Parul® University

NAAC A++ ACCREDITED UNIVERSITY

PARULINSTITUTEOF ENGINEERING &TECHNOLOGY
FACULTY OF ENGINEERING & TECHNOLOGY
PARULUNIVERSITY

Subject: Machine Learning

Unit 5 : Evaluation Metrics

Computer Science & Engineering

Jigar Sapkale (Assistant Prof. PIET-CSE)

Parul® University

NAAC A++ ACCREDITED UNIVERSITY

DIGITAL LEARNING CONTENT

Outline

- ROC Curves
  - Introduction of ROC
  - True Positive Rate (TPR)
  - False Positive Rate (FPR)
  - Use Cases of ROC Curves
- Evaluation Metrics
  - Classification Accuracy
  - Confusion Matrix
  - Recall
  - Precision
  - F1 Score

Parul® University

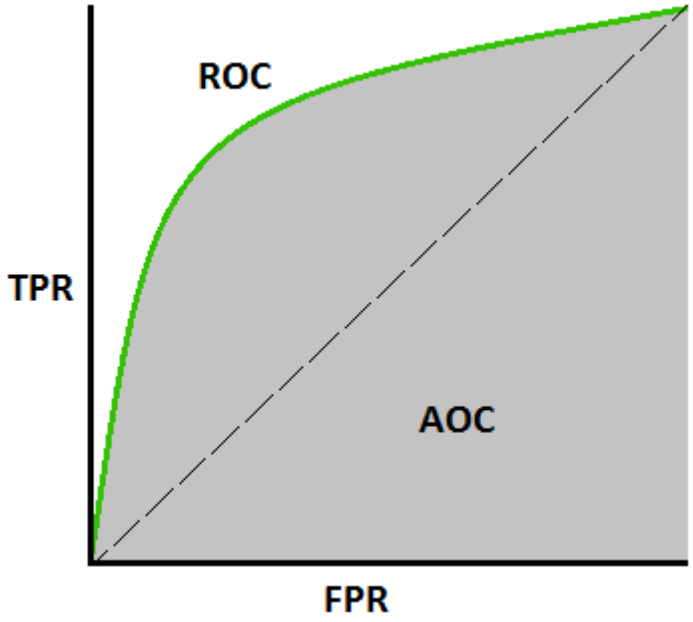
NAAC A++ ACCREDITED UNIVERSITY

Outline

- Error correction in Perceptrons

ROC Curve

Definition: An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds



True Positive Rate (TPR)

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

TRP tells us what proportion of the positive class got correctly classified.

EX: A simple example would be determining what proportion of the actual sick people were correctly detected by the model.

False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN}$$

TRP tells us what proportion of the positive class got correctly classified.

EX: A simple example would be determining what proportion of the actual sick people were correctly detected by the model.

Use Cases of ROC Curve

Medical Diagnostics: ROC curves help healthcare professionals evaluate the sensitivity and specificity of diagnostic procedures and optimize decision thresholds

Fraud Detection:ROC curves assist in balancing the trade-offs between correctly identifying fraudulent transactions (sensitivity) and minimizing false alarms (specificity).

Credit Scoring:ROC curves aid financial institutions in evaluating credit scoring models and setting appropriate thresholds for approving or denying credit.

Customer Churn Prediction:ROC curves assist in evaluating the performance of churn prediction models and optimizing retention strategies.

Sentiment Analysis:ROC curves help assess the performance of sentiment classification models in distinguishing between positive and negative sentiments.

Evaluation Metrics

An evaluation metric can be defined as a function that takes an ordered vector of relevance values, and returns a single numeric score, that summarizes those values

- Classification Accuracy
- Confusion Matrix
- F1 Score
- Recall
- Precision

Confusion Matrix

A confusion matrix is a table that visualizes the performance of a classification model by comparing predicted and actual values across different classes. It's a handy tool for evaluating the effectiveness of a model in terms of true positives, true negatives, false positives, and false negatives.

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

Confusion Matrix

- It creates a *N X N* matrix, where N is the number of classes or categories that are to be predicted. Here we have *N = 2*, so we get a *2 X 2* matrix.

