Chapter 30: Confusion Matrix

Introduction

In Artificial Intelligence and Machine Learning, evaluating the performance of a model is just as important as building it. One of the most common tools used to measure the performance of classification models is the **Confusion Matrix**. It allows us to visualize how well our AI model is making predictions — especially when dealing with binary or multi-class classification tasks.

This chapter will help you understand:

- What a confusion matrix is,
- Its components,
- How to read and interpret it,
- Important performance metrics derived from it, and
- Real-life examples and activities.

30.1 What is a Confusion Matrix?

A **confusion matrix** is a **table** that helps evaluate the performance of a classification algorithm by comparing the predicted results with the actual results.

It shows how many predictions your model got right and how many it got wrong, **categorized by each class**.

30.2 Structure of a Confusion Matrix

Let's take a simple example of **binary classification** – such as predicting whether an email is **spam** or **not spam**.

The confusion matrix for this would be a 2×2 table:

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

Let's understand each term:

- **True Positive (TP)**: Model correctly predicted **positive** class. *Example:* Spam email correctly identified as spam.
- **False Positive (FP)**: Model incorrectly predicted **positive** class. *Example:* Normal email wrongly marked as spam (Type I error).

- **True Negative (TN)**: Model correctly predicted **negative** class. *Example:* Normal email correctly marked as not spam.
- **False Negative (FN)**: Model incorrectly predicted **negative** class. *Example:* Spam email marked as not spam (Type II error).

30.3 Key Metrics Derived from a Confusion Matrix

From the matrix, we can calculate several **performance metrics** that help evaluate how good the model is.

30.3.1 Accuracy

Accuracy = (TP + TN) / (TP + TN + FP + FN)

It tells us **how often the classifier is correct**.

30.3.2 Precision

Precision = TP / (TP + FP)

It tells us **how many of the predicted positive results were actually positive**.

30.3.3 Recall (Sensitivity or True Positive Rate)

Recall = TP / (TP + FN)

It tells us **how many actual positives were correctly predicted**.

30.3.4 F1 Score

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

It is the **harmonic mean** of Precision and Recall. Useful when you need a balance between the two.

30.4 Example with Real Data

Suppose we test an AI model on 100 emails:

- 60 are **spam** (positive class)
- 40 are **not spam** (negative class)

Model prediction results:

- TP = 50
- FP = 5
- FN = 10
- TN = 35

Let's form the confusion matrix:

	Predicted Spam	Predicted Not Spam
Actual Spam	50 (TP)	10 (FN)
Actual Not Spam	5 (FP)	35 (TN)

Now compute the metrics:

- **Accuracy** = (50 + 35) / 100 = 85%
- **Precision** = 50 / (50 + 5) = 90.9%
- **Recall** = 50 / (50 + 10) = 83.3%
- **F1 Score** = 2 × (0.909 × 0.833) / (0.909 + 0.833) ≈ 87%

30.5 Use of Confusion Matrix in AI

- Helps detect whether a model is **biased** toward one class.
- Useful when **data is imbalanced** (e.g., 90% not spam, 10% spam).
- Helps in **model improvement** by identifying the types of errors.

30.6 Confusion Matrix for Multi-Class Classification

For more than two classes, the confusion matrix becomes larger (e.g., 3×3, 4×4, etc.)

Example (3-Class Problem: Cat, Dog, Rabbit):

	Predicted Cat	Predicted Dog	Predicted Rabbit
Actual Cat	30	5	2
Actual Dog	3	40	4
Actual Rabbit	1	2	35

Each row = actual class Each column = predicted class

30.7 Common Mistakes to Avoid

- Don't rely **only on accuracy**, especially for imbalanced datasets.
- Always check **precision and recall**, especially in critical applications (like medical diagnosis).
- Use **F1-score** when you need a balance between precision and recall.

30.8 Activity/Exercise

Try this small exercise:

An AI system predicts loan approval (Approve / Reject). Here are the results:

- Actual Approve: 80 cases
- Actual Reject: 20 cases
- Correct Approve predicted: 70
- Incorrect Approve predicted: 10
- Correct Reject predicted: 15
- Incorrect Reject predicted: 5

Task: Draw the confusion matrix and calculate:

- Accuracy
- Precision
- Recall
- F1 Score

Summary

- A **confusion matrix** is a powerful tool to evaluate classification models.
- It breaks down predictions into **true positives**, **true negatives**, **false positives**, and **false negatives**.
- From it, we derive important metrics like **accuracy, precision, recall, and F1 score**.
- It helps in better understanding of model performance, especially in **imbalanced data scenarios**.
