CS-470: Machine Learning

Week 5 — Precision, Recall, AUC-ROC

Instructor: Dr. Sajjad Hussain

Department of Electrical and Computer Engineering
SEECS, NUST

October 9th, 2025

Learning Objectives

- Understand common classification evaluation metrics.
- Recognize why accuracy can be misleading.
- Learn to read and interpret the confusion matrix.
- Define and interpret precision, recall and F1-score.
- Explore precision–recall tradeoffs and thresholding.
- Study ROC curves and AUC as threshold-independent metrics.

Why We Need Evaluation Metrics

- Classification models output discrete labels (or scores) instead of continuous predictions.
- We must select metrics that reflect the real-world cost of different errors.
- Different applications require emphasizing different types of errors (false positives vs false negatives).
- We begin with the most intuitive metric: **accuracy**, then generalize.

Accuracy: The Simplest Metric

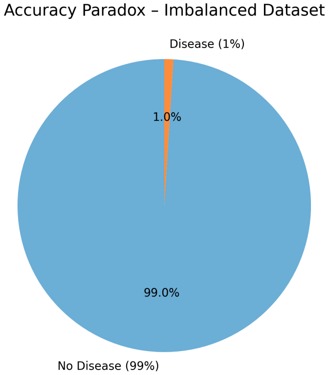
Definition

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Accuracy is easy to compute and understand.
- However, accuracy can be *misleading* when classes are imbalanced.

Accuracy Paradox (Imbalanced Data)

Consider a disease screening problem where only 1% of the population has the disease. A model that always predicts “no disease” achieves 99% accuracy, yet it never detects any sick patient. This shows that accuracy alone may hide poor performance on the important (rare) class.



Confusion Matrix — Basic Concept

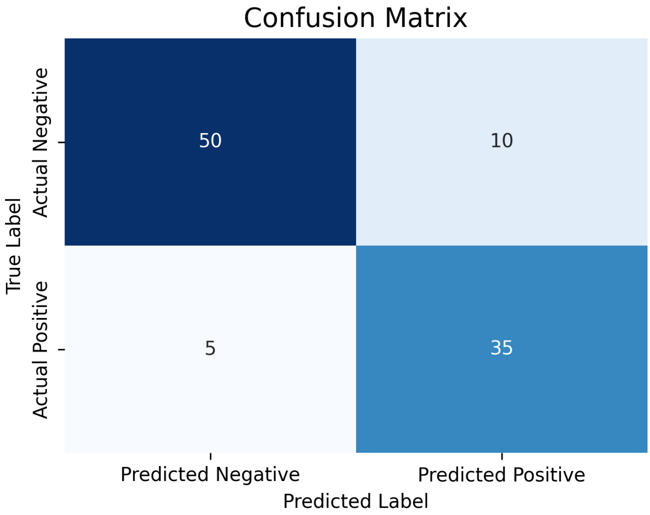
The confusion matrix gives a complete summary of prediction outcomes for binary classification:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Each cell has a clear interpretation:

- TP: correctly predicted positives,
- FP: false alarms (predicted positive but actually negative),
- FN: missed positives (predicted negative but actually positive),
- TN: correctly predicted negatives.

Visualizing the Confusion Matrix



Precision and Recall

Precision and **Recall** focus on the positive class and are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall (TPR)} = \frac{TP}{TP + FN}$$

Interpretations:

- Precision answers: "When the model predicts positive, how often is it correct?"
- Recall answers: "Of all actual positives, how many did the model find?"

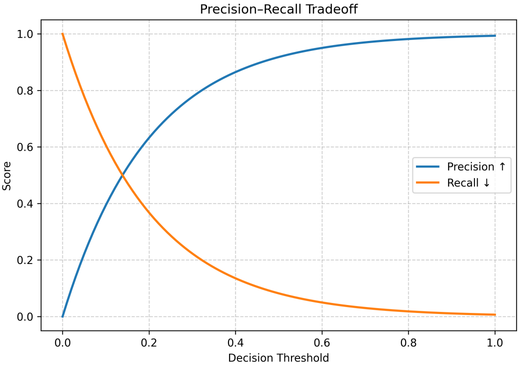
Precision vs. Recall: Intuition and Example

- Consider spam detection:
  - High precision means few legitimate emails are marked as spam (low false positives).
  - High recall means most spam emails are detected (low false negatives).
- The priority (precision or recall) depends on the application: customer-facing systems often prefer high precision; safety-critical systems often prefer high recall.

The Precision–Recall Tradeoff and Thresholding

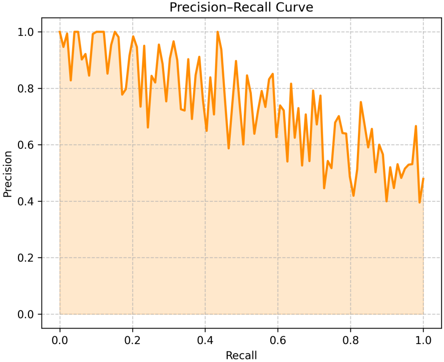
Many classifiers produce a probability score. Converting score to label requires choosing a decision threshold.

