AI Olympiad, February 2026

1. **Notation**

   Let us consider a multi-label classification problem. For each class c, TP_c represents the instances that are in class c correctly identified as belonging to c; TN_c represents the instances that are not in class c correctly identified as not belonging to c; FP_c represents the instances that are in not class c incorrectly labeled as belonging to c; FN_c represents the instances that are in class c incorrectly labeled as not belonging to c.

   We have the following metrics:

   - per-class **recall**:

   $$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} = P(\text{pred} = c \mid \text{true} = c);$$

   - per-class **precision**:

   $$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} = P(\text{true} = c \mid \text{pred} = c);$$

   - per-class **specificity**:

   $$\text{Spec}_c = \frac{TN_c}{TN_c + FP_c} = P(\text{pred} \neq c \mid \text{true} \neq c);$$

   - for each class-dependent error metric M_c, the **macro average error metric** is the arithmetic mean of the class specific scores

   $$\text{Macro} - M = \frac{\sum_c M_c}{\text{number of classes}};$$

   - for a distribution $\pi = (\pi_c)$ of the labels, the **accuracy**

   $$\text{acc}(\pi) = \sum_c \pi_c \text{Recall}_c;$$

   - if $C = (C_{ij})$ is a cost matrix where $C_{ij} = \text{Cost}(\text{pred} = j \mid \text{true} = i)$, the **expected cost** is

   $$\mathbb{E}_\pi(C) = \sum_{i,j} \pi_i P(\text{pred} = j \mid \text{true} = i) C_{ij}.$$

You can assume that every class has nonzero support and the classifier makes both correct and incorrect predictions for every class (non-degeneracy).

a) Let us assume that you are in a binary classifier, with classes 0 and 1. Prove that, if

$$\text{accuracy} = \text{precision}_1 = \text{recall}_1,$$

the support of the two classes must be the same.

[5 Marks]

For the remaining parts, you can assume the following context.

Context. A public-health triage system classifies incoming messages into one of three classes:

$$U = \text{Urgent}, \quad S = \text{Semi-urgent}, \quad N = \text{Non-urgent}.$$

The training dataset was intentionally enriched with urgent and semi-urgent cases. On the *training set*, the classifier produced the following confusion matrix (rows correspond to the true label, columns to the predicted label):

True \ Pred	U	S	N	Total
U	200	30	20	250
S	30	270	50	350
N	10	30	360	400

The corresponding training label distribution is

$$\pi_{\text{train}} = \big(P(U) = 0.25,\ P(S) = 0.35,\ P(N) = 0.40\big).$$

At deployment time, the real-world label distribution is different:

$$\pi_{\text{dep}} = \big(P(U) = 0.05,\ P(S) = 0.15,\ P(N) = 0.80\big).$$

You may assume that the classifier's conditional behavior is **stable**:

$$P_{\text{train}}(\text{pred} = j \mid \text{true} = i) = P_{\text{dep}}(\text{pred} = j \mid \text{true} = i) \quad \text{for all } i, j \in \{U, S, N\}.$$

We are given the following cost matrix (cost of predicting the column label when the true label is the row label):

$$C(\text{pred} = j \mid \text{true} = i) = \begin{array}{c|ccc} & U & S & N \\ \hline U & 0 & 20 & 100 \\ S & 5 & 0 & 10 \\ N & 1 & 1 & 0 \end{array}$$

b) Compute Precision_U under both π_{train} and π_{dep}. Explain why precision changes under distribution shift while recall does not.

[5 Marks]

c) Consider two strategies for choosing the deployed decision rule:

(1) maximizing training accuracy;

(2) minimizing expected deployment cost.

Explain why strategy (1) can lead to a higher expected cost at deployment than strategy (2).

[3 Marks]

[Total 13 Marks]