UPPSALA UNIVERSITET

Calibration of probabilistic predictive models

Machine Learning Journal Club, Gatsby Unit

David Widmann

Department of Information Technology, Uppsala University, Sweden
Centre for Interdisciplinary Mathematics, Uppsala University, Sweden

28 March 2022

Contact: david.widmann@it.uu.se

About me

TL;DR 📖

- 31 year old PhD student at Uppsala University
- On parental leave since September 2021
- Research on uncertainty quantification of probabilistic models
- Active member in the Julia community

About me

Education 🎓

2017—now: PhD student (Uppsala University)
2016—2017: MSc Mathematics (TU Munich)
2013—2016: BSc Mathematics (TU Munich)
2007—2013: Human medicine (LMU and TU Munich)

About me

Education 🎓

2017—now: PhD student (Uppsala University)

2016—2017: MSc Mathematics (TU Munich)

2013—2016: BSc Mathematics (TU Munich)

2007—2013: Human medicine (LMU and TU Munich)

Research interests 🔬

- Research topic: ”Uncertainty-aware deep learning”
- Statistics, probability theory, scientific machine learning, and computer science
- Julia programming, e.g., SciML and Turing

Papers

- J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics.* Vol. 89. Apr. 2019
  - Focus on multi-class classification, calibration lenses, calibration estimation and tests with ECE

Papers

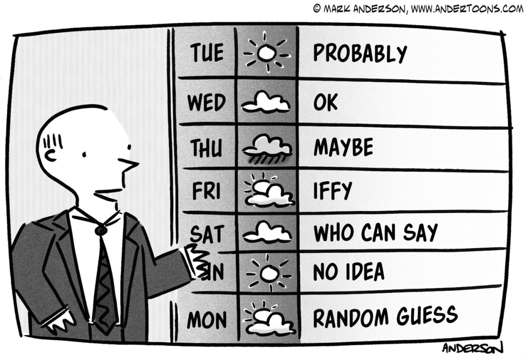
- J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics.* Vol. 89. Apr. 2019
  - Focus on multi-class classification, calibration lenses, calibration estimation and tests with ECE
- D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019
  - Calibration errors and tests for multi-class classification based on matrix-valued kernels

Papers

- J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics.* Vol. 89. Apr. 2019
  - Focus on multi-class classification, calibration lenses, calibration estimation and tests with ECE
- D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019
  - Calibration errors and tests for multi-class classification based on matrix-valued kernels
- D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021
  - Calibration errors and tests for probabilistic predictive models based on scalar-valued kernels

Calibration: Motivation and definition

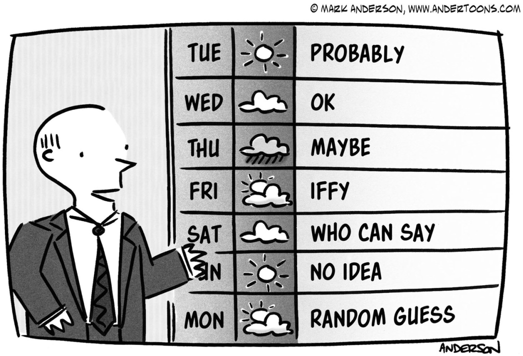
Example: Weather forecasts



"And now the 7-day forecast..."

E. Cooke. "Weighting forecasts." In: *Monthly Weather Review* 34.6 (June 1906), pp. 274–275

Example: Weather forecasts



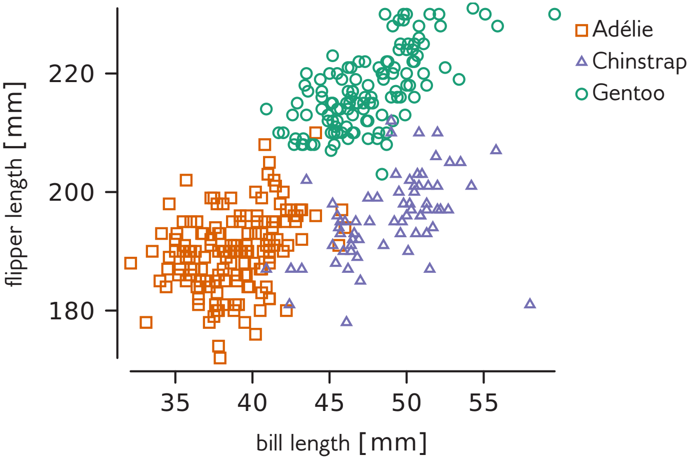
"And now the 7-day forecast..."

"Those forecasts which were marked 'doubtful' were the *best I could frame* under the circumstances. [...] If I make no distinction between these and others, I degrade the whole."

—E. Cooke

E. Cooke. "Weighting forecasts." In: *Monthly Weather Review* 34.6 (June 1906), pp. 274–275

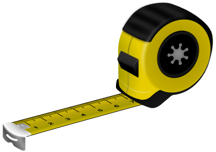
Motivation: Classification example



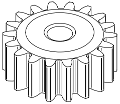
K. B. Gorman, T. D. Williams, and W. R. Fraser. "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)." In: *PLoS ONE* 9.3 (Mar. 2014), e90081

Motivation: Classification example

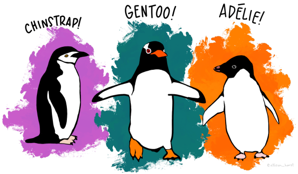
Input X



Model P



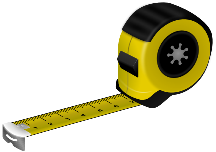
Prediction P_X
(distribution of target Y)

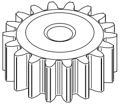


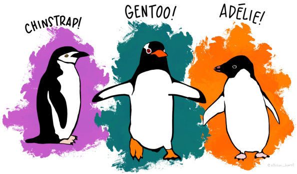
Artwork by @allison_horst

Motivation: Classification example

Input X **Model P** **Prediction P_X (distribution of target Y)**







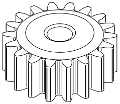
Artwork by @allison_horst

Example: Prediction P_X

Adélie	Chinstrap	Gentoo
80%	10%	10%

Calibration: Intuition

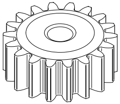
Model P



Calibration: Intuition

Model P

Input x_1 →

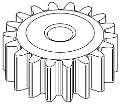


Calibration: Intuition

Model P

Prediction P_{x_1}

Input x_1 $\longrightarrow$



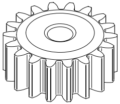
$\longrightarrow$

Adelie	Chinstrap	Gentoo
80%	10%	10%

Calibration: Intuition

Model P

Prediction P_{x_1}

Input x_1 →  →

Adelie	Chinstrap	Gentoo
80%	10%	10%

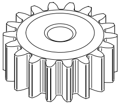
Empirical frequency

Adelie	Chinstrap	Gentoo
/		

Calibration: Intuition

Model P

Input x_2 →

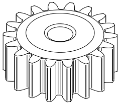


Empirical frequency

Adelie	Chinstrap	Gentoo
/		

Calibration: Intuition

Model P



Prediction P_{x_2}

Input x_2 → →

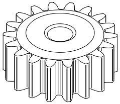
Adelie	Chinstrap	Gentoo
80%	10%	10%

Empirical frequency

Adelie	Chinstrap	Gentoo
$\mid$		

Calibration: Intuition

Model P **Prediction P_{x_2}**

Input x_2 →  →

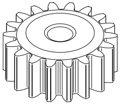
Adelie	Chinstrap	Gentoo
80%	10%	10%

Empirical frequency

Adelie	Chinstrap	Gentoo
/	/	

Calibration: Intuition

Model P

Input x_3 $\longrightarrow$ 

Empirical frequency

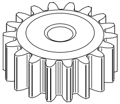
Adelie	Chinstrap	Gentoo
/	/	

Calibration: Intuition

Model P

Prediction P_{x_3}

Input x_3 →



→

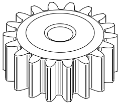
Adelie	Chinstrap	Gentoo
80%	10%	10%

Empirical frequency

Adelie	Chinstrap	Gentoo
/	/	

Calibration: Intuition

Model P **Prediction P_{x_3}**

Input x_3 →  →

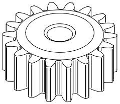
Adelie	Chinstrap	Gentoo
80%	10%	10%

Empirical frequency

Adelie	Chinstrap	Gentoo
//	/	

Calibration: Intuition

Model P **Prediction P_{x_i}**

Input x_i →  →

Adelie	Chinstrap	Gentoo
80%	10%	10%

Empirical frequency

Adelie	Chinstrap	Gentoo
𝍸 III ...	II ...	I ...

