Evaluation Metrics for Classification Models

21 November 2025 12:01

Evaluation metrics tell you **how well your classification model performs**. The choice of metric depends on the problem, dataset imbalance, and real-world constraints.

This tutorial covers:

1. **Core Concepts**
2. **Confusion Matrix**
3. **Basic Metrics (Accuracy, Precision, Recall, F1)**
4. **Advanced Metrics (AUC, ROC, PR curves, Log Loss)**
5. **Metrics for Imbalanced Data**
6. **Metrics for Multi-class & Multi-label Classification**

1. ⭐ Core Concepts

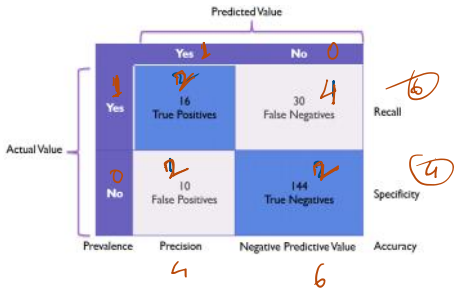
Classification problems output **discrete labels** (e.g., spam/not spam, disease/healthy, object categories).

A model's performance is measured using:

- **Predicted labels**
- **Ground-truth labels**
- **Probabilities (for probabilistic metrics)**

2. 🎯 Confusion Matrix — The Foundation

The **confusion matrix** is a 2×2 table for binary classification.



From this table, all major metrics are derived.

3. 📏 Basic Metrics

3.1 Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

✓ Good when classes are **balanced**

✗ Misleading for **imbalanced data**

Example: 99% accuracy when 99% samples belong to one class.

3.2 Precision

"How many predicted positives were actually positive?"

$$\text{Precision} = \frac{TP}{TP + FP}$$

Useful when:

- False positives are costly (e.g., spam filter marking real mail as spam).

3.3 Recall (Sensitivity/TPR)

"How many actual positives did we correctly identify?"

$$\text{Recall} = \frac{TP}{TP + FN}$$

Useful when:

- Missing positives is costly (e.g., cancer detection).

3.4 F1-Score

Harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

It is useful when:

- You want a balance between precision and recall.
- The dataset is imbalanced.

For **multi-class** and **multi-label** problems, F1-score comes in different variants.

1. Binary F1-Score

Used when:

- You have two classes (e.g., positive vs negative)
- You treat one class as “positive”

2. Macro F1-Score

$$Macro\ F1 = \frac{F1_1 + F1_2 + ... + F1_k}{k}$$

- Computes **F1 for each class independently**
- **Averages them equally**
- **Does NOT consider class imbalance**

✓ Good when class performance should be treated equally

✗ Not suitable for heavily imbalanced datasets

3. Micro F1-Score

$$Micro\ F1 = \frac{2 \times (TP_{total})}{2 \times TP_{total} + FP_{total} + FN_{total}}$$

- Aggregates global **TP, FP, FN** across all classes
- Computes precision and recall at the global level
- Equivalent to accuracy in multi-class classification

✓ Good when each sample has equal importance

✗ Not good when you need per-class fairness

4. Weighted F1-Score

$$Weighted\ F1 = \sum_{i=1}^{k} w_i \times F1_i$$

where

$$w_i = \frac{\text{support of class i}}{\text{total samples}}$$

- Computes F1 for each class
- Weighted by how many samples each class has (support)

✓ Best for **imbalanced datasets**

✓ Still gives per-class insight

✗ Majority class can dominate the metric

Summary Table of F1-Score Types

Type	Use Case	Considers Class Imbalance?	Notes
Binary F1	Single positive class	❌	For binary classification only
Macro F1	Equal class importance	❌	Good for fairness across classes
Micro F1	Overall performance	✓ (implicitly)	Treats all samples equally
Weighted F1	Imbalanced classes	✓	Most commonly used in real tasks

Which F1 Should You Use?

Task Type	Recommended F1
Binary classification	Binary F1
Balanced multi-class	Macro
Imbalanced multi-class	Weighted
Want a single score emphasizing performance on all samples	Micro

4. 🔲 Advanced Metrics

4.1 ROC Curve

- Plots **True Positive Rate (Recall)** vs **False Positive Rate**
- Shows model performance across thresholds.

