

Ch.12 Calibration

Conformal Prediction

Jikwang Kim

Seoul National University

January 5, 2026

Outline

- 1 Calibration: definition and methods
- 2 Properties of ECE and binned ECE
- 3 Properties of dCE
- 4 Venn-Abers Predictors

Outline

1 Calibration: definition and methods

2 Properties of ECE and binned ECE

3 Properties of dCE

4 Venn-Abers Predictors

Definition of Calibration

Definition 12.1: Perfect calibration

Consider a r.v. (X, Y) drawn from a distribution P on $\mathcal{X} \times \{0, 1\}$, and let $f : \mathcal{X} \rightarrow \{0, 1\}$ be an estimate of $\mathbb{P}(Y = 1 | X)$. The function f is **perfectly calibrated** if

$$\mathbb{P}(Y = 1 | f(X)) = f(X) \quad \text{a.s.}$$

Intuition) [Model Performance, Model Honesty]

Consider $\{X : f(X) = 0.2\}$, then the perfectly calibration satisfies

$$\mathbb{P}(Y = 1 | f(X) = 0.2) = 0.2$$

Example) [Oracle Model] $f(x) = \mathbb{P}(Y = 1 | X = x)$, [Constant Model] $f(x) = \mathbb{E}[Y]$, etc.

Question session 1

Q. In the book, well calibration does not imply that f gives a guarantee of accurate estimate for $\mathbb{P}(Y = 1 | X)$. Then, isn't it enough simply to control ECE for model prediction? Any additional procedure? A. Cite the notation in the book:

calibration is a conuterpart to conformal prediction, in that it is used to assess the quality of predictions from **machine learning models** and thus provide uncertainty quantification.

Cf)

| Method | Calibration | CP |
|--------------|---|---|
| Model | pre-trained | pre-trained |
| Data setting | $Y \in \{0, 1\}$ | $Y \in \mathbb{R}$ |
| Target | $\mathbb{P}[Y = 1 X]$ | Y |
| Goal | Gaurantee honesty of the model $\mathbb{E}[Y f(X)] = f(X)$ | Gaurantee coverage $\mathbb{P}[Y \in \hat{C}] \geq 1 - \alpha$ |

Goal : modify a given pretrained function to (approximately) satisfy perfect calibration.

Post-hoc Calibration Algorithm

For pretrained model $f : \mathcal{X} \rightarrow [0, 1]$, and n data points $(X_1, Y_1), \dots, (X_n, Y_n)$ $\stackrel{\text{i.i.d.}}{\sim} P$ that were not used for model training, (like the calibration set in the split CP)
we want to find $h : [0, 1] \rightarrow [0, 1]$ s.t.

$$h \circ f \quad (\text{approximately}) \text{ satisfies calibration.}$$

The three algorithms take as input a function f and calibration data and output such an h from some class of functions \mathcal{H} .

- ① Binning
- ② Isotonic regression
- ③ Temperature scaling

Binning

Concept) Grouping the probability values generated by the model f into intervals (bins), and overwriting the probabilities in each interval with the actual $Y = 1$ ratio.

Post-hoc Calibration: Binning

Let $[0, 1] = B_1 \cup \dots \cup B_K$ for some K , and $k(z)$ be the index of the bin containing $z \in [0, 1]$. For $n_k = \sum_{i=1}^n 1_{\{f(X_i) \in B_k\}}$, define

$$\hat{h}(z) = \begin{cases} \frac{1}{n_k} \sum_{i=1}^n Y_i \cdot 1_{\{f(X_i) \in B_{k(z)}\}}, & n_{k(z)} \geq 1, \\ \frac{1}{2}, & n_{k(z)} = 0. \end{cases}$$

This is equivalent to minimizing the squared error of $h \circ f$ on the calibration data:

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \sum_{i=1}^n ([h \circ f](X_i) - Y_i)^2$$

where $\mathcal{H} = \mathcal{H}_{\text{bin}}$, the set of functions that are constant within each bin B_k .

Post-hoc Calibration: Isotonic Regression

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \sum_{i=1}^n ([h \circ f](X_i) - Y_i)^2$$

where $\mathcal{H} = \mathcal{H}_{\text{iso}}$, the set of all nondecreasing functions.

It is reasonable why the initial model f is assumed to have larger outputs when $Y = 1$ confidence is higher.

Pros and Cons)

- don't need to specify bins.
- requires less calibration data to achieve a stable fit.
- only produce nondecreasing h .

Post-hoc Calibration: Temperature Scaling

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \sum_{i=1}^n ([h \circ f](X_i) - Y_i)^2$$

where $\mathcal{H} = \mathcal{H}_{\text{logistic}} = \{h_{\beta_0, \beta_1} : \beta_0, \beta_1 \in \mathbb{R}\}$, and

$$h_{\beta_0, \beta_1}(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{logit}(z))}}, \quad \text{logit}(z) = \log \left(\frac{z}{1-z} \right)$$

Pros and Cons)

- requires less calibration data to achieve a stable fit. (only two parameters)
- only produce logistic function h .

