Parul® University
Vadodara, Gujarat

NAAC GRADE A++

Information and Communication Technology

Machine Learning
SUBJECT CODE: 303105353
Unit-5 Evaluation Metrics

Dr. Vinod Patidar
Associate Professor
Computer Science and Engineering

Contents

- ❑Evaluation Metrics
- ❑ROC Curves
- ❑Significance tests
- ❑Error correction in Perceptrons

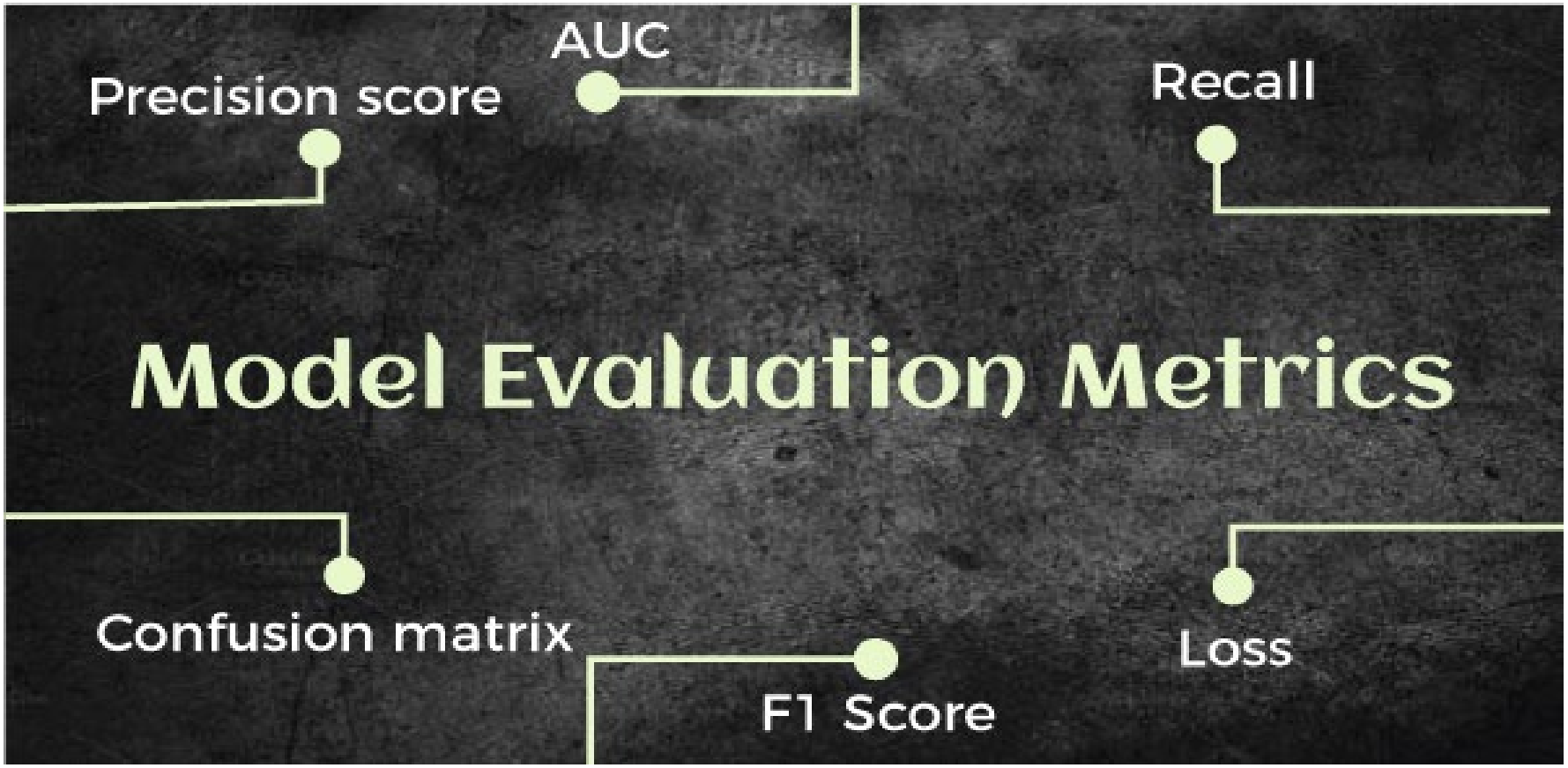
Evaluation Metrics

- Evaluating the performance of a Machine learning model is one of the important steps while building an effective ML model.
- ***To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics***
- These performance metrics help us understand how well our model has performed for the given data. In this way, we can improve the model's performance by tuning the hyper-parameters.