- Suppose there is a problem with our practice which is a binary classification. Samples of that classification belong to either *Yes* or *No*. So, we build our classifier which will predict the class for the new input sample. After that, we tested our model with *165* samples, and we get the following result.

Confusion Matrix

n = 165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

There are 4 terms you should keep in mind:

1.True Positives: It is the case where we predicted Yes and the real output was also yes.

2.True Negatives: It is the case where we predicted No and the real output was also No.

3.False Positives: It is the case where we predicted Yes but it was actually No.

4.False Negatives: It is the case where we predicted No but it was actually Yes.

Confusion Matrix

The accuracy of the matrix is always calculated by taking average values present in the ***main diagonal***

Ex:

Accuracy = (True Positive + True Negative) / Total Sample Accuracy
= (100 + 50) / 165
= 0.91

Accuracy = 0.91

Evaluation Metrics

	Actual Class	Predicted Class
1	Positive	Positive
2	Negative	Negative
3	Negative	Positive
4	Negative	Negative
5	Positive	Positive
6	Positive	Negative
7	Negative	Negative
8	Positive	Positive
9	Negative	Positive
10	Negative	Negative
11	Positive	Positive
12	Negative	Negative
13	Positive	Positive

Outcomes	Values
True Negative (TN)	5
False Negative (FN)	1
True Positive (TP)	5
False Positive (FP)	2

Table 1: Sample outcomes of a binary classification model

Classification Accuracy

Classification accuracy is the accuracy we generally mean, whenever we use the term accuracy. We calculate this by calculating the ratio of correct predictions to the total number of input Samples.

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

Let's use the data of the model outcomes from Table 1 to calculate the accuracy of a simple classification model

Accuracy = 5 + 5 / 5 + 1 + 5 + 2
= 10 / 13
= 0.77

Accuracy = 0.77

Classification Accuracy

an accuracy score above 0.7 describes an average model performance, whereas a score above 0.9 indicates a good model. However, the relevance of the score is determined by the task. Accuracy alone may not provide a complete picture of model performance, especially In scenarios where class imbalance exists in the dataset.

Therefore, to address the constraints of accuracy, precision, and recall metrics are used.

Precision

The precision metric determines the quality of positive predictions by measuring their correctness. It is the number of true positive outcomes divided by the sum of true positive and false positive predictions.

The formula applied in calculating precision is:

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

Using the classification model outcomes from Table 1 above, precision is calculated as

Precision

$$\text{Precision} = TP / TP + FP$$
$$= 5 / 5 + 2$$
$$= 5 / 7$$
$$= 0.71$$

Precision = 0.71

Precision can be thought of as a quality metric; higher precision indicates that an algorithm provides more relevant results than irrelevant ones. It is solely focused on the correctness of positive predictions, with no attention to the correct detection of negative predictions.

Recall

Recall, also called sensitivity, measures the model's ability to detect positive events correctly. It is the percentage of accurately predicted positive events out of all actual positive events. To calculate the recall of a classification model, the formula is

$$\textbf{Recall} = \textbf{TP} / \textbf{TP} + \textbf{FN}$$

Using the classification model outcomes from Table 1 above, recall is calculated as

Recall

Recall = TP / TP + FN
= 5 / 5 + 1
= 5 / 6
= 0.83

Recall = 0.83

A high recall score indicates that the classifier predicts the majority of the relevant results correctly. However, the recall metric does not take into account the potential repercussions of false positives

Recall

i.e., occurrences that are wrongly identified as positive – a false alarm. Typically, we would like to avoid such cases, especially in mission-critical applications such as intrusion detection, where a non-malicious false alarm increases the workload of overburdened security teams.

we want to build classifiers with high precision and recall. But that's not always possible. A classifier with high recall may have low precision, meaning it captures the majority of positive classes but produces a considerable number of false positives. Hence, we use the F1 score metric to balance this precision-recall trade-off.