Calibration

Prediction P_X

Adélie	Chinstrap	Gentoo
80%	10%	10%

Empirical frequency $\mathrm{law}(Y \mid P_X)$

Adélie	Chinstrap	Gentoo
~~IIII~~ III …	II …	I …

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations*. 2021

Calibration

Predictions consistent with empirically observed frequencies?

Prediction P_X

Adélie	Chinstrap	Gentoo
80%	10%	10%

$\overset{?}{=}$

Empirical frequency $\mathrm{law}(Y \mid P_X)$

Adélie	Chinstrap	Gentoo
~~||||~~ ||| …	|| …	| …

D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests in multi-class classification: A unifying framework.” In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests beyond classification.” In: *International Conference on Learning Representations.* 2021

Calibration

Predictions consistent with empirically observed frequencies?

Prediction P_X

Adélie	Chinstrap	Gentoo
80%	10%	10%

$\overset{?}{=}$

Empirical frequency $\mathrm{law}(Y \,|\, P_X)$

Adélie	Chinstrap	Gentoo
𝍸 ||| …	|| …		…

Definition

A probabilistic predictive model P is calibrated if

$$\mathrm{law}(Y \,|\, P_X) = P_X \qquad \text{almost surely.}$$

D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests in multi-class classification: A unifying framework.” In: *Advances in Neural Information Processing Systems 32*. 2019

D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests beyond classification.” In: *International Conference on Learning Representations*. 2021

Calibration

Predictions consistent with empirically observed frequencies?

Prediction P_X		
`Adélie`	`Chinstrap`	`Gentoo`
80%	10%	10%

$\stackrel{?}{=}$

Empirical frequency $\mathrm{law}(Y \mid P_X)$		
`Adélie`	`Chinstrap`	`Gentoo`
$\cancel{||||}$ $|||$ …	$||$ …	$|$ …

Definition

A probabilistic predictive model P is calibrated if

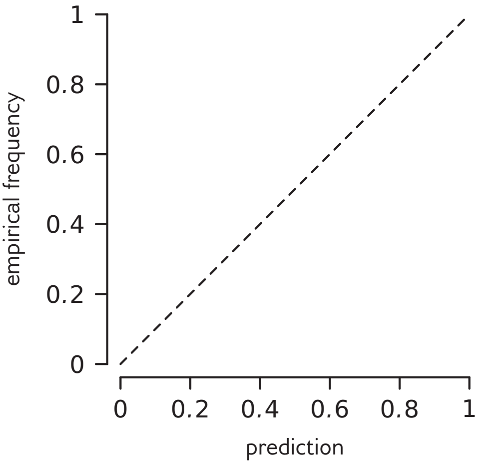
$$\mathrm{law}(Y \mid P_X) = P_X \qquad \text{almost surely.}$$

Notion captures also weaker confidence calibration

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

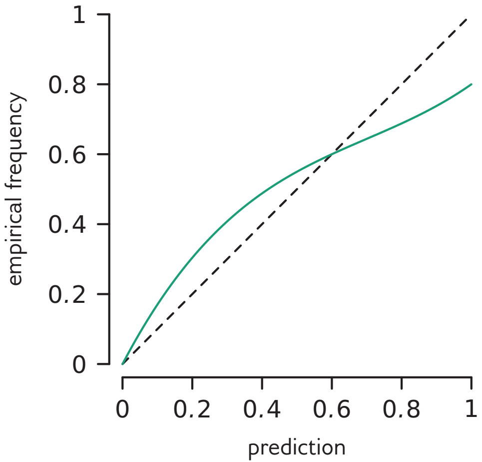
D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations*. 2021

Binary classification: Reliability diagram



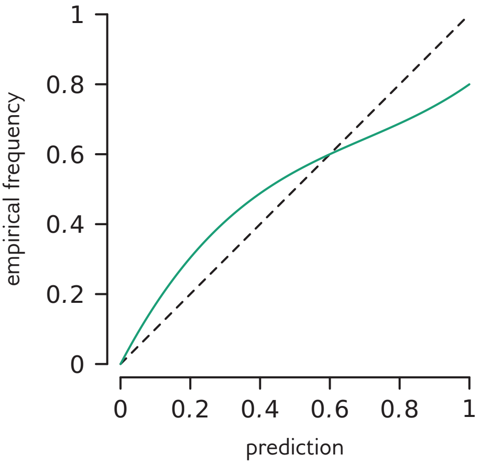
J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Binary classification: Reliability diagram



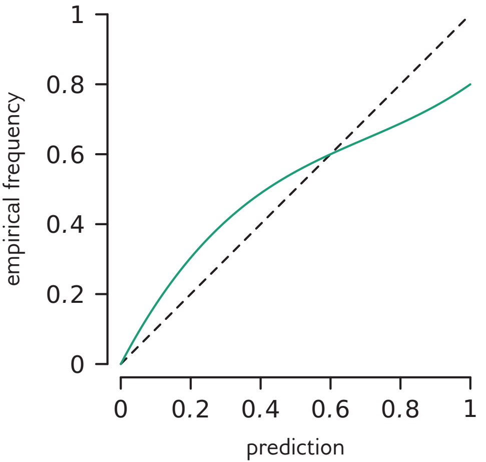
J. Vaicenavicius et al. “Evaluating model calibration in classification.” In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Binary classification: Reliability diagram



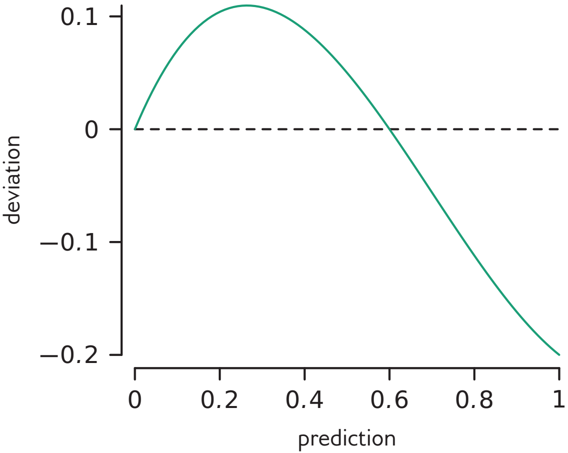
J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Binary classification: Reliability diagram



J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Binary classification: Reliability diagram

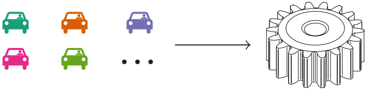


J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Multi-class classification: All scores matter!

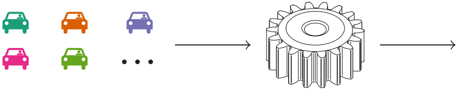
J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Multi-class classification: All scores matter!



