Comparing Calibration Error Metrics

1 Introduction

Calibration measures whether predicted probabilities match observed frequencies. For a predictor $f : \mathcal{X} \to [0,1]$ and distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, we say f is *calibrated* if for all $v \in [0,1]$:

$$\mathbb{E}[Y \mid f(X) = v] = v$$

When this doesn't hold, we measure calibration error in different ways. This note compares three common metrics.

2 Three Calibration Error Metrics

Let $(X,Y) \sim \mathcal{D}$ and $f(X) = V$ be the predicted probability. The calibration error at prediction value v is:

$$\mathrm{CE}(v) = \mathbb{E}[Y \mid V = v] - v$$

2.1 Expected Calibration Error (ECE)

Definition 1 (Expected Calibration Error)**.**

$$ECE(f) = \sum_{v:\Pr[V=v]>0} \Pr[V = v] \cdot |\mathbb{E}[Y \mid V = v] - v|$$

Interpretation: Average absolute calibration error, weighted by how often each prediction value occurs.

2.2 Squared Calibration Error (SCE)

Definition 2 (Squared Calibration Error)**.**

$$SCE(f) = \sum_{v:\Pr[V=v]>0} \Pr[V = v] \cdot (\mathbb{E}[Y \mid V = v] - v)^2$$

Interpretation: Mean squared calibration error, weighted by prediction frequency.

2.3 Maximum Calibration Error (MCE)

Definition 3 (Maximum Calibration Error)**.**

$$MCE(f) = \max_{v:\Pr[V=v]>0} \Pr[V = v] \cdot |CE(v)| = \max_{v:\Pr[V=v]>0} \Pr[V = v] \cdot |\mathbb{E}[Y \mid V = v] - v|$$

Interpretation: Maximum probability-weighted calibration error over all prediction values.

Design Choice: We weight the calibration error at each prediction value v by its probability $\Pr[V = v]$. This is a choice—an alternative would be the unweighted maximum $\max_v |\mathrm{CE}(v)|$, which treats all predictions equally regardless of frequency. The weighted version balances worst-case concerns with the practical importance of each prediction.

3 Relationships Between Metrics

3.1 Comparing Average vs. Squared (ECE vs. SCE)

Proposition 1 (ECE vs. SCE)**.** *For any predictor f:*

$$SCE(f) \leq ECE(f) \leq \sqrt{SCE(f)}$$

The first inequality holds when all calibration errors satisfy $|CE(v)| \leq 1$. The second holds always.

Proof. **First inequality:** If $|\mathrm{CE}(v)| \leq 1$ for all v, then $|\mathrm{CE}(v)|^2 \leq |\mathrm{CE}(v)|$, so:

$$\mathrm{SCE}(f) = \sum_v \Pr[V = v] \cdot |\mathrm{CE}(v)|^2 \leq \sum_v \Pr[V = v] \cdot |\mathrm{CE}(v)| = \mathrm{ECE}(f)$$

Second inequality: By Jensen's inequality (since $x \mapsto x^2$ is convex):

$$\mathrm{ECE}(f)^2 = \left(\sum_v \Pr[V = v] \cdot |\mathrm{CE}(v)| \right)^2 \leq \sum_v \Pr[V = v] \cdot |\mathrm{CE}(v)|^2 = \mathrm{SCE}(f)$$

Taking square roots: $\mathrm{ECE}(f) \leq \sqrt{\mathrm{SCE}(f)}$. □

3.2 Comparing Average vs. Maximum (ECE vs. MCE)

Proposition 2 (ECE vs. MCE)**.** *For any predictor f, let $m = |\{v : \Pr[V = v] > 0\}|$ be the number of distinct prediction values. Then:*

$$MCE(f) \leq ECE(f) \leq m \cdot MCE(f)$$

Proof. **First inequality:** MCE is the maximum of the same terms that ECE sums. Since all terms are non-negative:

$$\mathrm{MCE}(f) = \max_v \Pr[V = v] \cdot |\mathrm{CE}(v)| \leq \sum_v \Pr[V = v] \cdot |\mathrm{CE}(v)| = \mathrm{ECE}(f)$$

The sum of non-negative terms is at least as large as the maximum term.

Second inequality: ECE is a sum of m terms, each of which is at most MCE:

$$\mathrm{ECE}(f) = \sum_v \Pr[V = v] \cdot |\mathrm{CE}(v)| \leq \sum_v \mathrm{MCE}(f) = m \cdot \mathrm{MCE}(f)$$

□