

Week 7 - Evaluating Classification Models

Introduction:

Welcome to this lecture on the various methods used to evaluate classification models. Evaluating the performance of a classification model is crucial for assessing its accuracy and effectiveness. In this lecture, we will explore several key evaluation metrics and techniques used in the field of machine learning.

Importance of Model Evaluation:

Model evaluation helps us understand how well a classification model is performing. It provides insights into a model's strengths and weaknesses. Evaluation metrics guide model selection, hyperparameter tuning, and ultimately, decision-making in real-world applications.

Common Evaluation Metrics:

There are several common evaluation metrics used for classification models:

1. **Accuracy:** Accuracy is the most straightforward metric, measuring the ratio of correctly predicted instances to the total number of instances. It is suitable when classes are balanced, but it may not be informative for imbalanced datasets.
2. **Precision:** Precision measures the ratio of true positives (correctly predicted positive instances) to the total predicted positive instances. It focuses on the accuracy of positive predictions and is useful when minimizing false positives is essential.
3. **Recall (Sensitivity or True Positive Rate):** Recall measures the ratio of true positives to the total actual positive instances. It assesses the model's ability to identify all positive instances and is essential when minimizing false negatives is critical.
4. **Confusion Matrix:** A confusion matrix provides a detailed breakdown of a model's predictions. It includes counts of true positives, true negatives, false positives, and false negatives.

What types of prediction can you have?

True Positive (TP):

Definition: When the pregnancy test correctly identifies a positive result, indicating that the person is pregnant, and they are indeed pregnant. **Example:** A woman takes a pregnancy test and it accurately detects the presence of pregnancy hormones in her urine, confirming her pregnancy. **Calculation for Confusion Matrices:** Count the number of instances where the predicted value is true (positive, 1), and the actual value is true (positive, 1).

False Positive (FP):

A *Type 1* error occurs when the hypothesis test incorrectly rejects a true null hypothesis. This is also known as a “false positive” result. In other words, the test indicates that there is an effect or a difference when, in fact, there isn’t one.

Definition: When the pregnancy test incorrectly identifies a negative result as positive, indicating that the person is pregnant when they are not. **Example:** A woman takes a pregnancy test and it erroneously shows a positive result, suggesting pregnancy, but she is not actually pregnant. **Calculation for Confusion Matrices:** Count the number of instances when the predicted value is true (positive, 1), and the actual value is false (negative, 0).

True Negative (TN):

Definition: When the pregnancy test correctly identifies a negative result, indicating that the person is not pregnant, and they are indeed not pregnant. **Example:** A woman takes a pregnancy test, and it accurately shows a negative result, confirming that she is not pregnant, which aligns with her actual status. **Calculation for Confusion Matrices:** Count the number of instances where the predicted value is false (negative, 0), and the actual value is false (negative, 0).

False Negative (FN):

A *Type 2* error occurs when the hypothesis test fails to reject a false null hypothesis. This is known as a “false negative” result. In this case, the test suggests that there is no effect or difference when, in reality, there is one.

Definition: When the pregnancy test incorrectly identifies a positive result as negative, indicating that the person is not pregnant when they are. **Example:** A woman takes a pregnancy test, but it fails to detect the pregnancy hormones in her urine, showing a negative result, even though she is actually pregnant. **Calculation for Confusion Matrices:** Count the

number of instances when the predicted value is false (negative, 0), and the actual value is true (positive, 1).

Balancing Type 1 and Type 2 Errors

In practice, it is difficult to get 100 percent accuracy on any classification model. Therefore, there's often a trade-off between minimizing Type 1 and Type 2 errors and modellers will often have to make a conscious decision to bias the model in favour of one or the other.

The optimal balance depends on the specific context and the consequences of each type of error. For instance, in medical testing, it might be preferable to accept a higher risk of Type 1 errors (falsely identifying a condition) to reduce the risk of Type 2 errors (missing a diagnosis), given the potential health implications.

Worked Example

Suppose you have a binary classification problem with the following actual classes and predicted classes for a sample of 20 observations:

Actual Classes: 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1

Predicted Classes: 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1

Step 1: Create the Confusion Matrix:

	Predicted Negative (0)	Predicted Positive (1)
Actual Negative (0)	TN	FP
Actual Positive (1)	FN	TP

Using the provided actual and predicted classes, we can construct the confusion matrix:

	Predicted Negative (0)	Predicted Positive (1)
Actual Negative (0)	8	1
Actual Positive (1)	0	11

Step 2: Calculate Evaluation Metrics:

$$Accuracy = (TP + TN) / Total = (8 + 11) / 20 = 19 / 20 = 0.95$$

$$Precision = TP / (TP + FP) = 11 / (11 + 1) = 11 / 12 \approx 0.92$$

$$Recall(Sensitivity) = TP / (TP + FN) = 11 / (11 + 0) = 11 / 11 = 1$$

$$Specificity = TN / (TN + FP) = 8 / (8 + 1) = 8 / 9 \approx 0.89$$

$$F1Score = 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.92 * 1) / (0.92 + 1) \approx 0.96$$

Step 3: Interpretation:

- The accuracy of the model is 95%, indicating that 95% of the predictions are correct.
- The precision of the model is approximately 92%, implying that 92% of the samples predicted as positive are truly positive.
- The recall (sensitivity) of the model is 100%, meaning that 100% of the actual positive samples are correctly identified.
- The specificity of the model is approximately 89%, indicating that 89% of the actual negative samples are correctly identified.
- The F1 score, which combines precision and recall, is approximately 96%, suggesting overall good performance of the model.

This exercise provides a simplified scenario to create and evaluate a confusion matrix with a smaller dataset. It demonstrates the calculation of evaluation metrics to assess the performance of a classification model.

Conclusion:

Evaluating classification models is a critical step in the machine learning workflow. Various metrics, such as accuracy, precision, recall, F1-score, ROC curves, and confusion matrices, provide insights into model performance. Techniques like cross-validation, train-test splits, and stratified sampling help ensure robust evaluation.