J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

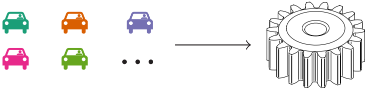
Multi-class classification: All scores matter!



object	human	animal
80%	0%	20%

J. Vaicenavicius et al. “Evaluating model calibration in classification.” In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Multi-class classification: All scores matter!

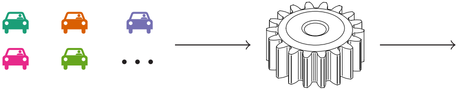


object	human	animal
80%	0%	20%

Common calibration evaluation techniques consider only the most-confident score

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Multi-class classification: All scores matter!



object	human	animal
80%	0%	20%

Common calibration evaluation techniques consider only the most-confident score

Common approaches do not distinguish between the two predictions even though the control actions based on these might be very different!

object	human	animal
80%	0%	20%
80%	20%	0%

J. Vaicenavicius et al. “Evaluating model calibration in classification.” In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Weaker notions of calibration and calibration lenses

Weaker notions

Weaker notions of calibration such as confidence calibration or calibration of marginal classifiers can be analyzed by considering calibration of induced predictive models.

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Weaker notions of calibration and calibration lenses

Weaker notions

Weaker notions of calibration such as confidence calibration or calibration of marginal classifiers can be analyzed by considering calibration of induced predictive models.

Definition (Calibration lenses)

Let ψ be a measureable function that defines targets $Z := \psi(Y, P_X)$. Then ψ induces a predictive model Q for targets Z with predictions

$$Q_X := \mathrm{law}\big(\psi(\tilde{Y}, P_X)\big)$$

where $\tilde{Y} \sim P_X$. Function ψ is called a *calibration lens*.

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Beyond classification

Definition (reminder)

A probabilistic predictive model P is calibrated if

$$\operatorname{law}(Y \mid P_X) = P_X \qquad \text{almost surely.}$$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Beyond classification

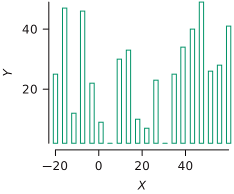
Definition (reminder)

A probabilistic predictive model P is calibrated if

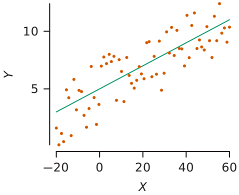
$$\operatorname{law}(Y \mid P_X) = P_X \qquad \text{almost surely.}$$

Examples of other target spaces

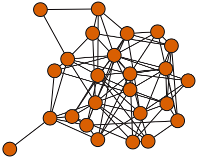
$\mathbb{N}_0$



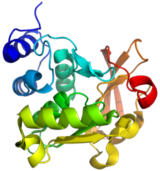
$\mathbb{R}^d$



graphs



protein structure



📖 D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests beyond classification.” In: *International Conference on Learning Representations.* 2021

Calibration errors

Expected calibration error (ECE)

Definition

The expected calibration error (ECE) with respect to distance measure d is defined as

$$\mathrm{ECE}_d := \mathbb{E}_{P_X} d\big(P_X, \mathrm{law}(Y \,|\, P_X)\big).$$

Expected calibration error (ECE)

Definition

The expected calibration error (ECE) with respect to distance measure d is defined as

$$\mathrm{ECE}_d := \mathbb{E}_{P_X} d\big(P_X, \mathrm{law}(Y \,|\, P_X)\big).$$

Choice of distance measure d

- For classification typically (semi-)metrics on the probability simplex (e.g., cityblock, Euclidean, or squared Euclidean distance)

Expected calibration error (ECE)

Definition

The expected calibration error (ECE) with respect to distance measure d is defined as

$$\mathrm{ECE}_d := \mathbb{E}_{P_X} d\big(P_X, \mathrm{law}(Y \,|\, P_X)\big).$$

Choice of distance measure d

- For classification typically (semi-)metrics on the probability simplex (e.g., cityblock, Euclidean, or squared Euclidean distance)
- For general probabilistic predictive models **statistical divergences**

Statistical divergences

Definition

Let $\mathcal{P}$ be a space of probability distributions. A function $d \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ that satisfies

- $d(P, Q) \geq 0$ for all $P, Q \in \mathcal{P}$,
- $d(P, Q) = 0$ if and only if $P = Q$,

is a statistical divergence.

Statistical divergences

Definition

Let $\mathcal{P}$ be a space of probability distributions. A function $d\colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ that satisfies

- $d(P, Q) \geq 0$ for all $P, Q \in \mathcal{P}$,
- $d(P, Q) = 0$ if and only if $P = Q$,

is a statistical divergence.

Note

- d dœs not need to be symmetric
- d dœs not need to satisfy the triangle inequality

Statistical divergences

Definition

Let $\mathcal{P}$ be a space of probability distributions. A function $d\colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ that satisfies

- $d(P, Q) \geq 0$ for all $P, Q \in \mathcal{P}$,
- $d(P, Q) = 0$ if and only if $P = Q$,

is a statistical divergence.

Note

- d dœs not need to be symmetric
- d dœs not need to satisfy the triangle inequality

Examples

- f-divergences, e.g., Kullback-Leibler divergence or total variation distance
- Wasserstein distance

Scoring rules: Definition

Definition

The expected score of a probabilistic predictive model P is defined as

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y)$$

where **scoring rule $s(p, y)$** is the reward of prediction p if the true outcome is y.

Scoring rules: Definition

Definition

The expected score of a probabilistic predictive model P is defined as

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y)$$

where **scoring rule $\boldsymbol{s(p, y)}$** is the reward of prediction p if the true outcome is y.

Examples for classification

- Brier score: $s(p, y) = -\int_\Omega \big((\delta_y - p)^2\big)(\mathrm{d}\omega)$
- Logarithmic score: $s(p, y) = \log p\big(\{y\}\big)$

Scoring rules: Decomposition

For proper scoring rules

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y) = \mathbb{E}_{P_X}\, d(\mathrm{law}(Y), \mathrm{law}(Y \mid P_X))$$

$$- \mathbb{E}_{P_X}\, d(P_X, \mathrm{law}(Y \mid P_X)) - S(\mathrm{law}(Y), \mathrm{law}(Y))$$

Expected score of P under Q

$$S(P, Q) := \int_\Omega s(P, \omega)\, Q(\mathrm{d}\omega)$$

Score divergence

$$d(P, Q) = S(Q, Q) - S(P, Q)$$

J. Bröcker. "Reliability, sufficiency, and the decomposition of proper scores." In: *Quarterly Journal of the Royal Meteorological Society* 135.643 (July 2009)

Scoring rules: Decomposition

For proper scoring rules

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y) = \underbrace{\mathbb{E}_{P_X}\, d(\mathrm{law}(Y), \mathrm{law}(Y \,|\, P_X))}_{\text{resolution}}$$
$$- \mathbb{E}_{P_X}\, d(P_X, \mathrm{law}(Y \,|\, P_X)) - S(\mathrm{law}(Y), \mathrm{law}(Y))$$

Expected score of P under Q

$$S(P, Q) := \int_\Omega s(P, \omega)\, Q(\mathrm{d}\omega)$$

Score divergence

$$d(P, Q) = S(Q, Q) - S(P, Q)$$

J. Bröcker. "Reliability, sufficiency, and the decomposition of proper scores." In: *Quarterly Journal of the Royal Meteorological Society* 135.643 (July 2009)

Scoring rules: Decomposition

For proper scoring rules

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y) = \underbrace{\mathbb{E}_{P_X}\, d(\mathrm{law}(Y), \mathrm{law}(Y \mid P_X))}_{\text{resolution}} - \underbrace{\mathbb{E}_{P_X}\, d(P_X, \mathrm{law}(Y \mid P_X))}_{\textbf{calibration}} - S(\mathrm{law}(Y), \mathrm{law}(Y))$$

