# Machine Learning

## Model Evaluation Metrics

Understanding Model Performance

## Class Objectives

By the end of this class, students will be able to:

- Interpret a Confusion Matrix
- Calculate and distinguish between Accuracy and Error Rate
- Understand Precision, Recall, and their trade-offs
- Compute and interpret the F1-Score
- Evaluate regression models using MAE, MSE, RMSE
- Understand and calculate $R^2$ (Coefficient of Determination)
- Choose the appropriate metric for different scenarios

## Why Model Evaluation Metrics?

Model evaluation metrics help us answer:

- **How well is my model performing?**
- **Is my model making the right predictions?**
- **Which model should I choose?**
- **What type of errors is my model making?**

Different metrics serve different purposes — choosing the right one depends on your problem domain and business requirements.

# 1. Confusion Matrix

## Definition

A confusion matrix is a table that visualizes the performance of a classification model by comparing actual vs. predicted values.

For **binary classification**:

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Actual Positive** | TP (True Positive) | FN (False Negative) |
| **Actual Negative** | FP (False Positive) | TN (True Negative) |

## Confusion Matrix Terms

- **True Positive (TP)**: Correctly predicted positive class
- **True Negative (TN)**: Correctly predicted negative class
- **False Positive (FP)**: Incorrectly predicted positive (Type I Error)
- **False Negative (FN)**: Incorrectly predicted negative (Type II Error)

**Example**: Email spam detection

- TP: Spam correctly classified as spam
- TN: Not spam correctly classified as not spam
- FP: Not spam incorrectly classified as spam
- FN: Spam incorrectly classified as not spam

# Confusion Matrix Example

Given predictions for 100 emails:

| | Predicted Spam | Predicted Not Spam |
|---|---|---|
| **Actual Spam** | 40 (TP) | 10 (FN) |
| **Actual Not Spam** | 5 (FP) | 45 (TN) |

- Total Spam emails: 40 + 10 = 50
- Total Not Spam emails: 5 + 45 = 50
- Total predictions: 100

## 2. Accuracy

### Definition

Accuracy measures the proportion of correct predictions out of total predictions.

### Formula

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

Or simply:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

## Accuracy Calculation Example

Using our spam detection confusion matrix:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{40 + 45}{40 + 45 + 5 + 10} = \frac{85}{100} = 0.85$$

**Accuracy = 85%**

The model correctly classified 85 out of 100 emails.

## When to Use Accuracy?

**Use when:**

- Classes are balanced (roughly equal distribution)
- All types of errors are equally important

**Do NOT use when:**

- Classes are imbalanced (e.g., 95% negative, 5% positive)
- Different types of errors have different costs

**Example of misleading accuracy**:
If 95% of emails are not spam, a model that always predicts "not spam" achieves 95% accuracy but is useless!

# Error Rate

## Definition

Error Rate (also called Misclassification Rate) is the proportion of incorrect predictions.

## Formula

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

Or:

$$\text{Error Rate} = 1 - \text{Accuracy}$$

## Error Rate Calculation Example

Using our spam detection confusion matrix:

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN} = \frac{5 + 10}{100} = \frac{15}{100} = 0.15$$

Or:

$$\text{Error Rate} = 1 - 0.85 = 0.15$$

**Error Rate = 15%**

The model incorrectly classified 15 out of 100 emails.

## 3. Precision

### Definition

Precision measures the proportion of positive predictions that were actually correct. It answers: **"Of all the instances we predicted as positive, how many were truly positive?"**

### Formula

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

Precision focuses on **minimizing False Positives**.

## Precision Calculation Example

Using our spam detection confusion matrix:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{40}{40 + 5} = \frac{40}{45} = 0.889$$

**Precision = 88.9%**

Of all emails predicted as spam, 88.9% were actually spam.
11.1% were false alarms (legitimate emails marked as spam).

## When to Use Precision?

**Use when:**

- **False Positives are costly**
- You want to be confident when you predict positive

**Examples:**

- **Email spam filter**: False Positive means important email goes to spam
- **Medical diagnosis**: False Positive means healthy person gets treatment
- **Criminal conviction**: False Positive means innocent person convicted

In these cases, we prefer to miss some positive cases rather than incorrectly flag negatives as positive.

## 4. Recall (Sensitivity/True Positive Rate)

### Definition

Recall measures the proportion of actual positives that were correctly identified. It answers: **"Of all the actual positive instances, how many did we correctly identify?"**

### Formula

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

Recall focuses on **minimizing False Negatives**.

