Problem 5

Considering a binary classification task where each instance is assigned a score by the classifier.

Notation:

- Let $P = \{x_1^+, x_2^+, \ldots, x_m^+\}$ be the set of m positive instances.
- Let $N = \{x_1^-, x_2^-, \ldots, x_n^-\}$ be the set of n negative instances.
- Let $s(x)$ be the score assigned to instance x by the classifier.

The AUC is calculated as the area under the curve defined by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for all possible classification thresholds.

Alternative Definition of AUC

Defining a function $\phi(x^+, x^-)$ over positive and negative instance pairs:

$$\phi(x^+, x^-) = \begin{cases} 1 & \text{if } s(x^+) > s(x^-) \\ 0.5 & \text{if } s(x^+) = s(x^-) \\ 0 & \text{if } s(x^+) < s(x^-) \end{cases}$$

Then the AUC can be defined as:

$$\text{AUC} = \frac{1}{mn} \sum_{x^+ \in P} \sum_{x^- \in N} \phi(x^+, x^-)$$

This definition calculates the proportion of positive-negative pairs (x^+, x^-) such that the positive instance has a higher score than the negative instance.

Equivalence to ROC Area

To see why this is equivalent to the ROC area:

- The ROC curve is constructed by varying the classification threshold from $+\infty$ to $-\infty$, and at each point computing TPR and FPR.
- Each time a positive instance is passed over, TPR increases by $\frac{1}{m}$.
- Each time a negative instance is passed over, FPR increases by $\frac{1}{n}$.

Thus, the ROC curve is a step function consisting of m vertical steps and n horizontal steps. The area under the ROC curve corresponds to the expected TPR for each FPR step, which can be interpreted as:

$$\text{AUC} = \Pr(s(x^+) > s(x^-)) + 0.5 \cdot \Pr(s(x^+) = s(x^-))$$

This is exactly the average value of $\phi(x^+, x^-)$ over all mn positive-negative pairs, as defined above.

Conclusion

Therefore, the AUC is equal to the proportion of positive-negative pairs that are correctly ranked by the classifier. This provides both a probabilistic and geometric interpretation of the AUC metric.