Expected score of P under Q

$$S(P, Q) := \int_\Omega s(P, \omega)\, Q(\mathrm{d}\omega)$$

Score divergence

$$d(P, Q) = S(Q, Q) - S(P, Q)$$

J. Bröcker. "Reliability, sufficiency, and the decomposition of proper scores." In: *Quarterly Journal of the Royal Meteorological Society* 135.643 (July 2009)

Scoring rules: Decomposition

For proper scoring rules

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y) = \underbrace{\mathbb{E}_{P_X}\, d(\mathrm{law}(Y), \mathrm{law}(Y\,|\,P_X))}_{\text{resolution}}$$
$$- \underbrace{\mathbb{E}_{P_X}\, d(P_X, \mathrm{law}(Y\,|\,P_X))}_{\textbf{calibration}} - \underbrace{S(\mathrm{law}(Y), \mathrm{law}(Y))}_{\text{uncertainty of } Y}$$

Expected score of P under Q

$$S(P,Q) := \int_\Omega s(P,\omega)\, Q(\mathrm{d}\omega)$$

Score divergence

$$d(P,Q) = S(Q,Q) - S(P,Q)$$

J. Bröcker. "Reliability, sufficiency, and the decomposition of proper scores." In: *Quarterly Journal of the Royal Meteorological Society* 135.643 (July 2009)

Scoring rules: Decomposition

For proper scoring rules

$$\mathbb{E}_{P_X,Y}\, s(P_X, Y) = \underbrace{\mathbb{E}_{P_X}\, d(\mathrm{law}(Y), \mathrm{law}(Y \mid P_X))}_{\text{resolution}}$$
$$- \underbrace{\mathbb{E}_{P_X}\, d(P_X, \mathrm{law}(Y \mid P_X))}_{\textbf{calibration}} - \underbrace{S(\mathrm{law}(Y), \mathrm{law}(Y))}_{\text{uncertainty of } Y}$$

Expected score of P under Q

$$S(P, Q) := \int_\Omega s(P, \omega)\, Q(\mathrm{d}\omega)$$

Score divergence

$$d(P, Q) = S(Q, Q) - S(P, Q)$$

Models can trade off calibration for resolution!

J. Bröcker. “Reliability, sufficiency, and the decomposition of proper scores.” In: *Quarterly Journal of the Royal Meteorological Society* 135.643 (July 2009)

An alternative definition of calibration

Theorem

A probabilistic predictive model P is calibrated if

$$(P_X, Y) \stackrel{d}{=} (P_X, Z_X),$$

where $Z_X \sim P_X$.

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

An alternative definition of calibration

Theorem

A probabilistic predictive model P is calibrated if

$$(P_X, Y) \stackrel{d}{=} (P_X, Z_X),$$

where $Z_X \sim P_X$.

Calibration error as distance between $\mathrm{law}\big((P_X, Y)\big)$ and $\mathrm{law}\big((P_X, Z_X)\big)$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Calibration error: Integral probability metric

$$\mathrm{CE}_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{P_X,Y} f(P_X, Y) - \mathbb{E}_{P_X,Z_X} f(P_X, Z_X) \right|$$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations*. 2021

Calibration error: Integral probability metric

$$\mathrm{CE}_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{P_X, Y} f(P_X, Y) - \mathbb{E}_{P_X, Z_X} f(P_X, Z_X) \right|$$

Examples

- 1-Wasserstein distance: $\mathcal{F} = \{f : \|f\|_{\mathrm{Lip}} \leq 1\}$
- Total variation distance: $\mathcal{F} = \{f : \|f\|_{\infty} \leq 1\}$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Calibration error: Integral probability metric

$$\mathrm{CE}_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{P_X, Y} f(P_X, Y) - \mathbb{E}_{P_X, Z_X} f(P_X, Z_X) \right|$$

Examples

- 1-Wasserstein distance: $\mathcal{F} = \{f : \|f\|_{\mathrm{Lip}} \leq 1\}$
- Total variation distance: $\mathcal{F} = \{f : \|f\|_{\infty} \leq 1\}$

Common choices of ECE_d in classification can be formulated in this way

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Kernel calibration error: Maximum mean discrepancy (MMD)

Choose $\mathcal{F} = \{f \in \mathcal{H} \colon \|f\|_{\mathcal{H}} \leq 1\}$ for some reproducing kernel Hilbert space $\mathcal{H}$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Kernel calibration error: Maximum mean discrepancy (MMD)

Choose $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ for some reproducing kernel Hilbert space $\mathcal{H}$

Reproducing kernel Hilbert space (RKHS)

- Hilbert space of functions that satisfy f close to $g \Rightarrow f(x)$ close to $g(x)$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Kernel calibration error: Maximum mean discrepancy (MMD)

Choose $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ for some reproducing kernel Hilbert space $\mathcal{H}$

Reproducing kernel Hilbert space (RKHS)

- Hilbert space of functions that satisfy f close to $g \Rightarrow f(x)$ close to $g(x)$
- Possesses a positive-definite function k as reproducing kernel

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Kernel calibration error: Maximum mean discrepancy (MMD)

Choose $\mathcal{F} = \{f \in \mathcal{H} \colon \|f\|_{\mathcal{H}} \leq 1\}$ for some reproducing kernel Hilbert space $\mathcal{H}$

Reproducing kernel Hilbert space (RKHS)

- Hilbert space of functions that satisfy f close to $g \Rightarrow f(x)$ close to $g(x)$
- Possesses a positive-definite function k as reproducing kernel

Definition

The kernel calibration error (KCE) of a model P with respect to kernel k is defined as

$$\mathrm{KCE}_k^2 := \mathrm{CE}_{\mathcal{F}}^2 = \int k\big((p, y), (\tilde{p}, \tilde{y})\big)\, \mu\big(\mathrm{d}(p, y)\big)\mu\big(\mathrm{d}(\tilde{p}, \tilde{y})\big),$$

where $\mu = \operatorname{law}\big((P_X, Y)\big) - \operatorname{law}\big((P_X, Z_X)\big)$.

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Choice of kernel

Observations

- Kernel k defined on the product space of predictions and targets

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Choice of kernel

Observations

- Kernel k defined on the product space of predictions and targets
- In multi-class classification, k can be identified with a matrix-valued kernel on the space of predictions

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Choice of kernel

Observations

- Kernel k defined on the product space of predictions and targets
- In multi-class classification, k can be identified with a matrix-valued kernel on the space of predictions
- For specific kernel choices, Z_X can be integrated out analytically

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations*. 2021

Choice of kernel

Observations

- Kernel k defined on the product space of predictions and targets
- In multi-class classification, k can be identified with a matrix-valued kernel on the space of predictions
- For specific kernel choices, Z_X can be integrated out analytically
- Otherwise numerical integration methods (e.g., Monte Carlo integration) can be used to integrate out Z_X

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations*. 2021

Choice of kernel

Observations

- Kernel k defined on the product space of predictions and targets
- In multi-class classification, k can be identified with a matrix-valued kernel on the space of predictions
- For specific kernel choices, Z_X can be integrated out analytically
- Otherwise numerical integration methods (e.g., Monte Carlo integration) can be used to integrate out Z_X
- Suggestive to use tensor product kernels $k = k_{\mathcal{P}} \otimes k_{\mathcal{Y}}$, where $k_{\mathcal{P}}$ and $k_{\mathcal{Y}}$ are kernels on the space of predictions and targets, respectively

