**Unit-2**

**Q.1] Enlist and explain in short Performance Metrics for Classification Problems and Regression Problem.**

* **Performance Metrics for Classification Problems-**
* **Confusion Matrix:** The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data.
* **Classification Accuracy:** The accuracy is used to find the portion of correctly classified values. It is the sum of all true values divided by total values.
* **Classification Report**: To find the Classification Report we Need Precision, Recall or Sensitivity, F1 Score methods
* **AUC (Area Under ROC curve) :** When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model’s performance.
* **Performance Metrics for Regression Problems-**
* **Mean Absolute Error (MAE):** It is the simplest error metric used in regression problems.
* **Mean Square Error (MSE):** MSE is like the MAE, but the only difference is that the it squares the difference of actual and predicted output values before summing them all instead of using the absolute value.
* **R Squared (R2):** R Squared metric is generally used for explanatory purpose and provides an indication of the goodness or fit of a set of predicted output values to the actual output values.

**Q.2] Describe following Performance Metrics for Classification Problems in detail.**

|  |
| --- |
| https://miro.medium.com/max/700/1*sluS7bgqYOjd4835VegbJw.png |

**i. Classification Accuracy:** The accuracy is used to find the portion of correctly classified values. It is the sum of all true values divided by total values.

**ii. Classification Report- Precision, Recall or Sensitivity, F1 Score,**

|  |
| --- |
| https://miro.medium.com/max/700/1*x16UX1WoAOX4g3EVG8XQmQ.png |

* **Report- Precision:** Precision is used to calculate the model's ability to classify positive values correctly. It is the true positives divided by total number of predicted positive values.

|  |
| --- |
| https://miro.medium.com/max/700/1*rMQvdpGIDSzsmDI7P0gHAg.png |

* **TPR (True Positive Rate) / Recall or Sensitivity:** It is used to calculate the model's ability to predict positive values. It is the true positives divided by total number of actual positive values.

|  |
| --- |
| https://miro.medium.com/max/700/1*FL6nadrpcxBmFV5YVbp76w.png |

* **F1 Score:** It is the harmonic mean of Recall and Precision. It is useful when you need to take both Precision and Recall into account.

|  |
| --- |
| Understanding AUC - ROC Curve | by Sarang Narkhede | Towards Data Science |

**iii. AUC (Area Under ROC curve):**

* AUC - ROC curve is a performance measurement for the classification problems at various threshold settings.
* ROC is a probability curve and AUC represents the degree or measure of separability.
* It tells how much the model is capable of distinguishing between classes.
* Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.
* By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.
* The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

**Q.3] Describe following Performance Metrics for Regression Problems in detail.**

**i. Mean Absolute Error (MAE)**

* It is the simplest error metric used in regression problems.
* It is basically the sum of average of the absolute difference between the predicted and actual values.
* In simple words, with MAE, we can get an idea of how wrong the predictions were.
* MAE does not indicate the direction of the model i.e. no indication about underperformance or overperformance of the model.
* **Mean Absolute Error = (1/n) \* ∑|yi – xi|**

Where, yi: Actual value for the ith observation

xi: Calculated value for the ith observation

n: Total number of observations

**ii. Mean Square Error (MSE):**

* MSE is like the MAE, but the only difference is that the it squares the difference of actual and predicted output values before summing them all instead of using the absolute value.
* The difference can be noticed in the following equation
* **Mean Square Error = (1/n) \* ∑(yi – xi)**
* Where, yi: Actual value for the ith observation

xi: Calculated value for the ith observation

n: Total number of observations

**iii. R Squared (R2):**

* R Squared metric is generally used for explanatory purpose and provides an indication of the goodness or fit of a set of predicted output values to the actual output values.

|  |
| --- |
| R-Squared as ration of SSR and SST |

* The following formula will help us understanding it
* In the above equation, numerator is MSE and the denominator is the variance in 𝑌 values.
* We can use r2\_score function of sklearn.metrics to compute R squared value.