Quantifying violations of perfect calibration

Definition: Calibration Measure

- The Expected Calibration Error:

$$\text{ECE}(f) := \mathbb{E}[|\mathbb{E}[Y | f(X)] - f(X)|]$$

- The binned ECE:

$$\text{binECE}(f) := \sum_{k=1}^K |\mathbb{E}[Y | f(X) \in B_k] - \mathbb{E}[f(X) | f(X) \in B_k]| \cdot \mathbb{P}(f(X) \in B_k),$$

where $[0, 1] = B_1 \cup \dots \cup B_K$ is some partition of the unit interval.

They are defined w.r.t. a probability distribution P , which is unknown and should be estimated.

Quantifying violations of perfect calibration

Definition: Empirically estimated calibration measure

The natural plug-in estimator of binECE:

$$\widehat{\text{binECE}}(f) = \sum_{k=1}^K \left| \frac{1}{n_k} \cdot \sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} (Y_i - f(X_i)) \right| \cdot \frac{n_k}{n},$$

where $n_k = |\{i : f(X_i) \in B_k\}|$.

Argue 1) Binned ECE (and its empirical estimate $\widehat{\text{binECE}}$) can be arbitrarily far from ECE.

Definition: Calibration Measure (Alternative Relaxation)

The distance to Calibration Error:

$$\text{dCE}(f) := \inf_{\substack{g: \mathcal{X} \rightarrow [0,1] \\ \mathbb{E}[Y|g(X)] = g(X)}} \mathbb{E}[|g(X) - f(X)|].$$

Argue 2) It is possible to give distribution-free confidence bounds for dCE, rather than ECE.

Outline

1 Calibration: definition and methods

2 Properties of ECE and binned ECE

3 Properties of dCE

4 Venn-Abers Predictors

Discontinuity of ECE

Example)

Let $(X, Y) \sim P$, where $X \sim \text{Unif}[0, 1]$, and

$$\mathbb{P}(Y = 1 \mid X) = \begin{cases} 0, & X \in [0, \frac{1}{4}] \cup [\frac{3}{4}, 1], \\ 1, & X \in (\frac{1}{4}, \frac{3}{4}). \end{cases}$$

Then the constant function $f(x) = \frac{1}{2}$ is perfectly calibrated, but for arbitrarily small ϵ ,

$$f_\epsilon(x) = \frac{1 - \epsilon}{2} + \epsilon x$$

is highly miscalibrated in terms of its ECE.

$$\therefore \|f - f_\epsilon\|_\infty \leq \epsilon, \text{ and } \|\text{ECE}(f) - \text{ECE}(f_\epsilon)\|_2 \not\leq \delta \quad \forall \delta,$$

and now we can consider the binned ECE instead, since its discretizations are easier to control the continuity errors.

binned ECE $\not\approx$ ECE

Prop 12.3: binned ECE cannot be larger than ECE

For any distribution P on $\mathcal{X} \times \{0, 1\}$, any function $f : \mathcal{X} \rightarrow [0, 1]$, and any partition $[0, 1] = B_1 \cup \dots \cup B_K$,

$$\text{binECE}(f) \leq \text{ECE}(f).$$

proof)

$$\begin{aligned}\text{binECE}(f) &= \sum_{k=1}^K |\mathbb{E}[Y - f(X) \mid f(X) \in B_k]| \cdot \mathbb{P}(f(X) \in B_k) \\ &= \sum_{k=1}^K \mathbb{E}[\mathbb{E}[|Y - f(X) \mid k(f(X))| \mid k(f(X)) = k]] \cdot \mathbb{P}(k(f(X)) = k) \\ &= \mathbb{E}|\mathbb{E}[Y - f(X) \mid k(f(X))]| \\ &= \mathbb{E}|\mathbb{E}[\mathbb{E}[Y - f(X) \mid f(X)] \mid k(f(X))]| \quad (\because \sigma(k(f(X))) \subset \sigma(f(X))) \\ &\leq \mathbb{E}[\mathbb{E}[|\mathbb{E}[Y - f(X) \mid f(X)]| \mid k(f(X))]] \quad (\because \text{by Jensen, since } |\cdot| \text{ is convex}) \\ &= \mathbb{E}|\mathbb{E}[Y - f(X) \mid f(X)]| \\ &= \text{ECE}(f)\end{aligned}$$



Cf) This inequality can be extremely loose. (Since we can control the partition arbitrarily.)

binECE Estimation

Since we want to certificate that the ECE (or binECE) is sufficiently small, then, in this chapter, our goal is to find the upper bound of them.