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Tensor product kernel

Construction of $k_{\mathcal{P}}$ with Hilbertian metrics

- For Hilbertian metrics of form $d_{\mathcal{P}}(p, \tilde{p}) = \|\phi(p) - \phi(\tilde{p})\|_2$ for some $\phi \colon \mathcal{P} \to \mathbb{R}^d$,

$$k_{\mathcal{P}}(p, \tilde{p}) = \exp\big(-\lambda d_{\mathcal{P}}^{\nu}(p, \tilde{p})\big), \tag{1}$$

  is valid kernel on the space of predictions for $\lambda > 0$ and $\nu \in (0, 2]$

D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests beyond classification.” In: *International Conference on Learning Representations*. 2021

Tensor product kernel

Construction of $k_{\mathcal{P}}$ with Hilbertian metrics

- For Hilbertian metrics of form $d_{\mathcal{P}}(p, \tilde{p}) = \|\phi(p) - \phi(\tilde{p})\|_2$ for some $\phi \colon \mathcal{P} \to \mathbb{R}^d$,

$$k_{\mathcal{P}}(p, \tilde{p}) = \exp\big(-\lambda d_{\mathcal{P}}^{\nu}(p, \tilde{p})\big), \tag{1}$$

  is valid kernel on the space of predictions for $\lambda > 0$ and $\nu \in (0, 2]$
- Parameterization of predictions gives rise to ϕ naturally

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Tensor product kernel

Construction of $k_{\mathcal{P}}$ with Hilbertian metrics

- For Hilbertian metrics of form $d_{\mathcal{P}}(p, \tilde{p}) = \|\phi(p) - \phi(\tilde{p})\|_2$ for some $\phi \colon \mathcal{P} \to \mathbb{R}^d$,

$$k_{\mathcal{P}}(p, \tilde{p}) = \exp\big(-\lambda d_{\mathcal{P}}^{\nu}(p, \tilde{p})\big), \tag{1}$$

  is valid kernel on the space of predictions for $\lambda > 0$ and $\nu \in (0, 2]$
- Parameterization of predictions gives rise to ϕ naturally
- For many mixture models, Hilbertian metrics of model components can be lifted to Hilbertian metric of mixture models

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests beyond classification." In: *International Conference on Learning Representations.* 2021

Estimation of calibration errors

Estimation of calibration errors

Task

Estimate the calibration error of a model P from a validation dataset $(X_i, Y_i)_{i=1,\ldots,n}$ of features and corresponding targets.

Estimation of calibration errors

Task

Estimate the calibration error of a model P from a validation dataset $(X_i, Y_i)_{i=1,\ldots,n}$ of features and corresponding targets.

Dataset of predictions and targets sufficient

- Calibration (errors) defined based only on predictions and targets
- Estimation can be performed with dataset (P_{X_i}, Y_i) of predictions and corresponding targets instead
- Highlights that structure of features and model is not relevant for calibration estimation

ECE: Estimation

Problem

The estimation of $\mathrm{law}(Y \mid P_X)$ is challenging.

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

ECE: Estimation

Problem

The estimation of $\operatorname{law}(Y \,|\, P_X)$ is challenging.

Binning predictions

- Common approach in classification
- Often leads to **biased and inconsistent** estimators

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

ECE: Experiments

10-class classification

For three models **M1**, **M2** and **M3**, 10^4 synthetic datasets $(P_{X_i}, Y_i)_{i=1,\ldots,250}$ are sampled according to

- $P_{X_i} = \mathrm{Cat}(p_i)$ with $p_i \sim \mathrm{Dir}(0.1, \ldots, 0.1)$,

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

ECE: Experiments

10-class classification

For three models **M1**, **M2** and **M3**, 10^4 synthetic datasets $(P_{X_i}, Y_i)_{i=1,\ldots,250}$ are sampled according to

- $P_{X_i} = \mathrm{Cat}(p_i)$ with $p_i \sim \mathrm{Dir}(0.1, \ldots, 0.1)$,
- Y_i conditionally on P_{X_i} from

$$\mathbf{M1}: P_{X_i}, \qquad \mathbf{M2}: 0.5 P_{X_i} + 0.5\delta_1, \qquad \mathbf{M3}: \mathrm{U}(\{1, \ldots, 10\}).$$

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

ECE: Experiments

10-class classification

For three models **M1**, **M2** and **M3**, 10^4 synthetic datasets $(P_{X_i}, Y_i)_{i=1,\ldots,250}$ are sampled according to

- $P_{X_i} = \mathrm{Cat}(p_i)$ with $p_i \sim \mathrm{Dir}(0.1, \ldots, 0.1)$,
- Y_i conditionally on P_{X_i} from

$$\mathbf{M1}: P_{X_i}, \qquad \mathbf{M2}: 0.5 P_{X_i} + 0.5\delta_1, \qquad \mathbf{M3}: \mathrm{U}(\{1, \ldots, 10\}).$$

Model **M1** is calibrated, and models **M2** and **M3** are uncalibrated.

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

ECE: Experiments

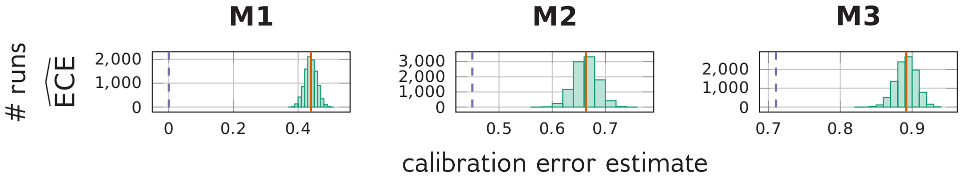
10-class classification

For three models **M1**, **M2** and **M3**, 10^4 synthetic datasets $(P_{X_i}, Y_i)_{i=1,\ldots,250}$ are sampled according to

- $P_{X_i} = \mathrm{Cat}(p_i)$ with $p_i \sim \mathrm{Dir}(0.1, \ldots, 0.1)$,
- Y_i conditionally on P_{X_i} from

$$\textbf{M1}: P_{X_i}, \qquad \textbf{M2}: 0.5P_{X_i} + 0.5\delta_1, \qquad \textbf{M3}: \mathrm{U}(\{1, \ldots, 10\}).$$

Model **M1** is calibrated, and models **M2** and **M3** are uncalibrated.



D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

Kernel calibration error: Estimation

- For the MMD unbiased and consistent estimators are available

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

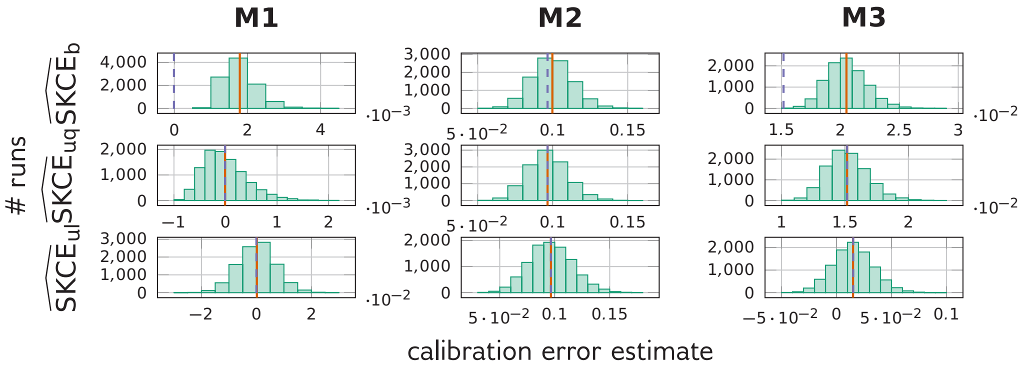
Kernel calibration error: Estimation

- For the MMD unbiased and consistent estimators are available
- Variance can be reduced by marginalizing out Z_X