Evaluation Metrics

- Each ML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new dataset.
- In machine learning, each task or problem is divided into **classification** and **Regression**.
- Not all metrics can be used for all types of problems; hence, it is important to know and understand which metrics should be used.
- Different evaluation metrics are used for both Regression and Classification tasks.

Evaluation Metrics



Performance Metrics for Classification

- In a classification problem, the category or classes of data is identified based on training data.
- The model learns from the given dataset and then classifies the new data into classes or groups based on the training.
- It predicts class labels as the output, such as *Yes or No, 0 or 1, Spam or Not Spam,* etc.

Performance Metrics for Classification

To evaluate the performance of a classification model, different metrics are used, and some of them are as follows:

- **Accuracy**
- **Confusion Matrix**
- **Precision**
- **Recall**
- **F-Score**
- **AUC(Area Under the Curve)-ROC**

Accuracy

- The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.

- It can be formulated as:

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

Accuracy

When to Use Accuracy?

- It is good to use the Accuracy metric when the target variable classes in data are approximately balanced.
- For example, if 60% of classes in a fruit image dataset are of Apple, 40% are Mango.
- In this case, if the model is asked to predict whether the image is of Apple or Mango, it will give a prediction with 97% of accuracy.

Accuracy

When not to use Accuracy?

- It is recommended not to use the Accuracy measure when the target variable majorly belongs to one class.
- For example, Suppose there is a model for a disease prediction in which, out of 100 people, only five people have a disease, and 95 people don't have one.
- In this case, if our model predicts every person with no disease (which means a bad prediction), the Accuracy measure will be 95%, which is not correct.

Confusion Matrix in Machine Learning

- A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known.
- The confusion matrix is simple to implement, but the terminologies used in this matrix might be confusing for beginners

Confusion Matrix in Machine Learning

- A typical confusion matrix for a binary classifier looks like the below image(However, it can be extended to use for classifiers with more than two classes).

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

Confusion Matrix in Machine Learning

We can determine the following from the above matrix:

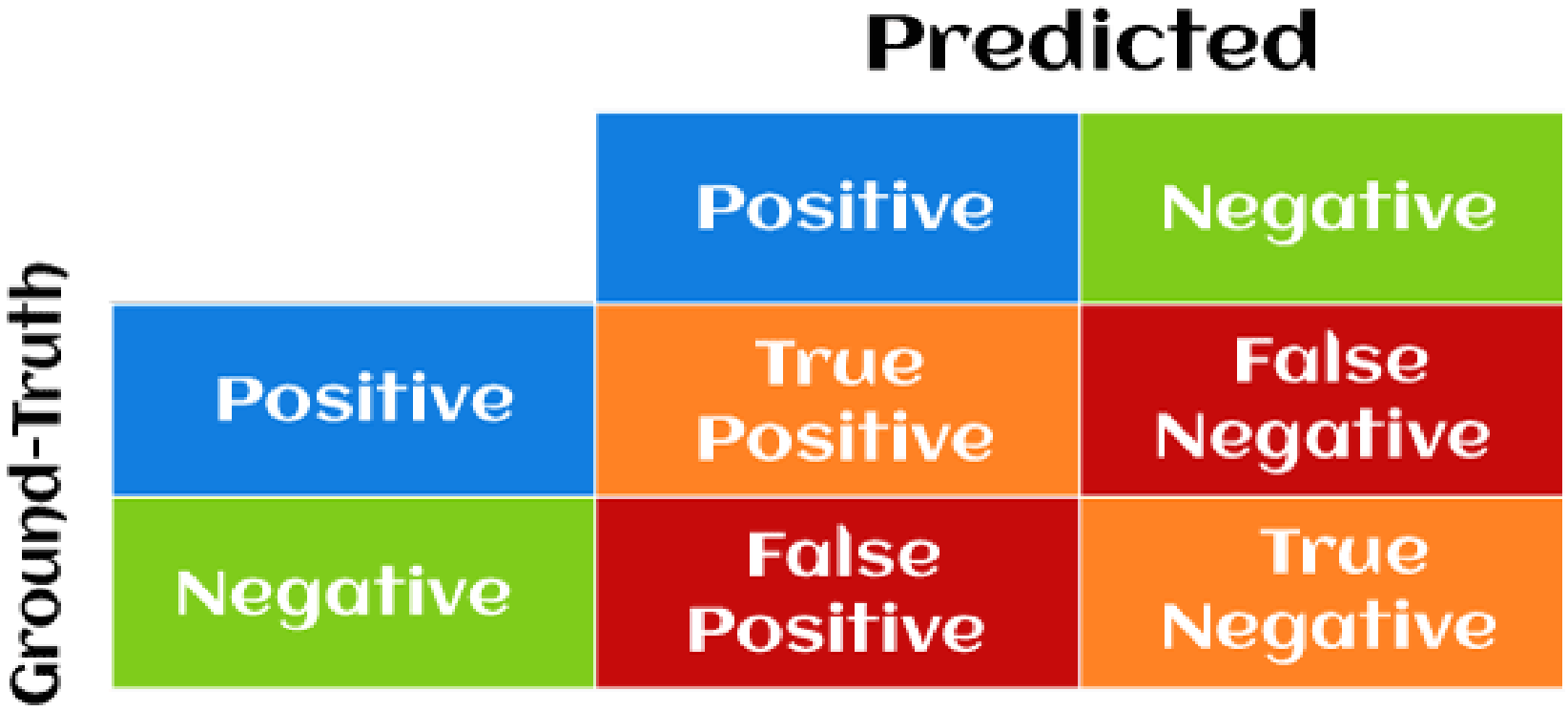
- In the matrix, columns are for the prediction values, and rows specify the Actual values. Here Actual and prediction give two possible classes, Yes or No. So, if we are predicting the presence of a disease in a patient, the Prediction column with Yes means, Patient has the disease, and for NO, the Patient doesn't have the disease.
- In this example, the total number of predictions are 165, out of which 110 time predicted yes, whereas 55 times predicted No.
- However, in reality, 60 cases in which patients don't have the disease, whereas 105 cases in which patients have the disease.

Confusion Matrix in Machine Learning

In general, the table is divided into four terminologies, which are as follows:

- **True Positive:** This combination tells us how many times a model correctly classifies a positive sample as Positive?
- **False Negative:** This combination tells us how many times a model incorrectly classifies a positive sample as Negative?
- **False Positive:** This combination tells us how many times a model incorrectly classifies a negative sample as Positive?
- **True Negative:** This combination tells us how many times a model correctly classifies a negative sample as Negative?

Confusion Matrix in Machine Learning



Need for Confusion Matrix in Machine learning

- It evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is.
- It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-II error.
- With the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy, precision, etc.

Precision

- The precision metric is used to overcome the limitation of Accuracy.
- Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

Precision = True Positive/True Positive + False Positive

Precision = TP/TP+FP

- **TP- True Positive**
- **FP- False Positive**

Precision

- The precision of a machine learning model will be low when the value of

  TP+FP (denominator) > TP (Numerator)

- The precision of the machine learning model will be high when Value of

  TP (Numerator) > TP+FP (denominator)

- **Hence, precision helps us to visualize the reliability of the machine learning model in classifying the model as positive.**

Examples to calculate the Precision in the machine learning model

Example 1- Let's understand the calculation of Recall

with four different cases where each case has the same

Recall as 0.667 but differs in the classification of negative samples. See how:

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

$$\text{Precision} = 2/2+1 = 2/3 = 0.667$$

Recall or Sensitivity

- It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly.
- It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

Recall or Sensitivity

- The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples.
- The recall measures the model's ability to detect positive samples.
- The higher the recall, the more positive samples detected.
- **Recall = True Positive/True Positive + False Negative**
- **Recall = TP/TP+FN**

Recall or Sensitivity

- TP- True Positive
- FN- False Negative
- Recall of a machine learning model will be low when the value of;
- TP+FN (denominator) > TP (Numerator)
- Recall of machine learning model will be high when Value of;
- TP (Numerator) > TP+FN (denominator)