Thm 12.4: binECE Estimation

Fixed n , and $f : \mathcal{X} \rightarrow [0, 1]$, let $[0, 1]$ be a fixed partition. For $\widehat{\text{binECE}}$, any distribution P on $\mathcal{X} \times \{0, 1\}$, and $\delta \in [0, 1]$,

$$\mathbb{P} \left(\widehat{\text{binECE}}(f) \geq \text{binECE}(f) - \sqrt{\frac{2 \log(1/\delta)}{n}} \right) \geq 1 - \delta$$

and

$$\mathbb{P} \left(\widehat{\text{binECE}}(f) \leq \text{binECE}(f) + \sqrt{\frac{K}{n}} + \sqrt{\frac{2 \log(1/\delta)}{n}} \right) \geq 1 - \delta$$

That is,

- distribution-free upper confidence bound

$$\text{binECE}(f) \leq \widehat{\text{binECE}}(f) + \sqrt{\frac{2 \log(1/\delta)}{n}} \quad \text{w.p. } 1 - \delta$$

- fair tightness of upper bound

$$\widehat{\text{binECE}}(f) + \sqrt{\frac{2 \log(1/\delta)}{n}} \leq \text{binECE}(f) + \sqrt{\frac{K}{n}} + 2\sqrt{\frac{2 \log(1/\delta)}{n}} \quad \text{w.p. } 1 - \delta$$

binECE Estimation

proof)

Let $\mu_k = \mathbb{E}[Y \mid f(X) \in B_k]$, $\mu_k^f = \mathbb{E}[f(X) \mid f(X) \in B_k]$, $p_k = \mathbb{P}(X \in B_K)$, and $n_k = |\{i \in [n] : f(X_i) \in B_k\}|$ for $k = 1, \dots, K$. Given n_k , for any k , (here, assuming $n_k \geq 1$)

$$\sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} (Y_i - f(X_i))$$

is a sum of n_k many i.i.d. terms, each with mean $\mu_k - \mu_k^f$ and variance ≤ 1 . Then,

$$n_k |\mu_k - \mu_k^f| \leq \mathbb{E} \left[\left| \sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} (Y_i - f(X_i)) \right| \middle| n_k \right] \leq n_k |\mu_k - \mu_k^f| + \sqrt{n_k}$$

Hence, marginalizing over n_k , since $(n_1, \dots, n_K) \sim \text{Multi}(n, (p_1, \dots, p_K))$,

$$\begin{aligned} \mathbb{E} [\widehat{\text{binECE}}(f)] &= \mathbb{E} \left[\frac{1}{n} \sum_{k=1}^K \left| \sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} (Y_i - f(X_i)) \right| \right] \\ &\geq \mathbb{E} \left[\sum_{k=1}^K \frac{n_k}{n} |\mu_k - \mu_k^f| \right] = \sum_{k=1}^K p_k |\mu_k - \mu_k^f| = \text{binECE}(f), \end{aligned}$$

and

$$\begin{aligned}\mathbb{E} \left[\widehat{\text{binECE}}(f) \right] &\leq \mathbb{E} \left[\sum_{k=1}^K \frac{n_k}{n} |\mu_k - \mu_k^f| + \frac{\sqrt{n_k}}{n} \right] = \text{binECE}(f) + \sum_{k=1}^K \frac{\mathbb{E}[\sqrt{n_k}]}{n} \\ &\leq \text{binECE}(f) + \sum_{k=1}^K \frac{\sqrt{np_k}}{n} \quad (\because \text{Jensen's ineq.}) \\ &\leq \text{binECE}(f) + \sqrt{\frac{K}{n}} \quad (\because p_k \leq 1 \ \forall k.).\end{aligned}$$

Finally, for $Z_i = (X_i, Y_i)$, since

$$\sup_{z_1, \dots, z_n, z'_i} \left| \widehat{\text{binECE}}(z_1, \dots, z_i, \dots, z_n) - \widehat{\text{binECE}}(z_1, \dots, z'_i, \dots, z_n) \right| \leq \frac{2}{n},$$

by McDiarmid's inequality,

$$\mathbb{P} \left(\left| \widehat{\text{binECE}}(f) - \mathbb{E} \left[\widehat{\text{binECE}}(f) \right] \right| \geq \epsilon \right) < 2 \exp \left(-\frac{n^2 \epsilon^2}{2} \right)$$



ECE Estimation

Thm 12.5: ECE Estimation is uninformative.

Fixed n , and $f : \mathcal{X} \rightarrow [0, 1]$, let $\widehat{\text{ECE}}(f) \in [0, 1]$ be a function of the observed data, satisfying

$$\mathbb{P}\left(\widehat{\text{ECE}}(f) \geq \text{ECE}(f)\right) \geq 1 - \delta \quad \forall P.$$

Then, for any distribution P on $\mathcal{X} \times \{0, 1\}$ for which $f(X)$ has a nonatomic distribution,

$$\mathbb{P}\left(\widehat{\text{ECE}}(f) \geq \mathbb{E}[|Y - f(X)|]\right) \geq 1 - \delta.$$

(Recall) Thm 11.11: Conti. regr. ftn. est. is uninformative.