## Recall Calculation Example

Using our spam detection confusion matrix:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{40}{40 + 10} = \frac{40}{50} = 0.80$$

**Recall = 80%**

Of all actual spam emails, we correctly identified 80%.
We missed 20% of spam emails (they went to inbox).

## When to Use Recall?

**Use when:**

- **False Negatives are costly**
- You want to catch as many positive cases as possible

**Examples:**

- **Disease detection**: False Negative means sick person not diagnosed
- **Fraud detection**: False Negative means fraud goes undetected
- **Airport security**: False Negative means threat not detected

In these cases, we prefer some false alarms over missing actual positive cases.

## Precision vs. Recall Trade-off

There is often a trade-off between Precision and Recall:

- **High Precision, Low Recall**: Model is conservative, only predicts positive when very confident (few false alarms, but misses many positives)
- **Low Precision, High Recall**: Model is aggressive, predicts positive liberally (catches most positives, but many false alarms)

You cannot maximize both simultaneously — you must choose based on your problem requirements.

# Precision-Recall Example Comparison

**Scenario 1: Strict Spam Filter** (High Precision)

- TP=30, FP=2, FN=20, TN=48
- Precision = 30/(30+2) = 93.75%
- Recall = 30/(30+20) = 60%
- Few false alarms, but misses many spam emails

**Scenario 2: Aggressive Spam Filter** (High Recall)

- TP=45, FP=15, FN=5, TN=35
- Precision = 45/(45+15) = 75%
- Recall = 45/(45+5) = 90%
- Catches most spam, but more false alarms

## 5. F1-Score

### Definition

The F1-Score is the harmonic mean of Precision and Recall. It provides a single metric that balances both concerns.

### Formula

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

Or equivalently:

$$\text{F1-Score} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

## F1-Score Calculation Example

Using our spam detection confusion matrix:

- Precision = 0.889
- Recall = 0.80

$$\text{F1-Score} = 2 \times \frac{0.889 \times 0.80}{0.889 + 0.80} = 2 \times \frac{0.7112}{1.689} = 2 \times 0.421 = 0.842$$

**F1-Score = 84.2%**

## Why Harmonic Mean?

The harmonic mean is used instead of arithmetic mean because:

- It **penalizes extreme values**
- If either Precision or Recall is very low, F1-Score will be low
- Ensures both metrics are reasonably good

**Example:**

- Precision = 1.0, Recall = 0.1
- Arithmetic mean = (1.0 + 0.1) / 2 = 0.55
- Harmonic mean (F1) = 2 × (1.0 × 0.1) / (1.0 + 0.1) = 0.18

The F1-Score better reflects that the model is not performing well overall.

# When to Use F1-Score?

**Use when:**

- You need to balance Precision and Recall
- Classes are imbalanced
- Both False Positives and False Negatives matter

**Examples:**

- Information retrieval systems
- Medical diagnosis where both types of errors are important
- When you need a single metric to compare models

F1-Score is particularly useful when you cannot afford to ignore either Precision or Recall.

# Classification Metrics Summary

| Metric | Formula | Focus | Use When |
|---|---|---|---|
| Accuracy | $(TP + TN)/Total$ | Overall correctness | Balanced classes |
| Error Rate | $1 - Accuracy$ | Overall errors | Balanced classes |
| Precision | $TP/(TP + FP)$ | Avoid false alarms | FP is costly |
| Recall | $TP/(TP + FN)$ | Catch all positives | FN is costly |
| F1-Score | $2 \times (P \times R)/(P + R)$ | Balance P & R | Both FP & FN costly |

## Regression Metrics Overview

For regression problems (predicting continuous values), we use different metrics:

- **MAE**: Mean Absolute Error
- **MSE**: Mean Squared Error
- **RMSE**: Root Mean Squared Error
- **$R^2$**: Coefficient of Determination

All these metrics evaluate how close predictions are to actual values.

# 6. Mean Absolute Error (MAE)

## Definition

MAE is the average of the absolute differences between predicted and actual values.

## Formula

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Where:

- $n$ = number of observations
- $y_i$ = actual value
- $\hat{y}_i$ = predicted value

# MAE Calculation Example

Predict house prices (in $1000s):

| Actual Price $(y_i)$ | Predicted Price $(\hat{y}_i)$ | Error $(y_i - \hat{y}_i)$ | Absolute Error $\lvert y_i - \hat{y}_i \rvert$ |
|---|---|---|---|
| 250 | 240 | 10 | 10 |
| 300 | 320 | -20 | 20 |
| 180 | 175 | 5 | 5 |
| 400 | 390 | 10 | 10 |
| 220 | 230 | -10 | 10 |

$$\text{MAE} = \frac{10 + 20 + 5 + 10 + 10}{5} = \frac{55}{5} = 11$$

Average error is $11,000 in either direction.