Difference between Precision and Recall

Precision	Recall
When a model classifies most of the positive samples correctly as well as many false-positive samples, then the model is said to be a high recall and low precision model.	When a model classifies a sample as Positive, but it can only classify a few positive samples, then the model is said to be high accuracy, high precision, and low recall model.
The precision of a machine learning model is dependent on both the negative and positive samples.	Recall of a machine learning model is dependent on positive samples and independent of negative samples.
In Precision, we should consider all positive samples that are classified as positive either correctly or incorrectly.	The recall cares about correctly classifying all positive samples. It does not consider if any negative sample is classified as positive.

Difference between Precision and Recall

- This question is very common among all machine learning engineers and data researchers. The use of Precision and Recall varies according to the type of problem being solved.
- If there is a requirement of classifying all positive as well as Negative samples as Positive, whether they are classified correctly or incorrectly, then use Precision.
- Further, on the other end, if our goal is to detect only all positive samples, then use Recall. Here, we should not care how negative samples are correctly or incorrectly classified the samples.

F1 Score

The F1 score or F-measure is described as the harmonic mean of the precision and recall of a classification model. The two metrics contribute equally to the score, ensuring that the F1 metric correctly indicates the reliability of a model.

The F1 score formula is

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} = 2 \times \frac{(Precision \times Recall) \times 1}{(Precision \times Recall) \times \frac{Precision + Recall}{Precision \times Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Using the classification model outcomes from Table 1, the F1 score is calculated as

F1 Score

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \approx \frac{2 \times 0.71 \times 0.83}{0.71 + 0.83} \approx \frac{1.1786}{1.54} \approx 0.77$$

Here, you can observe that the harmonic mean of precision and recall creates a balanced measurement, i.e., the model's precision is not optimized at the price of recall, or vice versa. As a result, the F1 score metric directs real-world decision-making more accurately

The F1 score ranges between 0 and 1, with 0 denoting the lowest possible result and 1 denoting a flawless result, meaning that the model accurately predicted each label.

F1 Score

A **high F1 score generally** indicates a well-balanced performance, demonstrating that the model can concurrently attain high precision and high recall. A **low F1 score** often signifies a trade-off between recall and precision,

F1 score	Interpretation
> 0.9	Very good
0.8 - 0.9	Good
0.5 - 0.8	OK
< 0.5	Not good

Parul® University
NAAC A++ ACCREDITED UNIVERSITY

F1 Score Application in Machine Learning

Medical Diagnostics

In medical diagnostics, it is important to acquire a high recall while correctly detecting positive occurrences, even if doing so necessitates losing precision. For instance, the F1 score of a cancer detection classifier should minimize the possibility of false negatives, i.e., patients with malignant cancer, but the classifier wrongly predicts as benign.

Sentiment Analysis

For natural language processing (NLP) tasks like sentiment analysis, recognizing both positive and negative sentiments in textual data allow businesses to assess public opinion, consumer feedback, and brand sentiment. Hence, the F1 score allows for an efficient evaluation of sentiment analysis models by taking precision and recall into account when categorizing sentiments.

F1 Score Application in Machine Learning

Fraud Detection

In fraud detection, by considering both precision (the accuracy with which fraudulent cases are discovered) and recall (the capacity to identify all instances of fraud), the F1 score enables practitioners to assess fraud detection models more accurately. For instance, the figure below shows the evaluation metrics for a credit card fraud detection model.

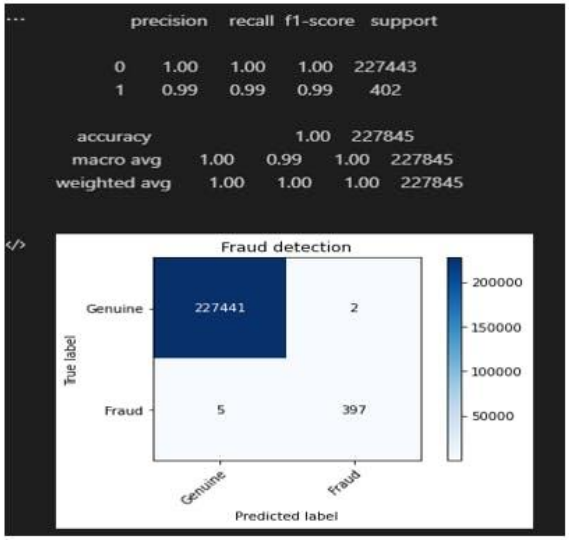


Figure: Confusion matrix for Random forest

F1 Score Limitations

Dataset Class Imbalance

For imbalanced data, when one class significantly outweighs the other, the regular F1 score metric might not give a true picture of the model's performance. This is because the regular F1 score gives precision and recall equal weight, but in datasets with imbalances, achieving high precision or recall for the minority class may result in a lower F1 score due to the majority class's strong influence.