Suppose $\hat{\mu}$, and $\hat{\epsilon}$ satisfies the distribution-free validity for the estimation of μ_P . Then, for any distribution P on $\mathcal{X} \times [a, b]$ for which P_X has a nonatomic distribution,

$$\mathbb{P}\left(\hat{\epsilon} \geq \frac{\mathbb{E}_P[\text{Var}_P(Y | X)]}{b - a}\right) \geq 1 - \delta.$$

proof) (Use sample-resample technique.)

We saw that we cannot meaningfully estimate the ECE of given training model f , whether its output is distributed discretely, or continuously. Then, how about the ECE of $\hat{h} \circ f$?

Thm 12.6: Distribution-free ECE control for binning

Fixed the partition B_1, \dots, B_K of $[0, 1]$, let \hat{h} be the binning post-hoc adjustment. Then,

$$\mathbb{E} [\text{ECE}(\hat{h} \circ f)] \leq \sqrt{\frac{K}{2n}},$$

and moreover, for any $\delta \in [0, 1]$,

$$\mathbb{P} \left(\text{ECE}(\hat{h} \circ f) \leq \frac{1}{\sqrt{2\delta}} \cdot \sqrt{\frac{K}{n}} \right) \geq 1 - \delta.$$

In binning setting, since we discard some of the information in f , (especially, about small values between a bin,) we can estimate ECE, which is discontinuous.

Binning

proof)

(Recall) Thm 11.10: Discrete regression function estimation

Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{x_1, \dots, x_K\}$, and $\mathcal{Y} \subseteq [a, b]$, and let $\delta \in [0, 1]$.

Let $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{i.i.d.}}{\sim} P$, $n_k = \sum_{i=1}^n 1_{\{X_i=x_k\}}$, and

$$\hat{\mu}(x_k) = \frac{1}{n_k} \sum_{i=1}^n Y_i \cdot 1_{\{X_i=x_k\}}.$$

Then, if $\hat{\epsilon} = \frac{b-a}{\sqrt{2\delta}} \cdot \sqrt{\frac{K}{n}}$, then

$$\mathbb{P}(\|\hat{\mu} - \mu_P\|_{L_1(P)} \leq \hat{\epsilon}) \geq 1 - \delta.$$

$$\left(\text{since } \mathbb{E} \left[\|\hat{\mu} - \mu_P\|_{L_1(P)}^2 \right] \leq \frac{(b-a)^2}{2} \cdot \frac{K}{n} \right)$$

For the sample $(f(X_1), Y_1), \dots, (f(X_n), Y_n)$, this two results follows directly from Thm 11.10. □

Post-hoc Calibration excluding Binning

To avoid this loss of information, we can instead consider the post-hoc calibration procedure that returns an injective function \hat{h}_n .

However, the post-hoc calibration w/ injective ftn. can't provide a distribution-free validity.

Thm 12.7

Fixed n , and $f : \mathcal{X} \rightarrow [0, 1]$, let \hat{h}_n be a post-hoc calibration that satisfies a distribution-free guarantee on ECE,

$$\mathbb{P}\left(\text{ECE}(\hat{h}_n \circ f) \leq \epsilon\right) \geq 1 - \delta \quad \forall P.$$

Then, for any distribution P on $\mathcal{X} \times \{0, 1\}$ for which $f(X)$ has a nonatomic distribution,

If $\epsilon < \mathbb{E}_P[\text{Var}_P(Y | X)]$, then $\mathbb{P}_{P^n}\left(\hat{h}_n \text{ is injective}\right) \leq \delta$.

That is, these two goals are incompatible a.s.:

- ① the post-hoc calibration function \hat{h}_n is injective, and
- ② the post-hoc calibration procedure provide a distribution-free ECE guarantee.

Post-hoc Calibration excluding Binning

proof) (Again to Thm 11.11: The inference on $\mathbb{E}[Y | \hat{h}_n(f(X))]$ is impossible if $\hat{h}_n(f(X))$ is nonatomic.)

Let $(f(X), Y) \sim \tilde{P}$, and $\mu_{\tilde{P}}$ be the regression function for this distribution. Then, for any injective $h : [0, 1] \rightarrow [0, 1]$,

$$\begin{aligned}\text{ECE}(h \circ f) &= \mathbb{E}_{\tilde{P}}[|\mathbb{E}_{\tilde{P}}[Y | h(f(X))] - h(f(X))|] \\ &= \mathbb{E}_{\tilde{P}}[|\mathbb{E}_{\tilde{P}}[Y | f(X)] - h(f(X))|] \quad (\because h \text{ is injective}) \\ &= \mathbb{E}_{\tilde{P}}[|\mu_{\tilde{P}}(Z) - h(Z)|] \quad (\because \text{change of var. w/ } Z = f(X)) \\ &= \|h - \mu_{\tilde{P}}\|_{L_1(\tilde{P})}\end{aligned}$$

On the other hand, if h is not injective, then we nonetheless have $\|h - \mu_{\tilde{P}}\|_{L_1(\tilde{P})} \leq 1$, so combining both cases,

$$\|h - \mu_{\tilde{P}}\|_{L_1(\tilde{P})} \leq \text{ECE}(h \circ f) \cdot I\{h \text{ is injective}\} + I\{h \text{ is not injective}\}.$$