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

Kernel calibration error: Estimation

- For the MMD unbiased and consistent estimators are available
- Variance can be reduced by marginalizing out Z_X



D. Widmann, F. Lindsten, and D. Zachariah. “Calibration tests in multi-class classification: A unifying framework.” In: *Advances in Neural Information Processing Systems 32*. 2019

Calibration tests

Problems with calibration errors

- Calibration errors have no meaningful unit or scale

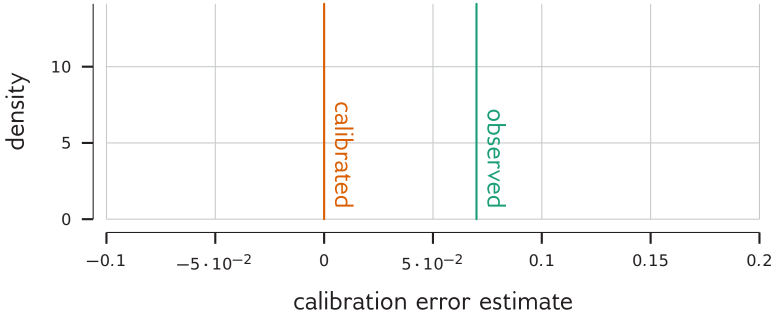
Problems with calibration errors

- Calibration errors have no meaningful unit or scale
- Different calibration errors rank models differently

Problems with calibration errors

- Calibration errors have no meaningful unit or scale
- Different calibration errors rank models differently
- Calibration error estimators are random variables

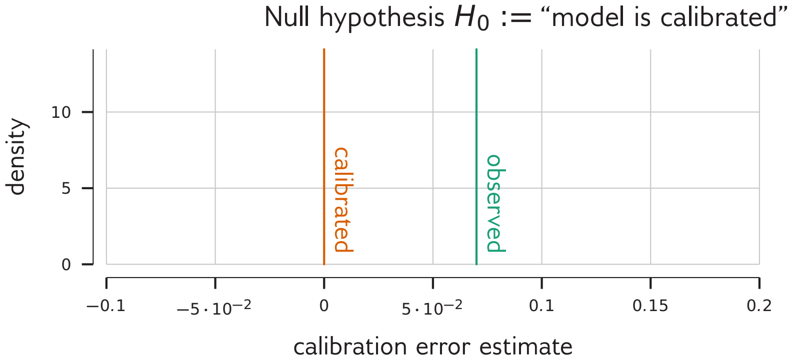
Calibration tests



J. Bröcker and L. A. Smith. “Increasing the reliability of reliability diagrams.” In: *Weather and Forecasting* (2007)

J. Vaicenavicius et al. “Evaluating model calibration in classification.” In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

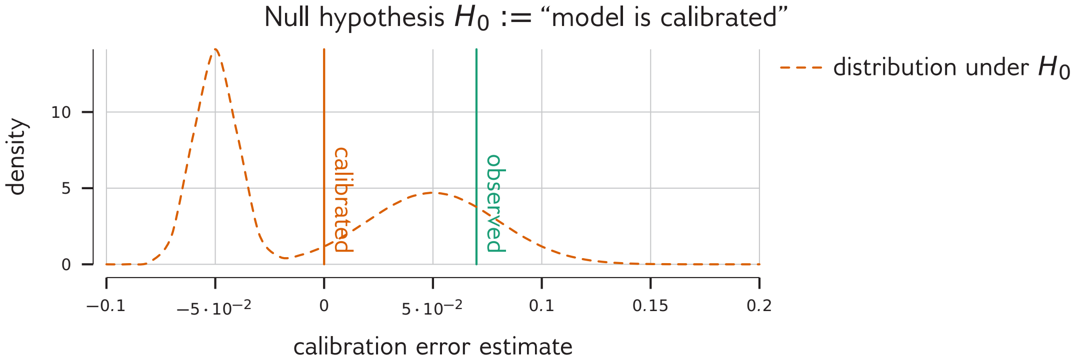
Calibration tests



J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

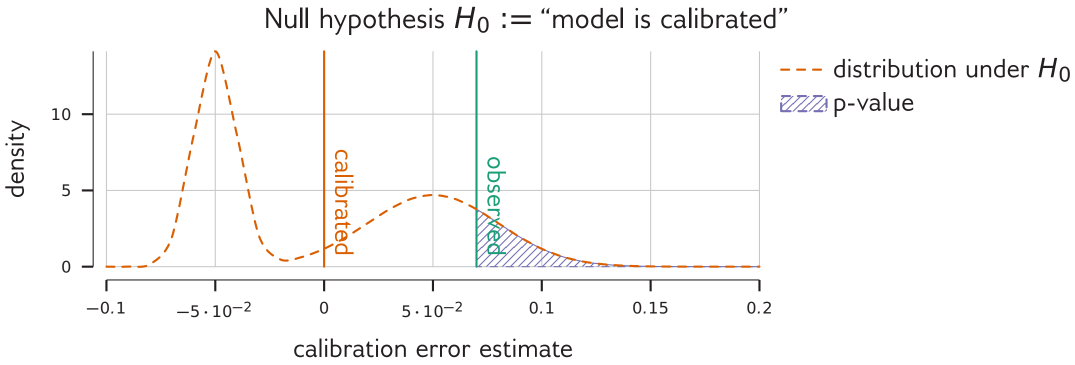
Calibration tests



J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

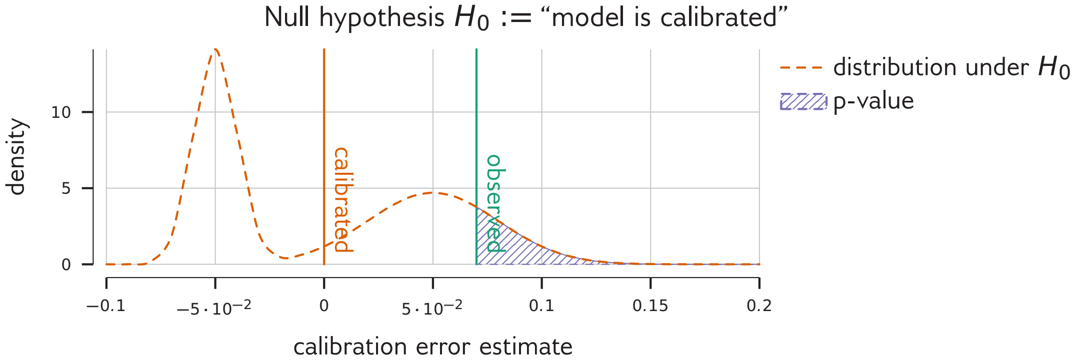
Calibration tests



J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

J. Vaicenavicius et al. "Evaluating model calibration in classification." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Calibration tests



Reject H_0 if p-value is small

J. Bröcker and L. A. Smith. “Increasing the reliability of reliability diagrams.” In: *Weather and Forecasting* (2007)

J. Vaicenavicius et al. “Evaluating model calibration in classification.” In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

Consistency resampling

Original dataset

$$\boxed{\begin{matrix} y_1 & P_{x_1} \\ \vdots & \vdots \\ y_n & P_{x_n} \end{matrix}}$$

J. Bröcker and L. A. Smith. “Increasing the reliability of reliability diagrams.” In: *Weather and Forecasting* (2007)

Consistency resampling

Original dataset

$$\boxed{\begin{matrix} y_1 & P_{x_1} \\ \vdots & \vdots \\ y_n & P_{x_n} \end{matrix}} \xrightarrow{i_j \sim \mathrm{U}(\{1, \ldots, n\})} \boxed{\begin{matrix} y_{i_1} & P_{x_{i_1}} \\ \vdots & \vdots \\ y_{i_n} & P_{x_{i_n}} \end{matrix}}$$