Cost Associated with False Prediction Outcomes

False positives and false negatives can have quite diverse outcomes depending on the application. In medical diagnostics, as discussed earlier, a false negative is more dangerous than a false positive. Hence, the F1 score must be interpreted carefully.

Contextual Dependence

The evaluation of the F1 score varies depending on the particular problem domain and task objectives. Various interpretations of what constitutes a high or low F1 score for different applications require various precision-recall criteria.

Performance Metrics for Regression

- Regression is a supervised learning technique that aims to find the relationships between the dependent and independent variables.
- A predictive regression model predicts a numeric or discrete value.
- The metrics used for regression are different from the classification metrics.
- It means we cannot use the Accuracy metric (explained above) to evaluate a regression model; instead, the performance of a Regression model is reported as errors in the prediction.

Performance Metrics for Regression

Following are the popular metrics that are used to evaluate the performance of Regression models.

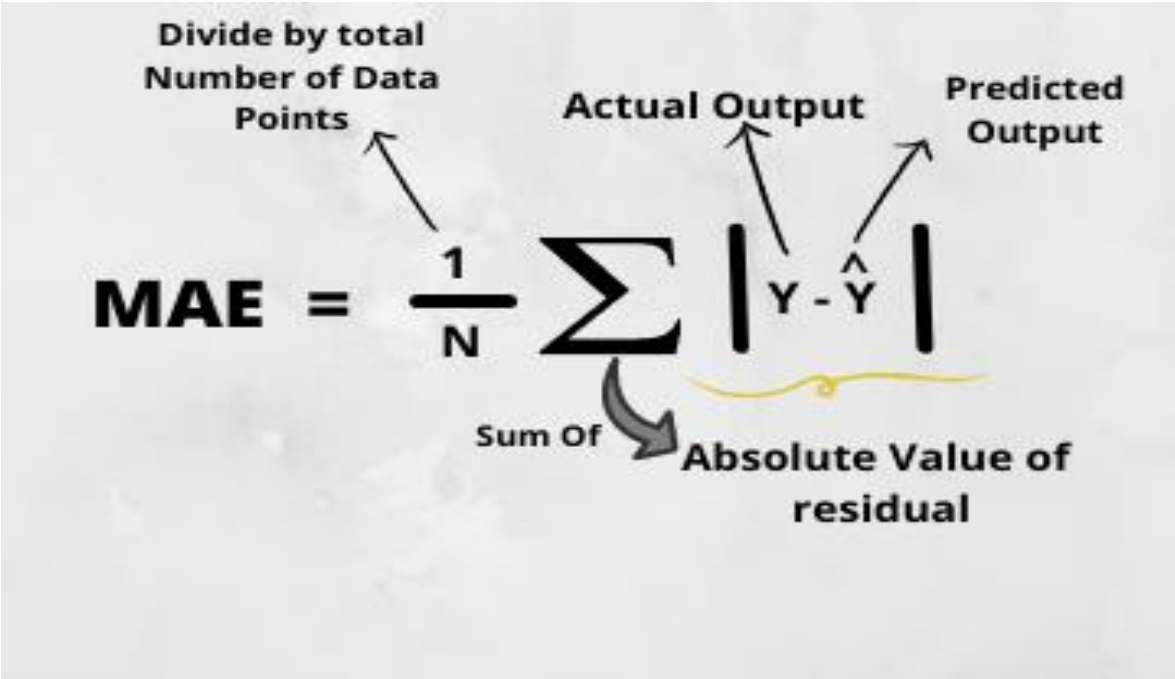
- **Mean Absolute Error**
- **Mean Squared Error**
- **R2 Score**
- **Adjusted R2**

Means Absolute Error (MAE)

- Mean Absolute Error or MAE is one of the simplest metrics, which measures the absolute difference between actual and predicted values, where absolute means taking a number as Positive.
- To understand MAE, let's take an example of Linear Regression, where the model draws a best fit line between dependent and independent variables.
- To measure the MAE or error in prediction, we need to calculate the difference between actual values and predicted values.
- But in order to find the absolute error for the complete dataset, we need to find the mean absolute of the complete dataset.