Now, for given ϵ in the setting, let

$$\hat{\epsilon} = \epsilon \cdot I\{\hat{h}_n \text{ is injective}\} + I\{\hat{h}_n \text{ is not injective}\},$$

then

$$\text{ECE}(\hat{h}_n \circ f) \leq \epsilon \Rightarrow \|\hat{h}_n - \mu_{\tilde{P}}\|_{L_1(\tilde{P})} \leq \hat{\epsilon},$$

and hence

$$\mathbb{P}_{P^n} \left(\|\hat{h}_n - \mu_{\tilde{P}}\|_{L_1(\tilde{P})} \leq \hat{\epsilon} \right) \geq \mathbb{P}_{P^n} \left(\text{ECE}(\hat{h}_n \circ f) \leq \epsilon \right) \geq 1 - \delta.$$

Therefore, by Thm 11.11.,

$$\mathbb{P}(\hat{\epsilon} \geq \mathbb{E}_P[\text{Var}_P(Y | X)]) \geq 1 - \delta,$$

and

$$\begin{aligned} \mathbb{P}_{P^n} \left(\hat{h}_n \text{ is injective} \right) &\leq \mathbb{P}_{P^n} (\hat{\epsilon} = \epsilon) \quad \forall \epsilon > 0 \\ &\leq \mathbb{P}_{P^n} (\hat{\epsilon} \leq \mathbb{E}_P[\text{Var}_P(Y | X)]) \\ &\leq \delta. \end{aligned}$$



Outline

1 Calibration: definition and methods

2 Properties of ECE and binned ECE

3 Properties of dCE

4 Venn-Abers Predictors

dCE v.s. ECE

Prop. 12.8: dCE is a weaker definition of calibration than ECE.

For any distribution P on $\mathcal{X} \times \{0, 1\}$ and any $f : \mathcal{X} \rightarrow [0, 1]$,

$$\text{dCE}(f) \leq \text{ECE}(f).$$

proof)

$$\text{dCE}(f) = \inf_{\substack{g: \mathcal{X} \rightarrow [0, 1] \\ \mathbb{E}[Y|g(X)] = g(X)}} \mathbb{E}[|g(X) - f(X)|] \leq \mathbb{E}[|\mathbb{E}[Y | f(X)] - f(X)|] = \text{ECE}(f)$$

□

Cf) This inequality can be extremely loose.

Prop.: dCE is continuous.

For any distribution P on $\mathcal{X} \times \{0, 1\}$ and any $f, g : \mathcal{X} \rightarrow [0, 1]$,

$$\text{dCE}(f) \leq \text{dCE}(g) + \mathbb{E}[|f(X) - g(X)|],$$

i.e., dCE is 1-Lipschitz w.r.t. $L_1(P)$ norm.

proof) It's easy by using the definition of dCE.

Definition : Estimator of dCE

Fixed $K \geq 1$,

$$\widehat{\text{dCE}}(f) := \frac{1}{n} \sum_{k=1}^K \left| \sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} \left(Y_i - \frac{k}{K} \right) \right|,$$

where the bins B_1, \dots, B_K partition of $[0, 1]$ into equal-length intervals,

$$B_1 = \left[0, \frac{1}{K} \right], \quad B_k = \left(\frac{k-1}{K}, \frac{k}{K} \right] \quad \forall k \geq 2$$

- ① dCE(f) is the distance between f and the closest perfect calibration function g from f , i.e., it means the smallest correction for calibration.
⇒ It is reasonable that g is on the discrete space.
- ② If $g(x) \equiv c_k$, $x \in B_k$, for perfect calibration, $c_k = \mathbb{E}[Y \mid f(X) \in B_k]$.
⇒ Since we don't know it, instead use the distribution-free value $c_k = \frac{k}{K}$.

dCE Estimation

Thm 12.10: dCE Estimation

Fixed n , and $f : \mathcal{X} \rightarrow [0, 1]$, let $[0, 1]$ be a fixed partition. For $\widehat{\text{dCE}}$, any distribution P on $\mathcal{X} \times \{0, 1\}$, and $\delta \in [0, 1]$,

$$\mathbb{P}\left(\widehat{\text{dCE}}(f) + \frac{1}{K} + \sqrt{\frac{2 \log(1/\delta)}{n}} \geq \text{dCE}(f)\right) \geq 1 - \delta.$$

proof)

Let $\tilde{g}(x) = \sum_k \frac{k}{K} \cdot I\{f(x) \in B_k\}$, and $g(x) = \sum_k \mu_k \cdot I\{f(x) \in B_k\}$ with $\mu_k = \mathbb{E}[Y \mid f(X) \in B_k]$, which is a perfectly calibrated function. Then,

$$n_k \left| \mu_k - \frac{k}{K} \right| \leq \mathbb{E} \left[\left| \sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} \left(Y_i - \frac{k}{K} \right) \right| \middle| n_k \right] \leq n_k \left| \mu_k - \frac{k}{K} \right| + \sqrt{n_k},$$

and hence, by marginalizing,

$$\mathbb{E}[|g(X) - \tilde{g}(X)|] \leq \mathbb{E}[\widehat{\text{dCE}}(f)] \leq \mathbb{E}[|g(X) - \tilde{g}(X)|] + \sqrt{\frac{K}{n}}$$

