CS7180/7880 Lecture 2: Calibration Measures

Batch:
1. Feature domain X
2. Ground-truth distribution D on data points $(x, y) \in X \times \{0,1\}$
3. Predictor $f: X \to [0,1]$

Def. 1. (Perfect calibration). We say f is perfectly calibrated if for every value $p \in [0,1]$ in the range of f,

$$\mathbb{E}[y \mid f(x) = p] = p.$$

Sequential.

For round $t = 1, \cdots, T$

1. Nature reveals $x_t \in X$
2. Forecaster predicts $p_t \in [0,1]$
3. Nature reveals true outcome $y_t \in \{0,1\}$.

Def 2. (Perfect Cal. for sequential predictions). We say $p_1, \dots, p_T$ are calibrated on $y_1, \dots, y_T$ if for every $p \in [0,1]$,

$$\sum_{t=1}^{T} y_t \mathbb{1}[p_t = p] = p \cdot \sum_{t=1}^{T} \mathbb{1}[p_t = p]$$

Calibration alone is easy to achieve by "trivial" predictions

Batch: $f(x) = \mathbb{E}[y]$.

Sequential: [FV98] Can achieve calibration error ECE $O(T^{-1/3}) \to 0$ as $T \to +\infty$ without observing x_t.

only by randomized algo

Calibration Measures.

In reality, perfect calibration is rare.

Goal: assess level of miscalibration.

Def.3 (ECE). Given predictor f, distribution D.

Define $\bar{y}_p : \mathbb{E}_D[y \mid f(x) = p]$ for every $p \in [0,1]$.

$$ECE(f, D) := \mathbb{E}_{p=f(x),\ \text{for } (x,y)\sim D}\left[\,|\bar{y}_p - p|\,\right].$$

Example: D is uniform on $\{(x_0, 0), (x_1, 1)\}$

$f_\alpha(x_0) = \alpha$, $\quad f_\alpha(x_1) = 1-\alpha$, for $\alpha \in [0,1]$

$$ECE(f_\alpha, D) = \begin{cases} \alpha, & \text{if } \alpha \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1] \\ 0, & \text{if } \alpha = \frac{1}{2} \end{cases}$$

$p = \alpha$ or $1-\alpha$

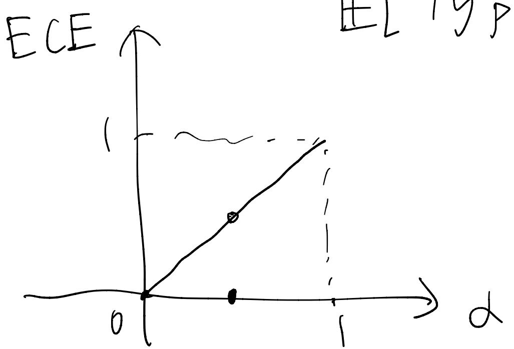
Case 1: $\alpha \neq 1-\alpha$. $\quad \bar{y}_\alpha = 0 \quad \bar{y}_{1-\alpha} = 1$

$$E[|\bar{y}_p - p|] = \frac{1}{2} \cdot |\bar{y}_\alpha - \alpha| + \frac{1}{2}|\bar{y}_{1-\alpha} - (1-\alpha)|$$

$$= \alpha.$$

Case 2: $\underbrace{\alpha = 1-\alpha}_{\alpha = \frac{1}{2}}$ $\quad \bar{y}_\alpha = \bar{y}_{1-\alpha} = \frac{1}{2}.$

$$E[|\bar{y}_p - p|] = |\tfrac{1}{2} - \tfrac{1}{2}| = 0$$



★ ECE in general cannot be estimated on a finite data set.

Reason: Need to see ≥ 2 examples x, x' with same pred.

$f(x) = f(x') = p$.

$\Pr[\text{collision}] = 0$

Binned ECE: Partition $[0,1]$ into k bins $B_1, \cdots, B_k$

Def 4 (Binned ECE). For $i = 1, \cdots, k$,

$$\bar{y}_i = \mathbb{E}[y \mid f(x) \in B_i], \quad \bar{p}_i = \mathbb{E}[f(x) \mid f(x) \in B_i].$$

$$\text{Binned ECE}(f, D) = \sum_{i=1}^{k} \Pr[f(x) \in B_i] \cdot |\bar{y}_i - \bar{p}_i|.$$

Distance to Calibration.

