CSE493S — Homework 2: Calibration

Marginal Mean/Quantile Consistency and Basic Theoretical Calibration

Due: Tue Nov 25, 2025 at 11:59pm **Name:** _____

Q1. Murphy's Brier Decomposition (10 pts). Suppose there are n examples $(x_1, y_1), \ldots, (x_n, y_n)$ with binary labels $y_i \in \{0, 1\}$ and probability forecasts $p_1, \ldots, p_n \in [0, 1]$. Partition the index set $\{1, \ldots, n\}$ arbitrarily into disjoint groups $S_1, \ldots, S_B$ whose union is $\{1, \ldots, n\}$. For each bin b, define $n_b = |S_b|$, $w_b = n_b/n$, the empirical event rate $r_b = \frac{1}{n_b} \sum_{i \in S_b} y_i$, and the average forecast $c_b = \frac{1}{n_b} \sum_{i \in S_b} p_i$. Let the overall event rate be $r = \frac{1}{n} \sum_{i=1}^{n} y_i$. Show the empirical identity

$$\text{Brier score} = \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i)^2 \;=\; \underbrace{r(1-r)}_{\text{uncertainty}} \;-\; \underbrace{\sum_{b=1}^{B} w_b (r_b - r)^2}_{\text{resolution}} \;+\; \underbrace{\sum_{b=1}^{B} w_b (c_b - r_b)^2}_{\text{reliability}} .$$

Provide a short derivation (e.g., add-and-subtract r_b, expand, and average within bins). Conclude that replacing each p_i by r_b for $i \in S_b$ (bin-wise recalibration) reduces the Brier score by exactly the reliability term and leaves uncertainty and resolution unchanged.

Q2. The Weather Factory (10 pts). Suppose nature draws a latent rain probability $B \sim D$ for some fixed but unknown distribution on $[0, 1]$. Then, conditioned on B, $Y \mid B \sim \text{Bernoulli}(B)$. You must forecast a single constant q (without assuming any parametric form for D). Prove $q^* = \mathbb{E}[B]$ (assume $\mathbb{E}[B]$ exists). Express the optimal squared-loss risk $\min_q \mathbb{E}[(Y - q)^2]$ in terms of $\text{Var}(Y)$ and $\text{Var}(B)$. For clarity, define *irreducible noise* as $\mathbb{E}[\, B(1-B)\,]$ and *heterogeneity* as $\text{Var}(B)$. Using the law of total variance, $\text{Var}(Y) = \mathbb{E}[\, B(1-B)\,] + \text{Var}(B)$; in particular, the optimal constant risk equals $\text{Var}(Y)$. Briefly explain the roles of irreducible noise and heterogeneity.

Q3. A/B Testing: One Constant Score (10 pts). A recommendation engine routes each user to arm A with probability α and to arm B with probability $1-\alpha$ (A/B testing). Let the binary outcome be Y (e.g., click/purchase). Under arm A the event rate is π_A; under arm B it is π_B. If forced to output a single constant score q for all users, show $q^* = \alpha\, \pi_A + (1 - \alpha)\, \pi_B$. Compute the Brier score increase versus arm-specific optimal constant scores (π_A on A and π_B on B) and show it equals $\alpha(1 - \alpha)(\pi_A - \pi_B)^2$. Provide one–two sentences of interpretation.

Q4. Label Shift: Should We Invert? (10 pts). Consider regression for a binary outcome task, where the prevalence of label 1 is π_{train} (and label 0 has $1-\pi_{\text{train}}$). At deployment, the marginal prevalence shifts to $\pi_{\text{deploy}} = 1 - \pi_{\text{train}}$ (label shift), but the model still outputs probabilities p learned under π_{train}.

- Derive the deployment Brier scores for using p and for using $1 - p$. Show that their difference satisfies
$$\Delta = \mathbb{E}\big[(Y - (1-p))^2\big] - \mathbb{E}\big[(Y-p)^2\big] = 1 - 2\,\mathbb{E}[Y] - 2\,\mathbb{E}[p] + 4\,\mathbb{E}[pY]\,,$$
where expectations are under the deployment distribution.
- Give conditions under which flipping to $1 - p$ strictly *improves* the Brier score after the shift (i.e., $\Delta < 0$). Identify any degenerate cases where neither choice changes the Brier score (e.g., constant $p \equiv 0.5$).

Q5. The Tie Zone (8 pts). For a discrete distribution with atoms near the median, characterize the full set of 0.5-quantiles and explain why the pinball risk is flat on that interval. Provide a concrete example (e.g., a five-point support) and identify all medians.

Q6. Mean vs. Median Target (8 pts). Switch to absolute loss $L(m) = \mathbb{E}[|Y - m|]$. Show that any minimizer is a median (0.5-quantile), contrasting with squared loss targeting the mean. Give one operational implication when choosing loss.

Q7. Calibration vs. Sharpness Trade-off (12 pts). You are given per-model bin summaries (each model's bins are formed by its own predicted probabilities). For both models, the overall event rate is the same: $\bar{r} = 0.5$.

Model A bins

$$w^A = (0.30,\ 0.40,\ 0.30), \quad r^A = (0.20,\ 0.50,\ 0.80), \quad c^A = (0.18,\ 0.50,\ 0.82).$$

Model B bins

$$w^B = (0.25,\ 0.50,\ 0.25), \quad r^B = (0.10,\ 0.50,\ 0.90), \quad c^B = (0.15,\ 0.50,\ 0.85).$$

(a) Verify that $\sum_b w_b^A r_b^A = \sum_b w_b^B r_b^B = \bar{r} = 0.5$. Compute the uncertainty term $\bar{r}(1 - \bar{r})$.

(b) Using Murphy's decomposition with each model's own bins, compute the resolution for each model: $\sum_b w_b (r_b - \bar{r})^2$. Which model is sharper (higher resolution)?

(c) Compute the reliability for each model: $\sum_b w_b (c_b - r_b)^2$. Which model is better calibrated (lower reliability)?

(d) Compute each model's Brier score via Brier $= \bar{r}(1 - \bar{r}) - \text{resolution} + \text{reliability}$. Which model has the lower Brier score on this dataset?

(e) In 3–5 sentences, explain how one model can be *more calibrated but less sharp*, while the other can be *less calibrated but sharper*. Conclude why calibration does not directly determine sharpness (and vice versa).