AUC-ROC

Area under ROC curve

$$AUC \approx P(\text{positive class has higher score than negative class})$$

✓ Best for:

- Balanced data
- Binary classification

4.2 Precision–Recall Curve

Plots **Precision vs Recall** at different thresholds.

✓ Better than ROC for **highly imbalanced** data.

Average Precision (AP)

Area under PR curve.

4.3 Log Loss (Cross-Entropy Loss)

Measures how well probabilities match actual labels.

$$LogLoss = -\frac{1}{N}\sum(y\log(p) + (1-y)\log(1-p))$$

✓ Probabilistic metric

✓ Lower is better

Used in:

- Kaggle competitions
- Well-calibrated probability models

5. ⚠ Metrics for Imbalanced Classification

For skewed datasets (fraud detection, medical diagnosis), use:

Sensitivity (Recall)

$$TP/(TP+FN)$$

Specificity

$$TN/(TN+FP)$$

Balanced Accuracy

$$\frac{Recall_{positive} + Recall_{negative}}{2}$$

F2 or F0.5 Score

Weights recall higher or precision higher.

Matthews Correlation Coefficient (MCC)

Best single metric for imbalanced problems.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

✓ Range: −1 to +1

✓ Robust even when classes are extreme (e.g., 99:1)

6. Multi-Class Metrics

For multi-class classification, metrics are computed using:

- **Macro** (unweighted mean of per-class scores)
- **Micro** (global counts)
- **Weighted** (weighted by class frequency)

Common metrics:

- Macro F1
- Weighted accuracy
- Top-k accuracy (e.g., Top-5 accuracy used in ImageNet)
- Cohen's Kappa (measures inter-class agreement)

7. Multi-Label Classification Metrics

In multi-label problems, predictions assign **multiple labels** to one sample.

Metrics include:

- **Hamming Loss**
- **Jaccard Index**
- **Subset Accuracy** (strictest—exact match required)

Summary Table

Metric	Good For	Bad For	Notes
Accuracy	Balanced data	Imbalanced data	Overall correctness
Precision	FP costly	FN costly	"How precise are positive predictions?"
Recall	FN costly	FP acceptable	"How many positives did we catch?"
F1	Precision = Recall	Probabilistic metrics	Harmonic mean
AUC-ROC	Balanced binary	Extreme imbalance	Rank-based
AUC-PR	Imbalanced data	Balanced	Focus on minority class
Log Loss	Probability calibration	Hard-label only	Penalizes wrong confidence
MCC	Imbalanced data	Very large datasets	Best single metric

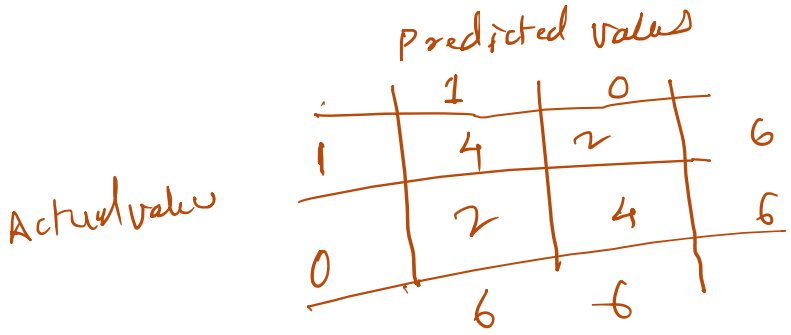
Patient	True Label (Actual y)	Predicted Probability (ŷ)	Pred label
1	1	0.95	1
2	0	0.85	1
3	1	0.90	1
4	0	0.10	0
5	1	0.80	1
6	0	0.05	0
7	1	0.60	1
8	1	0.30	0
9	0	0.70	1
10	0	0.40	0
11	1	0.20	0
12	0	0.25	0

TP → 4

TN → 4

FP → 2

FN → 2



Accuracy Imbalance (1000)

(950 → positive (1))

50 → Negative (0)

predict → 1000 positive (1)