Moreover, by construction,

$$|\tilde{g}(x) - f(x)| \leq \max_k \sup_{t \in B_k} \left| t - \frac{k}{K} \right| \leq \frac{1}{K} \quad \forall x.$$

Hence, since g is perfectly calibrated,

$$\text{dCE}(f) \leq \mathbb{E}[|g(X) - f(X)|] \leq \mathbb{E}[|g(X) - \tilde{g}(X)|] + \mathbb{E}[|\tilde{g}(X) - f(X)|] \leq \mathbb{E}[\widehat{\text{dCE}}(f)] + \frac{1}{K},$$

and finally, for $Z_i = (X_i, Y_i)$, since

$$\sup_{z_1, \dots, z_n, z'_i} |\widehat{\text{dCE}}(z_1, \dots, z_i, \dots, z_n) - \widehat{\text{dCE}}(z_1, \dots, z'_i, \dots, z_n)| \leq \frac{2}{n},$$

by McDiarmid's inequality,

$$\mathbb{P}\left(|\widehat{\text{dCE}}(f) - \mathbb{E}[\widehat{\text{dCE}}(f)]| \geq \epsilon\right) < 2 \exp\left(-\frac{n^2 \epsilon^2}{2}\right)$$



Question session 2

Q. Why are these metrics based on L_1 distance primarily used instead of L_2 distance, which is commonly used in statistics and ML? Is it about the theoretical characteristics of the calibration or the practical difficulties?

A. We have to estimate ECE (or dCE), which is $ECE_2^2 = \mathbb{E}[|\mathbb{E}[Y | f(X)] - f(X)|^2]$. We can observe only noisy Y_i , then it's more difficult to estimate the square of bias, i.e., if we use the naïve plug-in estimator

$$\sum_{k=1}^K \left| \frac{1}{n_k} \cdot \sum_{\substack{i \in [n] \\ f(X_i) \in B_k}} (Y_i - f(X_i)) \right|^2 \cdot \left(\frac{n_k}{n} \right)^2,$$

then it would have the serious upward bias. So, we have to take the strategy

- ① for bias correction: using U-statistics, $\sum_{a \neq b} U^{(a)} U^{(b)}$,
- ② for variance correction: case distinctions (calibrated/miscalibrated).

Algorithm 1 Confidence Interval for the ℓ_2 Expected Calibration Error

Input: Calibration data set $\{(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)})\}$, model f , number k of top classes for which to check calibration, number of bins ℓ_m , significance level $\alpha \in (0, 1)$.

Output: Confidence interval C_m

Record predicted probabilities $Z^{(i)} = f(X^{(i)})$ and prediction errors $U^{(i)} = Y_{r_{1:k}}^{(i)} - Z_{(1:k)}^{(i)}$, $i \in [n]$.

For the partition $\{B_1, \dots, B_{\ell_m}\}$ from (3) from Section 2.1, define the indices of datapoints in each bin: $\mathcal{I}_{m,i} = \{j : Z_{(1:k)}^{(j)} \in B_i, 1 \leq j \leq n\}$, $i \in [\ell_m]$.

Find debiased top-1-to- k calibration error estimator $T_m = \frac{1}{n} \sum_{1 \leq i \leq \ell_m, |\mathcal{I}_{m,i}| \geq 2} \frac{1}{|\mathcal{I}_{m,i}| - 1} \sum_{a \neq b \in \mathcal{I}_{m,i}} U^{(a)\top} U^{(b)}$

Compute the variance of a calibrated model: $\sigma_0^2 = 2 \int_{\Delta(K,k)} (\|Z_{(1:k)}\|_2^2 - 2\|Z_{(1:k)}\|_3^3 + \|Z_{(1:k)}\|_2^4) dZ_{(1:k)}$,

For $i \in [\ell_m]$, define per-bin estimators of mean and covariance of the prediction error:

$$\mathbb{E}_n[U]^{(i)} = \frac{1}{|\mathcal{I}_{m,i}|} \sum_{j \in \mathcal{I}_{m,i}} U^{(j)}, \quad \text{Cov}_n[U]^{(i)} = \frac{1}{|\mathcal{I}_{m,i}|} \sum_{j \in \mathcal{I}_{m,i}} (U^{(j)\top} U^{(j)} - \mathbb{E}_n[U]^{(i)} \mathbb{E}_n[U]^{(i)\top}). \quad (6)$$