J. Bröcker and L. A. Smith. “Increasing the reliability of reliability diagrams.” In: *Weather and Forecasting* (2007)

Consistency resampling

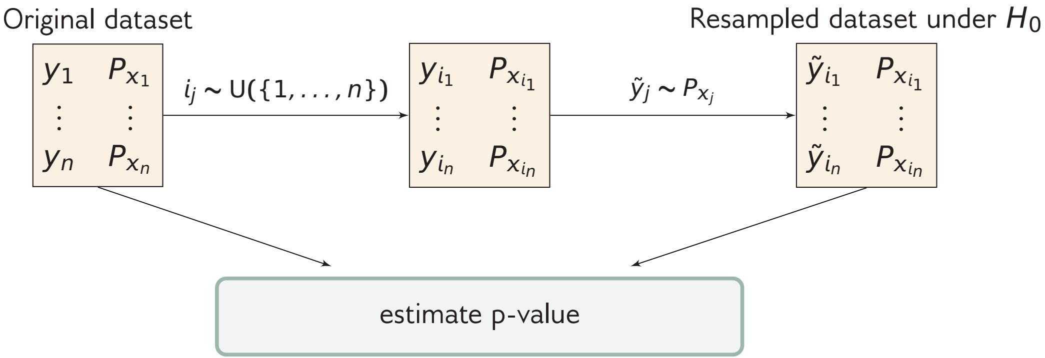
Original dataset

Resampled dataset under H_0

$$\boxed{\begin{array}{cc} y_1 & P_{x_1} \\ \vdots & \vdots \\ y_n & P_{x_n} \end{array}} \xrightarrow{i_j \sim \mathrm{U}(\{1,\ldots,n\})} \boxed{\begin{array}{cc} y_{i_1} & P_{x_{i_1}} \\ \vdots & \vdots \\ y_{i_n} & P_{x_{i_n}} \end{array}} \xrightarrow{\tilde{y}_j \sim P_{x_j}} \boxed{\begin{array}{cc} \tilde{y}_{i_1} & P_{x_{i_1}} \\ \vdots & \vdots \\ \tilde{y}_{i_n} & P_{x_{i_n}} \end{array}}$$

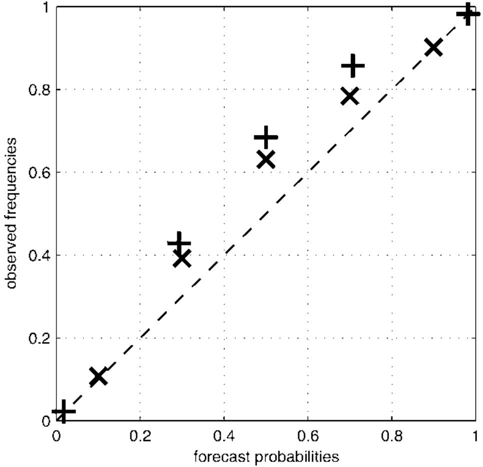
J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

Consistency resampling



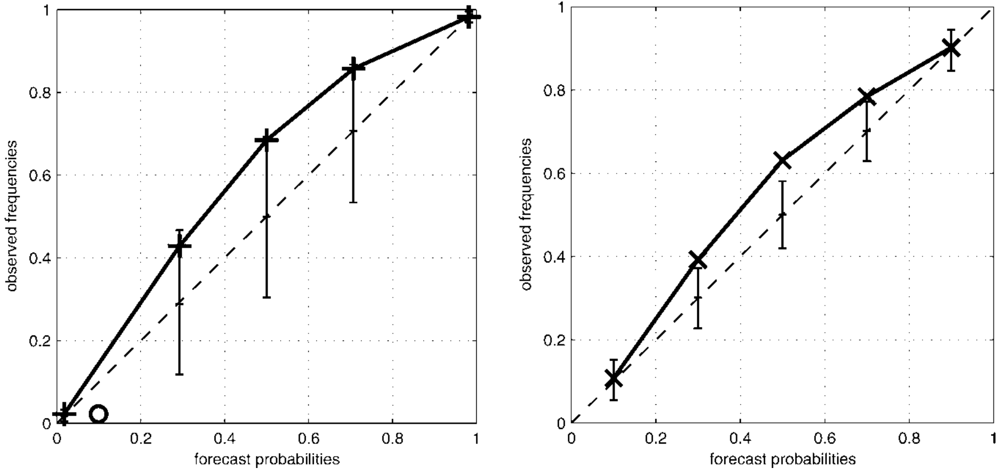
J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

Consistency bars



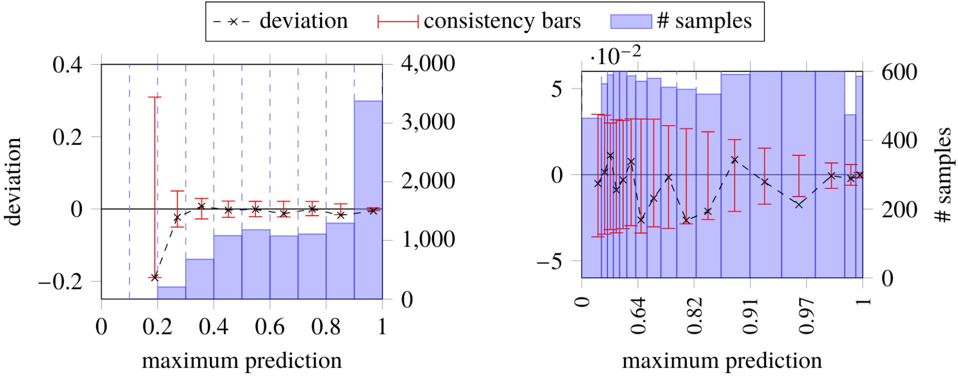
J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

Consistency bars



J. Bröcker and L. A. Smith. "Increasing the reliability of reliability diagrams." In: *Weather and Forecasting* (2007)

Variant



(a) Equally-sized bins

(b) Data-dependent bins

J. Vaicenavicius et al. “Evaluating model calibration in classification.” In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Apr. 2019

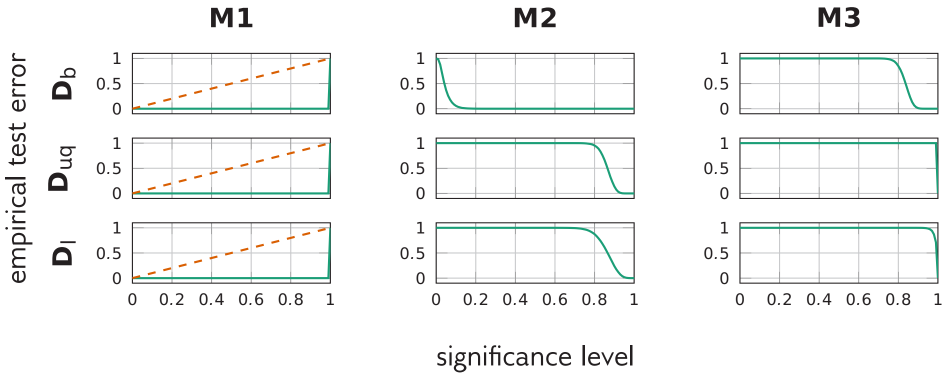
Kernel calibration error: Distribution-free tests

