Part 6:

Setup

From Part 5, you already have an optimal threshold t^* (chosen on the **test** set).

Now, in Part 6, you:

- Fix the classifier and the threshold t^*.
- Evaluate its performance on the **validation** data.
- Compute:
  1. The **empirical cost** on the validation set.
  2. A **99% confidence interval** for the true expected cost using **Hoeffding's inequality**.

Step 1: Define the cost on one observation

Let (X_i, Y_i), $i = 1, \ldots, n$, be the validation data, where

- $Y_i \in \{0, 1\}$ is the true label,
- X_i is the tweet.

The model gives a probability

$$p_i = P(Y = 1 \mid X_i) \approx \text{model.predict\_proba}(X_i) \text{ (second column)}.$$

Using the fixed threshold t^*, the prediction is

$$\hat{Y}_i = \begin{cases} 1, & \text{if } p_i \geq t^*, \\ 0, & \text{if } p_i < t^*. \end{cases}$$

The cost for observation i is

- 1 if it is a **false negative**: $Y_i = 1$ and $\hat{Y}_i = 0$,
- 5 if it is a **false positive**: $Y_i = 0$ and $\hat{Y}_i = 1$,
- 0 if the prediction is correct.

Define a random variable Z_i as the cost of the i-th observation:

$$Z_i = \begin{cases} 1, & \text{if } Y_i = 1 \text{ and } \hat{Y}_i = 0, \\ 5, & \text{if } Y_i = 0 \text{ and } \hat{Y}_i = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, each Z_i is **bounded**:

$$0 \leq Z_i \leq 5.$$

Step 2: Empirical average cost on the validation set

The **empirical (observed) average cost** on the validation set is

$$\widehat{C}_{\text{valid}} = \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

This is what you conceptually compute when you call `cost(model, t^*, X_valid, Y_valid)` in code.

Interpretation:

- $\widehat{C}_{\text{valid}}$ is your estimator of the **true expected cost** C of this classifier (with fixed threshold t^*) on new data from the same distribution.

Step 3: Assumptions for Hoeffding's inequality

We now want a **confidence interval** for the unknown true expected cost

$$C = \mathbb{E}[Z_i].$$

We assume:

1. The pairs (X_i, Y_i) (and thus the Z_i) are i.i.d. samples.
2. The random variables Z_i are **bounded**, which we already know:

$$a = 0 \leq Z_i \leq 5 = b.$$

So the conditions for Hoeffding's inequality are satisfied.

Step 4: State Hoeffding's inequality

Hoeffding's inequality says that for i.i.d. random variables $Z_1, \ldots, Z_n$ with $a \leq Z_i \leq b$,

$$\mathbb{P}\left(\left|\widehat{C}_{\text{valid}} - C\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right).$$

Here,

- $\widehat{C}_{\text{valid}} = \frac{1}{n}\sum Z_i$,
- $C = \mathbb{E}[Z_i]$,
- $a = 0, b = 5$.

We want a **99% confidence interval**, i.e.

$$\mathbb{P}\left(\left|\widehat{C}_{\text{valid}} - C\right| \leq \varepsilon\right) \geq 0.99.$$

Equivalently, we want

$$2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \leq 0.01.$$

Let $\delta = 0.01$ (so confidence $1 - \delta = 0.99$).

Step 5: Solve for ε

From Hoeffding's bound:

$$2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) = \delta.$$

Take natural logarithm:

$$-\frac{2n\varepsilon^2}{(b-a)^2} = \ln\left(\frac{\delta}{2}\right).$$

Multiply by -1:

$$\frac{2n\varepsilon^2}{(b-a)^2} = -\ln\left(\frac{\delta}{2}\right) = \ln\left(\frac{2}{\delta}\right).$$

Solve for ε^2:

$$\varepsilon^2 = \frac{(b-a)^2}{2n}\ln\left(\frac{2}{\delta}\right).$$

So

$$\varepsilon = (b-a)\sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)}.$$

Plug in $a = 0, b = 5, \delta = 0.01$:

$$\varepsilon = 5\sqrt{\frac{1}{2n}\ln\left(\frac{2}{0.01}\right)} = 5\sqrt{\frac{1}{2n}\ln(200)}.$$

That is the **radius** of the confidence interval.

Step 6: Final 99% confidence interval

We now have:

- Point estimate: $\widehat{C}_{\text{valid}}$ (empirical cost on validation set).
- Error bound: $\varepsilon = 5\sqrt{\dfrac{\ln(200)}{2n}}$.

So the **99% Hoeffding confidence interval** for the true expected cost C is

$$\left(\widehat{C}_{\text{valid}} - \varepsilon,\ \widehat{C}_{\text{valid}} + \varepsilon\right),$$

i.e.

$$\left(\widehat{C}_{\text{valid}} - 5\sqrt{\frac{\ln(200)}{2n}},\ \widehat{C}_{\text{valid}} + 5\sqrt{\frac{\ln(200)}{2n}}\right).$$

This is exactly what you'd write in an exam-style, hand-written solution:

1. Define Z_i and $\widehat{C}_{\text{valid}}$.
2. Note bounds $0 \le Z_i \le 5$.
3. State Hoeffding's inequality.
4. Choose $\delta = 0.01$, solve for ε.
5. Give the final interval formula around $\widehat{C}_{\text{valid}}$.