## 7. Mean Squared Error (MSE)

### Definition

MSE is the average of the squared differences between predicted and actual values.

### Formula

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

# MSE Calculation Example

Using the same house price data:

| Actual $(y_i)$ | Predicted $(\hat{y}_i)$ | Error $(y_i - \hat{y}_i)$ | Squared Error $(y_i - \hat{y}_i)^2$ |
|---|---|---|---|
| 250 | 240 | 10 | 100 |
| 300 | 320 | -20 | 400 |
| 180 | 175 | 5 | 25 |
| 400 | 390 | 10 | 100 |
| 220 | 230 | -10 | 100 |

$$\text{MSE} = \frac{100 + 400 + 25 + 100 + 100}{5} = \frac{725}{5} = 145$$

## 8. Root Mean Squared Error (RMSE)

### Definition

RMSE is the square root of MSE, bringing the error metric back to the original units.

### Formula

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

## RMSE Calculation Example

Using the MSE from our house price example:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{145} \approx 12.04$$

The RMSE is approximately $12,040.

**Interpretation:** The model's predictions deviate from actual prices by about $12,000 on average, with larger errors weighted more heavily.

## MAE vs. MSE vs. RMSE Comparison

Using our house price example:

- **MAE = 11** (average absolute error)
- **MSE = 145** (average squared error)
- **RMSE = 12.04** (root average squared error)

**Key Difference:**

- MAE treats all errors equally: $|10|, |20|, |5|, |10|, |10|$
- MSE/RMSE penalize large errors more: $100, 400, 25, 100, 100$

The error of 20 has 4× the impact in MSE compared to error of 10.

# MAE vs. RMSE: Which to Use?

| Aspect | MAE | RMSE |
|---|---|---|
| **Interpretation** | Average absolute error | Root average squared error |
| **Units** | Same as target | Same as target |
| **Outlier Sensitivity** | Low | High |
| **Large Error Penalty** | Linear | Quadratic |
| **Use When** | Outliers are noise | Large errors are critical |

**Rule of thumb:**

- If all errors are equally bad → Use MAE
- If large errors are worse → Use RMSE

## 9. $R^2$ (R-Squared / Coefficient of Determination)

### Definition

$R^2$ represents the proportion of variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1 (can be negative for poor models).

### Formula

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

- $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (Residual Sum of Squares)
- $SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (Total Sum of Squares)
- $\bar{y}$ = mean of actual values

# $R^2$ Calculation Example

House prices: $y = [250, 300, 180, 400, 220]$

Mean: $\bar{y} = \frac{250+300+180+400+220}{5} = \frac{1350}{5} = 270$

| $y_i$ | $\hat{y}_i$ | $(y_i - \hat{y}_i)^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|
| 250 | 240 | 100 | 400 |
| 300 | 320 | 400 | 900 |
| 180 | 175 | 25 | 8100 |
| 400 | 390 | 100 | 16900 |
| 220 | 230 | 100 | 2500 |

## $R^2$ Calculation (Continued)

$$SS_{res} = 100 + 400 + 25 + 100 + 100 = 725$$

$$SS_{tot} = 400 + 900 + 8100 + 16900 + 2500 = 28800$$

$$R^2 = 1 - \frac{725}{28800} = 1 - 0.0252 = 0.9748$$

**$R^2$ = 0.9748 or 97.48%**

The model explains 97.48% of the variance in house prices.

## Interpreting $R^2$

- **$R^2 = 1.0$**: Perfect predictions (all points on the line)
- **$R^2 = 0.9$**: 90% of variance explained (very good)
- **$R^2 = 0.7$**: 70% of variance explained (good)
- **$R^2 = 0.5$**: 50% of variance explained (moderate)
- **$R^2 = 0.0$**: Model no better than predicting the mean
- **$R^2 < 0$**: Model worse than predicting the mean

**Higher $R^2$ = Better model fit**

But beware: High $R^2$ doesn't always mean a good model (could be overfitting).

## Regression Metrics Summary

| Metric | Formula | Range | Interpretation | Use When |
|---|---|---|---|---|
| MAE | $\frac{1}{n}\sum\lvert y_i - \hat{y}_i\rvert$ | $[0, \infty)$ | Avg absolute error | Equal error weight |
| MSE | $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$ | $[0, \infty)$ | Avg squared error | Penalize large errors |
| RMSE | $\sqrt{MSE}$ | $[0, \infty)$ | Root avg sq. error | Interpretable + penalize |
| $R^2$ | $1 - \frac{SS_{res}}{SS_{tot}}$ | $(-\infty, 1]$ | Variance explained | Model comparison |

Lower is better for MAE, MSE, RMSE. Higher is better for $R^2$.