Define variance estimator for a mis-calibrated model by

$$\hat{\sigma}_1^2 = \sum_{i=1}^{\ell_m} \frac{|\mathcal{I}_{m,i}|}{n} \left\| \mathbb{E}_n[U]^{(i)} \right\|^4 - \left(\sum_{i=1}^{\ell_m} \frac{|\mathcal{I}_{m,i}|}{n} \left\| \mathbb{E}_n[U]^{(i)} \right\|^2 \right)^2 + 4 \sum_{i=1}^{\ell_m} \frac{|\mathcal{I}_{m,i}|}{n} \mathbb{E}_n[U]^{(i)\top} \text{Cov}_n[U]^{(i)} \mathbb{E}_n[U]^{(i)}. \quad (7)$$

Define positive part $T_m^+ = \max\{T_m, 0\}$ and normal quantiles $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, $z_\alpha = \Phi^{-1}(1 - \alpha)$.

Define adjusted positive confidence interval

$$C_{m,1} = \begin{cases} [T_m^+ - z_{\alpha/2} \hat{\sigma}_1 / \sqrt{n}, \quad T_m^+ + z_{\alpha/2} \hat{\sigma}_1 / \sqrt{n}], & \text{if } T_m^+ / 2 \leq T_m^+ - z_{\alpha/2} \hat{\sigma}_1 / \sqrt{n}, \\ [\max\{0, T_m^+ - z_\alpha \hat{\sigma}_1 / \sqrt{n}\}, \quad T_m^+ + z_{\alpha/2} \hat{\sigma}_1 / \sqrt{n}] \setminus \{0\}, & \text{if } T_m^+ - z_\alpha \hat{\sigma}_1 / \sqrt{n} < T_m^+ / 2, \\ [T_m^+/2, \quad T_m^+ + z_{\alpha/2} \hat{\sigma}_1 / \sqrt{n}], & \text{otherwise.} \end{cases} \quad (8)$$

Let $C_m = C_{m,1}$, and if the estimator value is small enough that $T_m^+ < z_\alpha \sigma_0 / (n \sqrt{\text{Vol}(B_1)})$, then include zero: $C_m = C_m \cup \{0\}$.

return C_m

Figure: Sun, Yan, et al. "A Confidence Interval for the ℓ_2 Expected Calibration Error." arXiv preprint arXiv:2408.08998 (2024).

Question session 3

Q. I felt that ECE and dCE are very similar to Bayes/Minimax structure, then can we apply a strategy to ECE that gives weight to the accuracy in a specific region of $f(X)$? Also, if possible, I wonder if it would be possible to show the validity of that estimator similarly.

A. If we use the absolute mean loss,

Bayes rule:

$$\delta^B := \arg \min_{\delta} \int_{\Theta} \mathbb{E}_P[|\theta - \delta(X)|] d\pi(\theta),$$

Minimax rule:

$$\delta^M := \arg \min_{\delta} \sup_{\theta \in \Theta} \mathbb{E}_P[|\theta - \delta(X)|]$$

That is, you can feel that those two are similar, but the calibrated errors are corresponding to the risk values.

Question session 4

Q. The book explains that in continuous situations, estimation that is impossible with ECE is possible with dCE. What are the pros and cons of choosing dCE over ECE, and which indicator is more commonly used in practice?

A.

pros: Above.

cons: Conservative. (By prop. 12.8)

In practice, especially in ML of the classifier, you can often see the following notation:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} \cdot |\text{acc}(B_m) - \text{conf}(B_m)|,$$

where

$$\text{acc}(B_m) = \frac{1}{|B_m|} \cdot \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i),$$

and

$$\text{conf}(B_m) = \frac{1}{|B_m|} \cdot \sum_{i \in B_m} p_i.$$

Outline

- 1 Calibration: definition and methods
- 2 Properties of ECE and binned ECE
- 3 Properties of dCE
- 4 Venn-Abers Predictors

- ① We want to estimate

$$\mathbb{P}[Y = 1 \mid f(X)]$$

to certify the perfect calibration, using post-hoc calibration.

- ② For the miscalibration measure, ECE (and binECE), we found out that it can't estimate precisely by using the finite-sampled empirical estimator.

⇒ We instead target

$$\mathbb{P}[Y = 1 \mid (\hat{h} \circ f)(X)].$$

- ③ Estimate the upper bound of those measures w.r.t. the calibrated probability, so that we can indirectly check whether this model is perfectly calibrated.

⇒ But this method is proper only using binning or dCE.

- ④ On the other hand, we can take the other strategy to interval-estimate the above probability.

⇒ Venn-Abers Prediction.

- By exchangeability, we can treat the universe of $(X_{n+1}, 0)$, same as that of $(X_{n+1}, 1)$.
- Model two isotonic regression for the calibrated probability, respectively: $\hat{p}_{n+1}^0, \hat{p}_{n+1}^1$.
- We have confidence in that the true calibrated probability will be in $[\hat{p}_{n+1}^0, \hat{p}_{n+1}^1]$.