Recall or Sensitivity

- TP- True Positive
- FN- False Negative
- Recall of a machine learning model will be low when the value of;
- TP+FN (denominator) > TP (Numerator)
- Recall of machine learning model will be high when Value of;
- TP (Numerator) > TP+FN (denominator)
- Unlike Precision, Recall is independent of the number of negative sample classifications. Further, if the model classifies all positive samples as positive, then Recall will be 1

Examples to calculate the Recall in the machine learning model

Below are some examples for calculating Recall in machine learning as follows

- **Example 1-** Let's understand the calculation of Recall with four different cases where each case has the same Recall as 0.667 but differs in the classification of negative samples. See how:

Recall or Sensitivity

- Recall = True Positive/True Positive + False Negative

Recall = TP/TP+FN

=2/(2+1)

=2/3

=0.667

Difference between Precision and Recall in Machine Learning

Precision	Recall
It helps us to measure the ability to classify positive samples in the model.	It helps us to measure how many positive samples were correctly classified by the ML model.
While calculating the Precision of a model, we should consider both Positive as well as Negative samples that are classified.	While calculating the Recall of a model, we only need all positive samples while all negative samples will be neglected.

Difference between Precision and Recall in Machine Learning

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.

Why use Precision and Recall in Machine Learning models?

- 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.

F-Scores

- F-score or F1 Score is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class.
- It is calculated with the help of Precision and Recall.
- It is a type of single score that represents both Precision and Recall. So, the **F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them**.

F-Scores

- The formula for calculating the F1 score is given below:

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

F-Scores

When to use F-Score?

- As F-score make use of both precision and recall, so it should be used if both of them are important for evaluation, but one (precision or recall) is slightly more important to consider than the other.
- For example, when False negatives are comparatively more important than false positives, or vice versa.

AUC-ROC Curve in Machine Learning

- In Machine Learning, only developing an ML model is not sufficient as we also need to see whether it is performing well or not.
- It means that after building an ML model, we need to evaluate and validate how good or bad it is, and for such cases, we use different Evaluation Metrics. *AUC-ROC curve is such an evaluation metric that is used to visualize the performance of a classification model*.
- It is one of the popular and important metrics for evaluating the performance of the classification model

What is AUC-ROC Curve?

AUC-ROC curve is a performance measurement metric of a classification model at different threshold values. Firstly, let's understand ROC (Receiver Operating Characteristic curve) curve.

ROC Curve :

- An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds.
- In the others words ROC or Receiver Operating Characteristic curve represents a probability graph to show the performance of a classification model at different threshold levels.

What is AUC-ROC Curve?

ROC Curve :

- The curve is plotted between two parameters, which are:

  - **True Positive Rate or TPR**
  - **False Positive Rate or FPR**

In the curve, TPR is plotted on Y-axis, whereas FPR is on the X-axis.

ROC Curve

TPR:

TPR or True Positive rate is a synonym for Recall, which can be calculated as:

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

ROC Curve

False Positive Rate (**FPR**) is defined as follows

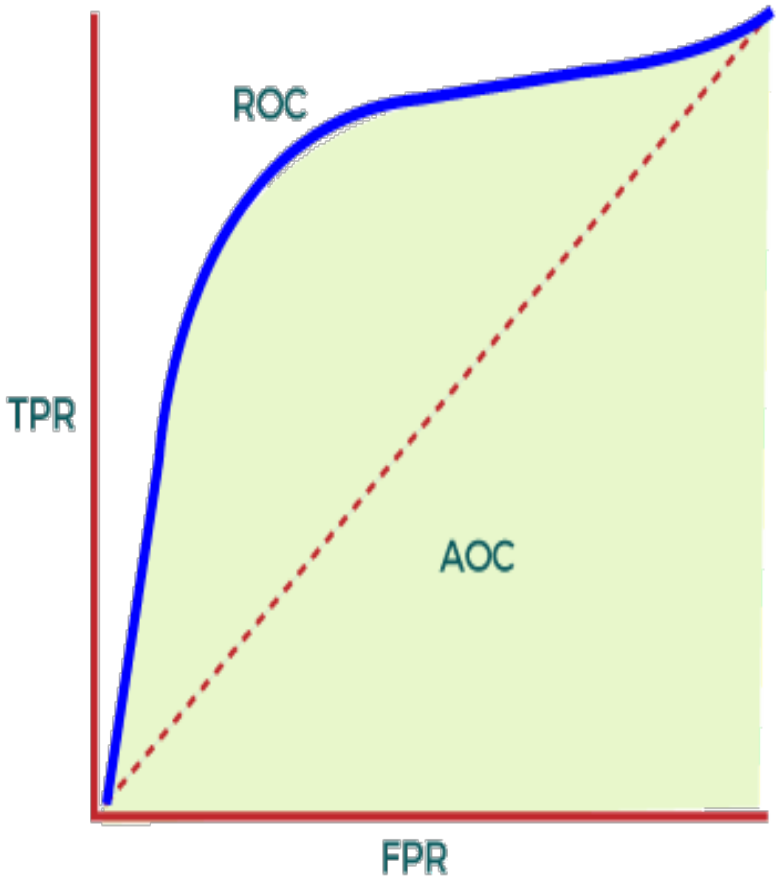
$$FPR = \frac{FP}{FP + TN}$$

Here, **TP**: True Positive
FP: False Positive
TN: True Negative
FN: False Negative
Now, to efficiently calculate the values at any threshold level, we need a method, which is AUC.

AUC: Area Under the ROC curve

AUC is known for **Area Under the ROC curve**.
As its name suggests, AUC calculates the two-dimensional area under the entire ROC curve ranging from (0,0) to (1,1), as shown below image:

AUC calculates the performance across all the thresholds and provides an aggregate measure. The value of AUC ranges from 0 to 1. It means a model with 100% wrong prediction will have an AUC of 0.0, whereas models with 100% correct predictions will have an AUC of 1.0.



AUC: Area Under the ROC curve

When to Use AUC-ROC

AUC is preferred due to the following cases:

- AUC is used to measure how well the predictions are ranked instead of giving their absolute values. Hence, we can say AUC is **Scale-Invariant.**
- It measures the quality of predictions of the model without considering the selected classification threshold. It means AUC is **classification-threshold-invariant.**

When not to use AUC-ROC

- AUC is not preferable when we need to calibrate probability output.
- Further, AUC is not a useful metric when there are wide disparities in the cost of false negatives vs false positives, and it is difficult to minimize one type of classification error.

Applications of AUC-ROC Curve

Although the AUC-ROC curve is used to evaluate a classification model, it is widely used for various applications. Some of the important applications of AUC-ROC are given below:

Classification of 3D model

The curve is used to classify a 3D model and separate it from the normal models. With the specified threshold level, the curve classifies the non-3D and separates out the 3D models.

Healthcare

The curve has various applications in the healthcare sector. It can be used to detect cancer disease in patients. It does this by using false positive and false negative rates, and accuracy depends on the threshold value used for the curve.

Binary Classification

AUC-ROC curve is mainly used for binary classification problems to evaluate their performance.

Classification Accuracy

It is one of the important parameters to determine the accuracy of the classification problems. It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. The formula is given below:

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

Misclassification rate:

It is also termed as Error rate, and it defines how often the model gives the wrong predictions. The value of error rate can be calculated as the number of incorrect predictions to all number of the predictions made by the classifier. The formula is given below:

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

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**

Mean 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.

Mean Absolute Error (MAE)

The below formula is used to calculate MAE:

$$MAE = 1/N \sum |Y - Y'|$$

Here,
Y is the Actual outcome,
Y' is the predicted outcome, and
N is the total number of data points.

Mean Squared Error

- 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.

Mean Squared Error

- The formula for calculating MSE is given below:

$$MSE = 1/N \sum (Y - Y')^2$$

Here,
Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points.

R Squared Score

- 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.

$$R^2 = 1 - \frac{MSE(Model)}{MSE(Baseline)}$$

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

We can calculate the adjusted R squared as follows:

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

Here,
n is the number of observations
k denotes the number of independent variables
and R_a^2 denotes the adjusted R^2

Significance Tests

- 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 Tests 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 Tests and 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 Tests and 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 GRADE A++

https://paruluniversity.ac.in/