# Choosing the Right Metric: Decision Guide

**For Classification:**

1. Are classes balanced? → **Accuracy**
2. Is FP costly? → **Precision**
3. Is FN costly? → **Recall**
4. Are both FP & FN costly? → **F1-Score**
5. Need detailed analysis? → **Confusion Matrix**

**For Regression:**

1. Want simple interpretation? → **MAE**
2. Want to penalize large errors? → **RMSE**
3. Want to measure explained variance? → **$R^2$**
4. For optimization? → **MSE**

## Real-World Examples

### Medical Diagnosis (Cancer Detection):

- Metric: **Recall** (minimize False Negatives)
- Reason: Missing a cancer diagnosis is catastrophic

### Credit Card Fraud Detection:

- Metric: **F1-Score** or **Precision** and **Recall**
- Reason: Balance catching fraud vs. blocking legitimate transactions

### House Price Prediction:

- Metric: **RMSE** and **$R^2$**
- Reason: Penalize large prediction errors, measure overall fit

## Common Mistakes to Avoid

1. **Using Accuracy with imbalanced data**
   - Can be misleading when one class dominates
2. **Ignoring the cost of different errors**
   - FP and FN often have different impacts
3. **Using MSE for interpretation**
   - Squared units are hard to understand; use RMSE
4. **Relying only on $R^2$**
   - Doesn't show prediction error magnitude
5. **Not using Confusion Matrix**
   - Always start with confusion matrix for classification

## Practice Problem 1: Classification

A medical test for a rare disease (5% prevalence) gives these results for 1000 patients:

| | Predicted Positive | Predicted Negative |
|---|---|---|
| **Actual Positive** | 45 | 5 |
| **Actual Negative** | 95 | 855 |

Calculate:

1. Accuracy
2. Precision
3. Recall
4. F1-Score

## Solution: Practice Problem 1

Given: TP=45, FN=5, FP=95, TN=855

1. **Accuracy** = $\frac{45+855}{1000} = \frac{900}{1000} = 0.90$ or **90%**
2. **Precision** = $\frac{45}{45+95} = \frac{45}{140} = 0.321$ or **32.1%**
3. **Recall** = $\frac{45}{45+5} = \frac{45}{50} = 0.90$ or **90%**
4. **F1-Score** = $2 \times \frac{0.321 \times 0.90}{0.321+0.90} = 2 \times \frac{0.289}{1.221} = 0.473$ or **47.3%**

## Analysis: Practice Problem 1

- **High Accuracy (90%)**: But misleading! Predicting everyone as negative gives 95% accuracy.
- **Low Precision (32.1%)**: Many false alarms — only 1 in 3 positive predictions is correct.
- **High Recall (90%)**: Good at catching actual positive cases.
- **Moderate F1 (47.3%)**: Reflects the imbalance between precision and recall.

**Conclusion**: This test has too many false positives. Might need confirmation test for positive results.

## Practice Problem 2: Regression

Predict student exam scores:

| Actual | Predicted |
|---|---|
| 85 | 80 |
| 90 | 92 |
| 75 | 78 |
| 95 | 90 |
| 70 | 68 |

Calculate: MAE, MSE, RMSE, $R^2$

# Solution: Practice Problem 2 (Part 1)

Mean actual: $\bar{y} = \frac{85+90+75+95+70}{5} = \frac{415}{5} = 83$

| $y_i$ | $\hat{y}_i$ | $\lvert y_i - \hat{y}_i \rvert$ | $(y_i - \hat{y}_i)^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|
| 85 | 80 | 5 | 25 | 4 |
| 90 | 92 | 2 | 4 | 49 |
| 75 | 78 | 3 | 9 | 64 |
| 95 | 90 | 5 | 25 | 144 |
| 70 | 68 | 2 | 4 | 169 |
| **Sum** | | **17** | **67** | **430** |

## Solution: Practice Problem 2 (Part 2)

**MAE** = $\frac{17}{5} = 3.4$ points

**MSE** = $\frac{67}{5} = 13.4$ points²

**RMSE** = $\sqrt{13.4} \approx 3.66$ points

**R²** = $1 - \frac{67}{430} = 1 - 0.156 = 0.844$ or **84.4%**

**Interpretation**: Predictions are off by ~3.4 points on average (MAE) or ~3.66 points with large errors weighted more (RMSE). The model explains 84.4% of score variance.