Upper bound the p-value

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

Kernel calibration error: Distribution-free tests

Upper bound the p-value



D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32.* 2019

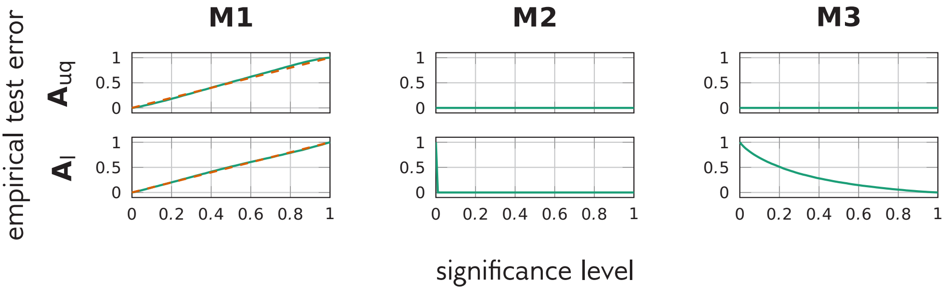
Kernel calibration error: Asymptotic tests

Approximate the p-value based on the **asymptotic** distribution

D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

Kernel calibration error: Asymptotic tests

Approximate the p-value based on the **asymptotic** distribution



D. Widmann, F. Lindsten, and D. Zachariah. "Calibration tests in multi-class classification: A unifying framework." In: *Advances in Neural Information Processing Systems 32*. 2019

Calibration: Software packages

`CalibrationAnalysis.jl`

Summary

- Suite for analyzing calibration of probabilistic predictive models
- Written in Julia, with interfaces in Python (`pycalibration`) and R (`rcalibration`)

`CalibrationAnalysis.jl`

Summary

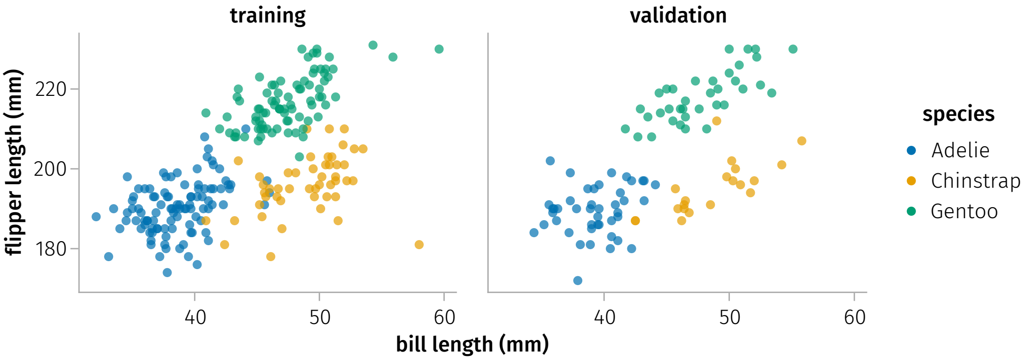
- Suite for analyzing calibration of probabilistic predictive models
- Written in Julia, with interfaces in Python (`pycalibration`) and R (`rcalibration`)

Features

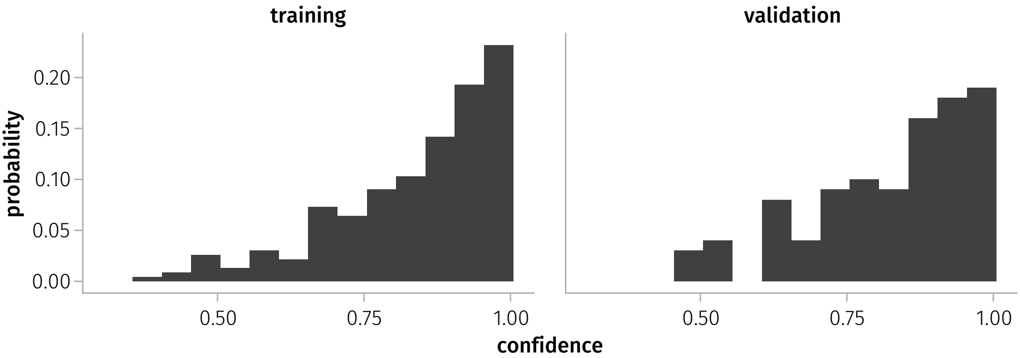
- Supports classification and regression models
- Reliability diagrams (`ReliabilityDiagrams.jl`)
- Estimation of calibration errors such as ECE and KCE (`CalibrationErrors.jl`)
- Calibration tests (`CalibrationTests.jl`)
- Integration with Julia ecosystem: Supports `Plots.jl` and `Makie.jl`, `KernelFunctions.jl`, and `HypothesisTests.jl`

Calibration analysis: Penguins example

We train a naive Bayes classifier of penguin species based on bill depth, bill length, flipper length, and body mass.



Binary predictions

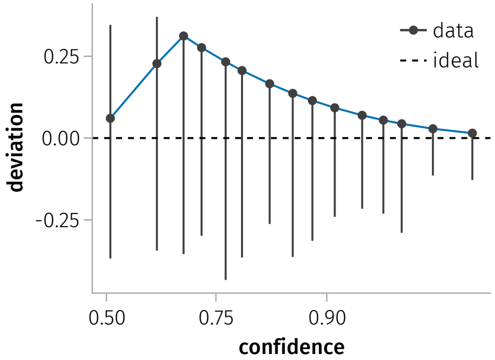


Reliability diagram

Code

```
julia> using CalibrationAnalysis, CairoMakie

julia> reliability(
           confidence,
           outcome;
           binning=EqualMass(; n=15),
           deviation=true,

           ↪  consistencybars=ConsistencyBars(),
       )
```

Polished result



Expected calibration error: Code

```julia
julia> ece = ECE(UniformBinning(5), TotalVariation());

julia> ece(confidence, outcome)
0.06594437403598197

julia> ece(predictions, observations)
0.15789651955832515
```

Kernel calibration error: Code

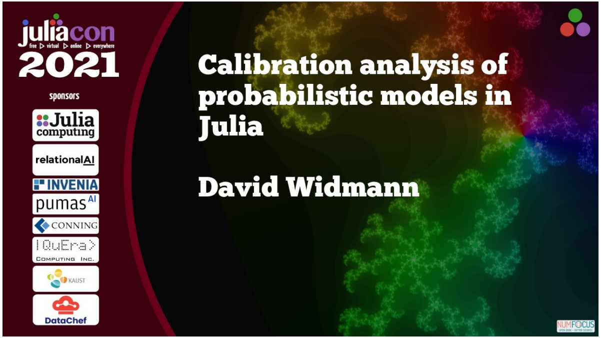
```julia
julia> kernel = GaussianKernel() ⊗ WhiteKernel();

julia> skce = SKCE(kernel);

julia> skce(predictions, observations)
0.0032631144705774404

julia> skce = SKCE(kernel; unbiased=false);

julia> skce(predictions, observations)
0.0042021131168416 22

julia> skce = SKCE(kernel; blocksize=5);

julia> skce(predictions, observations)
-0.005037270862051889
```

Calibration test: Code

```
julia> AsymptoticSKCETest(kernel, predictions, observations)
Asymptotic SKCE test
--------------------
Population details:
    parameter of interest:   SKCE
    value under h_0:         0.0
    point estimate:          0.00326311

Test summary:
    outcome with 95% confidence: reject h_0
    one-sided p-value:           0.0150

Details:
    test statistic: -0.0009060378940361157

julia> test = ConsistencyTest(ece, predictions, observations);

julia> pvalue(test; bootstrap_iters=10_000)
0.0188
```

Additional resources

- Online documentation: `https://devmotion.github.io/CalibrationErrors.jl/`
- Talk at JuliaCon 2021: `https://youtu.be/PrLsXFvwzuA`



Slides available at `https://talks.widmann.dev/2021/07/calibration/`

Concluding remarks

Important takeaways

- More fine-grained analysis of calibration can be important
- MMD-like kernel calibration error can be applied to probabilistic models beyond classification
- Estimators of kernel calibration error have appealing properties
- Calibration errors and reliability diagrams can be misleading