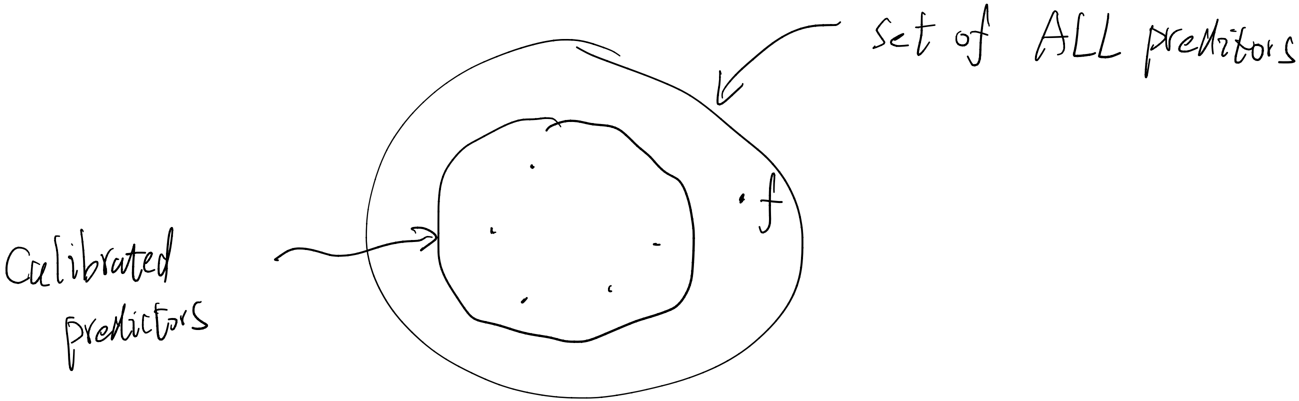
① Completeness: If f is perfectly calibrated on D

then $CAL(f, D) = 0$.

② Soundness: If f is mis-calibrated on D

then $CAL(f, D) > 0$.

③ Robustness: Completeness and soundness should be preserved under small perturbations to predictions.



Distance to Calibration.

$$\mathrm{DTC}(f, D) = \inf_{f'} \mathbb{E}[\,|f(x) - f'(x)|\,]$$

where inf is over all calibrated predictors f' on D.

Example. Same as before. D is uniform on $\{(x_0, 0), (x_1, 1)\}$.

$$f_\alpha(x_0) = \alpha, \qquad f_\alpha(x_1) = 1 - \alpha.$$

$$\mathrm{DTC}(f_\alpha, D) = \begin{cases} \alpha, & \text{if } \alpha \in [0, \frac{1}{4}] \\ |\alpha - \frac{1}{2}|, & \text{if } \alpha \in [\frac{1}{4}, 1]. \end{cases}$$

There are only 2 calibrated predictors

$$f_1(x_0) = 0, \quad f_1(x_1) = 1.$$

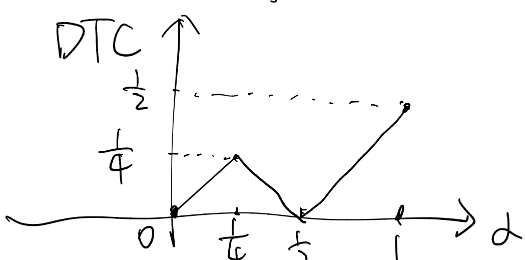
$$f_2(x_0) = f_2(x_1) = \frac{1}{2}.$$

When $\alpha \le \frac{1}{4}$, f_1 is closer. $\mathrm{DTC} = \mathbb{E}\,|f_\alpha(x) - f_1(x)|$

$= \alpha$

When $\alpha \ge \frac{1}{4}$, f_2 is closer. $\mathrm{DTC} = \mathbb{E}\,|f_\alpha(x) - f_2(x)|$

$= |\alpha - \frac{1}{2}|$



Def (Consistency) We say CAL is Consistent if there are constants $c_1, c_2, \alpha_1, \alpha_2 > 0$, s.t.

$$c_1 \cdot DTC(f,D)^{\alpha_1} \le CAL(f,D) \le c_2 \cdot DTC(f,D)^{\alpha_2}$$

Can we achieve
1. Continuous
2. Consistent
3. "Easy" to compute / estimate

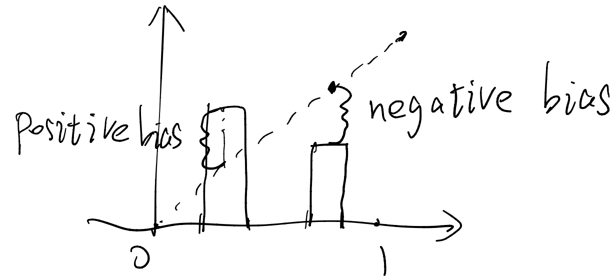
Def (Smooth Calibration Error). Define W as the set of 1-Lipschitz func. $w: [0,1] \to [-1,1]$.

weight func.

$w(v_1) - w(v_2) \le |v_1 - v_2|, \forall v_1, v_2$

$\in [0,1]$

$$\mathrm{SmCE}(f, D) = \sup_{w \in W} \mathbb{E}\left[(f(x) - y) \cdot w(f(x)) \right].$$

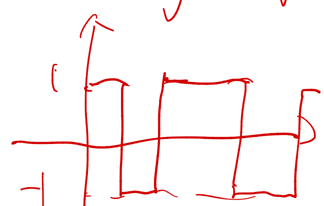


$$|f(x) - y| = \sup_{\sigma \in \{-1, 1\}} (f(x) - y) \cdot \sigma$$

abs. of bias

"weight"
"auditor"
"distinguisher"
"test"

For Binned ECE: weight func:



For SmCE

Thm: $\frac{1}{32} \cdot DTC^2 \leq SmCE \leq 2 \cdot DTC$

Desired properties.

① Completeness

② Soundness

③ Consistency

④ Continuity

⑤ Testability (Can estimate using a small data set)

6. Efficiency (Efficient to compute)
7. Actionability (Next lecture)
8. Truthfulness (Incentivize true/honest predictions)