TP → 950

FP → 50

FN → 0

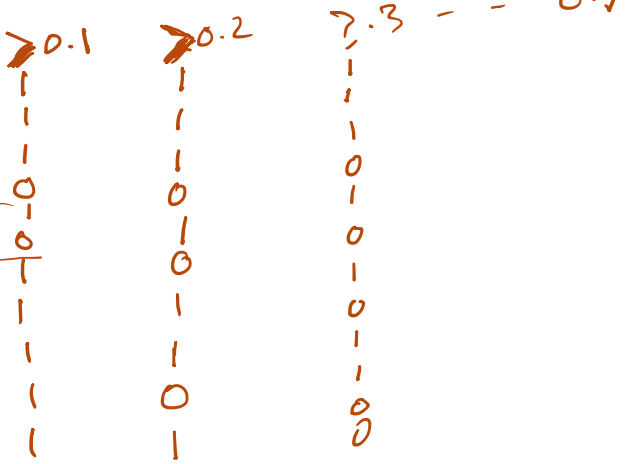
TN → 0

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{950 + 0}{950 + 0 + 50 + 0} = \frac{950}{1000} = 95\%$$

$= \frac{950}{1000} = 95\%$

ROC

Patient	True Label (Actual y)	Predicted Probability (ŷ)
1	1	0.95
2	0	0.85
3	1	0.90
4	0	0.10
5	1	0.80
6	0	0.05
7	1	0.60
8	1	0.30
9	0	0.70
10	0	0.40
11	1	0.20
12	0	0.25



$TPR = \frac{TP}{TP+FN}$ $\quad FPR = \frac{FP}{FP+TN}$

0.1

TP → 6
FN → 0
TN → 2
FP → 4

$TPR = \frac{6}{6+0} = 1$

$FPR = \frac{4}{2+4} = 0.66$

0.2

$TP = 5$
$TN = 2$
$FP = 4$
$FN = 1$

$TPR = \frac{5}{5+1} = \frac{5}{6} = 0.83$

$FPR = \frac{4}{4+2} = \frac{4}{6} = 0.66$

0.3

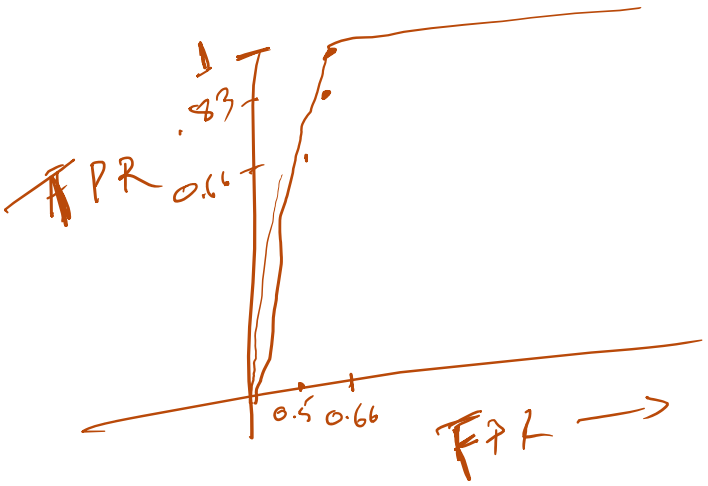
$TPR = \frac{TP}{TP+FN} = \frac{4}{4+2} = \frac{4}{6} = 0.66$

0.3

TP → 4
TN → 3
FP → 3
FN → 2

TPR = $\frac{TP}{TP+FN}$

FPR = $\frac{FP}{FP+TN} = \frac{3}{3+3} = \frac{1}{2} = 0.5$



Task 4: ROC Curve Table

Generate a table of **TPR and FPR at different thresholds** (e.g., 0.9, 0.8, ..., 0.1), and create a rough sketch of the ROC curve (or compute AUC numerically if you prefer).

Task 5: Compute AUC-ROC

Using the TPR and FPR values from Task 4:

- Plot the **ROC curve** (TPR vs FPR).
- Use the **trapezoidal rule** to compute **AUC**:

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR \approx \sum \left(\frac{TPR_i + TPR_{i+1}}{2} \cdot (FPR_{i+1} - FPR_i) \right)$$

(0, 0.1)

(0.1, 0.2)

0.1 → TPR = 1
FPR = 0.66

0.2 → TPR = 0.83
FPR = 0.66

$\frac{1 + 0.83}{2} (0.66 - 0.6)$

0

(0.2, 0.3)

^


(0.9 → 1)