Algorithm 12.11: Venn-Abers Prediction

- ① Input: Augmented calibration data with binary label

$\mathcal{D}_{n+1}^y = \{(X_1, Y_1), \dots, (X_{n+1}, y)\}$ for $y \in \{0, 1\}$, and pretrained model $f : \mathcal{X} \rightarrow [0, 1]$.

- ② Perform isotonic regressions, respectively,

$$\hat{h}^y = \arg \min_{h \in \mathcal{H}_{\text{iso}}} \left[\sum_{i=1}^n ([h \circ f](X_i) - Y_i)^2 + ([h \circ f](X_{n+1}) - y)^2 \right],$$

and return the fitted values

$$\hat{p}^y = ([\hat{h}^y \circ f](X_1), \dots, [\hat{h}^y \circ f](X_{n+1}))$$

- ③ Output: The interval $[\hat{p}_{n+1}^0, \hat{p}_{n+1}^1]$.

Distribution-free guarantee of Venn-Abers prediction

Thm 12.12: Distribution-free Guarantee for Perfect Calibration of Venn-Abers Predictor

Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{X} \times \{0, 1\}$ be exchangeable, and $f : \mathcal{X} \rightarrow [0, 1]$ be a pre-trained model. For \hat{p}_{n+1}^0 , and \hat{p}_{n+1}^1 , which is got by the Algorithm 12.11,

$$\mathbb{P}\left(Y_{n+1} = 1 \mid \hat{p}_{n+1}^{Y_{n+1}}\right) = \hat{p}_{n+1}^{Y_{n+1}},$$

i.e., \exists r.v. $W \in [\hat{p}_{n+1}^0, \hat{p}_{n+1}^1]$ that is perfectly calibrated. ($\mathbb{E}[Y_{n+1} \mid W] = W$)

proof)

Let $\hat{h} = \hat{h}^{Y_{n+1}}$. By definition of the isotonic regression problem, the data points

$$\left(([\hat{h} \circ f](X_1), Y_1), \dots, ([\hat{h} \circ f](X_{n+1}), Y_{n+1})\right)$$

are exchangeable.

Without proof, denote that the fitted values of isotonic regression are piecewise constant, and the value within each bin is equal to the sample mean within the bin, i.e.,

for some partition $[n + 1] = I_1 \cup \dots \cup I_M$, $[\hat{h} \circ f](X_i) = \bar{Y}_{I_m} \forall m \in [M], i \in I_m$.

Using this, for any function $g : [0, 1] \rightarrow [0, 1]$,

$$\begin{aligned}\mathbb{E} \left[\left(Y_{n+1} - \hat{p}_{n+1}^{Y_{n+1}} \right) \cdot g(\hat{p}_{n+1}^{Y_{n+1}}) \right] &= \mathbb{E} \left[\left(Y_{n+1} - [\hat{h} \circ f](X_{n+1}) \right) \cdot g([\hat{h} \circ f](X_{n+1})) \right] \\ &= \frac{1}{n+1} \mathbb{E} \left[\sum_{i=1}^{n+1} \left(Y_i - [\hat{h} \circ f](X_i) \right) \cdot g([\hat{h} \circ f](X_i)) \right] \\ &\quad (\because \text{exchangeability}) \\ &= \frac{1}{n+1} \mathbb{E} \left[\sum_{m=1}^M \sum_{i \in I_m} \left(Y_i - [\hat{h} \circ f](X_i) \right) \cdot g([\hat{h} \circ f](X_i)) \right] \\ &= \frac{1}{n+1} \mathbb{E} \left[\sum_{m=1}^M \sum_{i \in I_m} \left(Y_i - \bar{Y}_{I_m} \right) \cdot g(\bar{Y}_{I_m}) \right] \\ &= \frac{1}{n+1} \mathbb{E} \left[\sum_{m=1}^M 0 \cdot g(\bar{Y}_{I_m}) \right] = 0,\end{aligned}$$

and hence

$$\mathbb{E}[Y_{n+1} - \hat{p}_{n+1}^{Y_{n+1}} \mid \hat{p}_{n+1}^{Y_{n+1}}] = 0 \quad \text{a.s.}$$



Question session 5

Q. In the book, "If the interval is wide, it signifies high uncertainty in our probabilistic prediction." What are the width criteria? Are they different by the domain?

A. We don't need the width criteria, and to consider the domain, since our target is just a probability(scalar).

To help you understand, let's see the following example:

By the isotonic regression for test, assume that

$$\hat{p}_{n+1}^0 = 0.3, \hat{p}_{n+1}^1 = 0.7.$$

Consider the random variable $W = \begin{cases} 0.3, & \text{w.p. } \frac{1}{2} \\ 0.7, & \text{w.p. } \frac{1}{2}, \end{cases}$ then $W \in [0.3, 0.7]$, and

if $Y_{n+1}|W \sim \text{Bern}(W)$, then W is perfectly calibrated.

But in the other model, if $\hat{p}_{n+1}^0 = 0.1, \hat{p}_{n+1}^1 = 0.9$, the r.v. of estimated probability W take values in $[0.1, 0.9]$, more widely.