- Lower threshold $\rightarrow$ more predictions labeled positive $\Rightarrow$ higher recall, lower precision.
- Higher threshold $\rightarrow$ fewer positives $\Rightarrow$ higher precision, lower recall.



Precision–Recall Curve

- The precision–recall (PR) curve plots Precision vs Recall for different thresholds.
- Useful for imbalanced problems: it focuses on the positive class and ignores the abundance of true negatives.



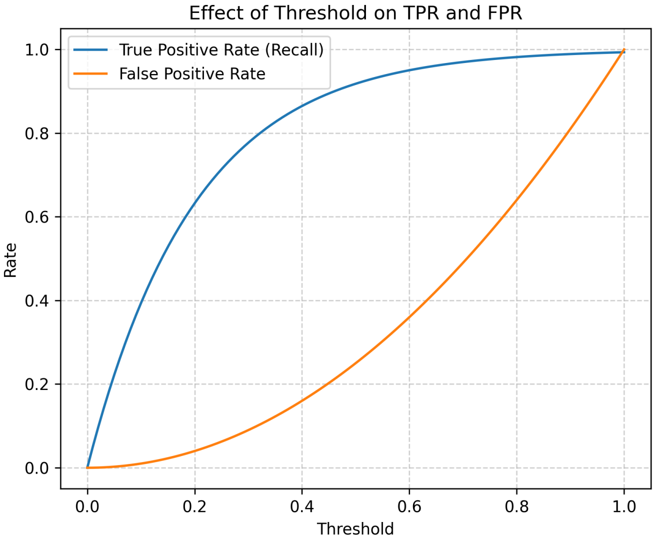
F1 Score — Single-number Summary

When you need a single metric that balances precision and recall, use the F1 score:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is the harmonic mean of precision and recall. It penalizes extreme imbalances: a classifier with very high precision but very low recall (or vice versa) will have a low F1.

Threshold Effects on TPR and FPR



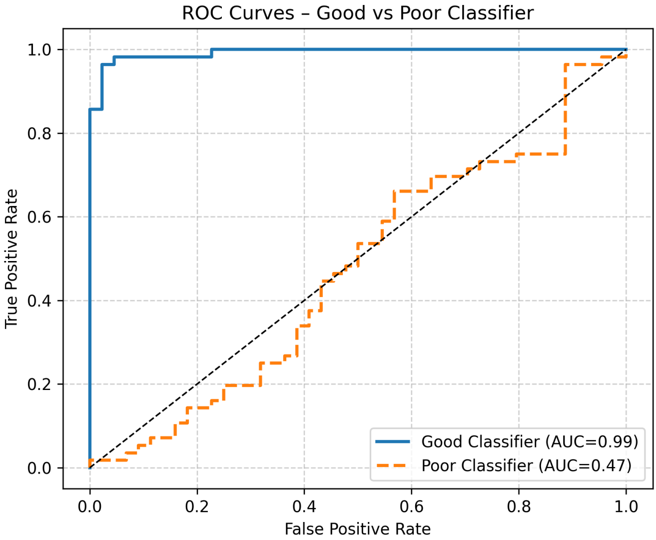
ROC Curve — Receiver Operating Characteristic

- The ROC curve plots **True Positive Rate (TPR)** vs **False Positive Rate (FPR)** for all possible thresholds:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}.$$

- Each point on the ROC corresponds to a threshold.
- A random classifier yields the diagonal line from (0,0) to (1,1).
- Better classifiers produce curves that bow toward the upper left corner.

ROC Curve Example



AUC — Area Under the ROC Curve

- **AUC** measures the area under the ROC curve and summarizes the model's discriminative power across all thresholds.
- Interpretation: *Probability that a randomly chosen positive instance ranks higher than a randomly chosen negative instance.*
- AUC ranges from 0.5 (random) to 1.0 (perfect). Values closer to 1.0 indicate better ranking performance.

When to Use PR Curve vs ROC Curve

- **ROC** is appropriate when classes are roughly balanced or when you care about both TPR and FPR across thresholds.
- **PR** is more informative for highly imbalanced datasets or when the positive class is the primary focus.
- Compare models using the metric most aligned with your application goals (AUC-ROC, AUC-PR, F1, etc.).

Practical Tips

- Always inspect the confusion matrix first — it reveals the types of mistakes your model makes.
- Choose the metric that aligns with domain costs (e.g., in medical screening, prefer high recall).
- When comparing models, evaluate across multiple metrics (AUC, F1, precision at fixed recall, etc.).
- Use cross-validation to get robust estimates of these metrics.

Summary

- Accuracy is simple but can be misleading on imbalanced data.
- Confusion matrix is the basis for deriving precision, recall, and F1.
- Thresholding changes precision/recall — PR curves visualize the tradeoff.
- ROC and AUC provide threshold-independent evaluation; AUC is useful for ranking comparisons.
- Always pick metrics based on application requirements and costs of different error types.

Next Lecture

- Week 6: Support Vector Machines (SVMs)
- Topics preview:
  - Large margin classification,
  - Soft margins and regularization,
  - Kernel trick for nonlinear decision boundaries.