Math Basics for Binary Medical Image Classification

1 Data and Model

We have input images $x \in \mathcal{X}$ with binary labels $y \in \{0, 1\}$ (e.g., `pneumonia` vs. `normal`). A neural network with parameters θ maps x to two *logits*:

$$\boldsymbol{z}(x;\theta) = \big(z_0(x;\theta),\, z_1(x;\theta)\big) \in \mathbb{R}^2.$$

We convert logits to class probabilities using the *softmax*:

$$p_\theta(y = k \mid x) \;=\; \frac{e^{z_k(x;\theta)}}{e^{z_0(x;\theta)} + e^{z_1(x;\theta)}} \quad (k = 0, 1).$$

The prediction is $\hat{y} = \arg\max_k p_\theta(y = k \mid x)$, and the model's confidence is $\max_k p_\theta(y = k \mid x)$.

2 Training Objective: Cross-Entropy

Given a training set $\{(x_i, y_i)\}_{i=1}^n$, we minimize the average *cross-entropy (CE)* loss:

$$\mathcal{L}_{\mathrm{CE}}(\theta) \;=\; -\frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}\{y_i = 1\}\,\log p_\theta(1 \mid x_i) \;+\; \mathbf{1}\{y_i = 0\}\,\log p_\theta(0 \mid x_i)\Big].$$

This encourages the model to assign high probability to the true class. We optimize θ by gradient-based methods (e.g., AdamW).

3 Evaluation: Accuracy

On a test set $\{(x_j, y_j)\}_{j=1}^m$, the *accuracy* is

$$\mathrm{Acc} \;=\; \frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\{\hat{y}_j = y_j\}.$$

Accuracy checks "how often we are correct" under a fixed decision threshold (argmax).

4 Evaluation: AUROC (Discrimination)

For binary tasks, let $s(x) = p_\theta(1 \mid x)$ be the positive-class score. The *ROC* curve varies a threshold t and plots True Positive Rate vs. False Positive Rate. The *Area Under the ROC* (*AUROC*) can be understood as

$$\mathrm{AUROC} \;=\; \Pr\big[s(x^+) > s(x^-)\big],$$

the probability that a randomly chosen positive example x^+ scores higher than a randomly chosen negative example x^-. AUROC measures *ranking quality* regardless of any fixed threshold.

5 Calibration and ECE

A model is *well-calibrated* if its predicted confidence matches the empirical accuracy. Let $c_j = \max_k p_\theta(y = k \mid x_j)$ be the confidence of prediction for x_j. Divide $[0, 1]$ into B bins, e.g. $[0, \frac{1}{B}), [\frac{1}{B}, \frac{2}{B}), \ldots$. For bin b, define

$$\mathrm{conf}(b) = \frac{1}{|b|} \sum_{j \in b} c_j, \qquad \mathrm{acc}(b) = \frac{1}{|b|} \sum_{j \in b} \mathbf{1}\{\hat{y}_j = y_j\},$$

where $|b|$ is the number of test points whose confidences fall in bin b. The *Expected Calibration Error (ECE)* is the weighted average of the absolute gap:

$$\mathrm{ECE} \; = \; \sum_{b=1}^{B} \frac{|b|}{m} \left| \mathrm{acc}(b) - \mathrm{conf}(b) \right|.$$

Small ECE means "when the model says 0.8, it is correct about 80% of the time".

Reliability Diagram. Plot $\mathrm{acc}(b)$ (vertical) versus $\mathrm{conf}(b)$ (horizontal) with the diagonal line $y = x$. Points close to the diagonal indicate better calibration.

6 Post-hoc Temperature Scaling

To improve calibration without changing the classifier's ranking, we adjust logits by a single *temperature* $T > 0$ (learned on a validation set):

$$\tilde{\boldsymbol{z}}(x; \theta, T) \; = \; \frac{\boldsymbol{z}(x; \theta)}{T}, \qquad \tilde{p}_\theta(y = k \mid x; T) \; = \; \frac{e^{\tilde{z}_k(x;\theta,T)}}{\sum_\ell e^{\tilde{z}_\ell(x;\theta,T)}}.$$

We choose T to minimize the validation negative log-likelihood (NLL):

$$T^\star \; = \; \arg\min_{T>0} \; - \sum_{(x,y) \in \mathrm{Val}} \log \tilde{p}_\theta(y \mid x; T).$$

Applying $T^\star$ typically reduces ECE (better-calibrated probabilities) while leaving AUROC almost unchanged (ranking preserved).

7 Generalization: Train/Validation/Test

To estimate performance fairly:

- **Train** set fits parameters θ by minimizing $\mathcal{L}_{\mathrm{CE}}$.
- **Validation** set tunes choices (e.g., early stopping, temperature T).
- **Test** set is used once for the final report (no tuning).

Good models have high AUROC/accuracy on validation and test, and low ECE (honest probabilities).