binary classification: $\mathscr{Y} := \{0, 1\}$
- Loss function: $\ell(y', y) = \mathbb{1}(y' \neq y)$
- Prediction: $y_\ell^{\boldsymbol{\theta}} = \mathbb{1}(\theta_y|\boldsymbol{x} > 0.5)$

Empirical Binning (Exercise 8)

Basic Setup:

- Binary classification: $\mathscr{Y} := \{0, 1\}$
- Loss function: $\ell(y', y) = \mathbb{1}(y' \neq y)$
- Prediction: $y_\ell^{\boldsymbol{\theta}} = \mathbb{1}(\theta_y | \boldsymbol{x} > 0.5)$

Steps:

- Apply equal-width binning to $\theta_1 | \boldsymbol{x}$ on $\mathbf{D}$
- For each bin $\mathbf{B}_m \longrightarrow$ use $\overline{y}(\mathbf{B}_m)$

Empirical Binning (Exercise 8)

Basic Setup:

- Binary classification: $\mathcal{Y} := \{0, 1\}$
- Loss function: $\ell(y', y) = \mathbb{1}(y' \neq y)$
- Prediction: $y_\ell^{\boldsymbol{\theta}} = \mathbb{1}(\theta_y | \boldsymbol{x} > 0.5)$

Steps:

- Apply equal-width binning to $\theta_1 | \boldsymbol{x}$ on $\mathbf{D}$
- For each bin $\mathbf{B}_m \longrightarrow$ use $\overline{y}(\mathbf{B}_m)$

Question: Empirical Binning optimizes binary-ECE($\mathbf{D}$)?

Outline

- Conformal Prediction
  - Notions

Outline

- Conformal Prediction
  - Notions

Marginal Coverage (Exercise 9)

- Prove that if we always predict $Y_{\text{test}} := \mathcal{Y}$ we can always produce perfect conformal predictions w.r.t. the notion of marginal coverage with any chosen significance level $\sigma \in [0,1]$.

Marginal Coverage (Exercise 9)

- Prove that if we always predict $Y_{\text{test}} := \mathscr{Y}$ we can always produce perfect conformal predictions w.r.t. the notion of marginal coverage with any chosen significance level $\sigma \in [0, 1]$.
- Prove that if we know the prior distribution $P(\mathscr{Y})$, we can always produce perfect conformal predictions w.r.t. the notion of marginal coverage with any chosen significance level $\sigma \in [0, 1]$.

References I

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger.
On calibration of modern neural networks.
In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[2] M. Kull and P. Flach.
Novel decompositions of proper scoring rules for classification: score adjustment as precursor to calibration.
In *Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 68–85, 2015.

[3] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach.
Classifier calibration: a survey on how to assess and improve predicted class probabilities.
Machine Learning, pages 1–50, 2023.

[4] B. Zadrozny and C. Elkan.
Transforming classifier scores into accurate multiclass probability estimates.
In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (SIGKDD)*, pages 694–699, 2002.