Means Absolute Error (MAE)

MAE calculation formula is:



Means Squared Error (MSE)

- Mean Squared error or MSE is one of the most suitable metrics for Regression evaluation.
- It measures the average of the Squared difference between predicted values and the actual value given by the model.
- Since in MSE, errors are squared, therefore it only assumes non-negative values, and it is usually positive and non-zero.
- Moreover, due to squared differences, it penalizes small errors also, and hence it leads to over-estimation of how bad the model is.
- MSE is a much-preferred metric compared to other regression metrics as it is differentiable and hence optimized better.

Means Squared Error (MSE)

MSE calculation formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

where:

- x_i represents the actual or observed value for the i-th data point.
- y_i represents the predicted value for the i-th data point.

R Squared Error

- R squared error is also known as Coefficient of Determination, which is another popular metric used for Regression model evaluation.
- The R-squared metric enables us to compare our model with a constant baseline to determine the performance of the model.
- To select the constant baseline, we need to take the mean of the data and draw the line at the mean.
- The R squared score will always be less than or equal to 1 without concerning if the values are too large or small.

R Squared Error

R-squared score is as follows:

$$R^2 = 1 - \frac{SSR}{SST}$$

Where:

- R^2 is the R-Squared.
- SSR represents the sum of squared residuals between the predicted values and actual values.
- SST represents the total sum of squares, which measures the total variance in the dependent variable.

R Squared Error

R-squared score is as follows:

$$R^2 = 1 - \frac{SSR}{SST}$$

Where:

- R^2 is the R-Squared.
- SSR represents the sum of squared residuals between the predicted values and actual values.
- SST represents the total sum of squares, which measures the total variance in the dependent variable.

Adjusted R Squared

- Adjusted R squared, as the name suggests, is the improved version of R squared error.
- R square has a limitation of improvement of a score on increasing the terms, even though the model is not improving, and it may mislead the data scientists.
- To overcome the issue of R square, adjusted R squared is used, which will always show a lower value than R^2.
- It is because it adjusts the values of increasing predictors and only shows improvement if there is a real improvement.

Adjusted R Squared

Adjusted R-squared is as follows:

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

Significance Test

- Significance tests, also known as hypothesis tests, are statistical techniques used to assess the validity of a hypothesis or to determine if observed results are statistically significant.
- In Statistics, tests of significance are the method of reaching a conclusion to reject or support the claims based on sample data.
- In the context of machine learning model evaluation, significance tests help us make informed decisions about the performance of our models and whether observed differences are meaningful or due to random chance.

Why Significance Test in Machine Learning ?

•In machine learning, we often compare different models, algorithms, or variations of a model to select the best one for a specific task.
•Significance tests help us answer questions like:

- Are the differences in accuracy between Model A and Model B statistically significant, or could they have occurred by random chance?
- Does a new model's performance improvement over an old model represent a real improvement, or is it merely a chance variation?

Common Significance Test & Metrics

T-Tests: T-tests are commonly used when comparing the means of two groups, such as comparing the performance metrics of two different machine learning models.

•Paired t-tests are used when the same subjects are used for both groups (e.g., comparing a model's performance before and after an improvement).

Common Significance Test & Metrics

P-Values : P-values indicate the probability of observing results as extreme as those obtained if the null hypothesis (usually stating no difference) were true.

•A low p-value (typically < 0.05) suggests that the observed differences are unlikely to have occurred by random chance, leading to the rejection of the null hypothesis.

•A high p-value suggests that the observed differences could reasonably occur due to random variation, leading to the acceptance of the null hypothesis.

Parul® University

NAAC A++ ACCREDITED UNIVERSITY

DIGITAL LEARNING CONTENT

Introduction to NLP

Thank You!!!

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversi