Model Evaluation Metrics in Machine Learning

Model Evaluation:

- Important component in Data Science lifecycle which occurs post Model Training.
- Assess performance of a model on a dataset.
- Determine how well the model is generalize to unseen data.
- To select best performing model.

Model Evaluation Metrics:

1. **Classification:**
    - **Accuracy**
    - **Precision (Positive prediction Value)**
    - **Recall (True Positive Rate or Sensitivity)**
    - **F1-Score**
    - **Confusion Matrix**
    - **AUC-ROC**

2. **Regression:**
    - **Mean Squared Error**
    - **Root Mean Squared Error**
    - **Mean Absolute Error**
    - **R-Squared Score (R2)**
    - **Adjusted R-Squared Score**

3. **Clustering:**
    - **Silhouette Coefficient**

Classification:

Confusion Matrix: A Table that summarizes model predictions against Actual values (ground truth)

Predictions **Actual Values**

	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Accuracy: Measures overall Correctness of the Model.

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN})}$$

Note: Useful for only Balanced Dataset.

Precision: Measures ability of Model to identify True Positives out of all positive predictions.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})}$$

1. It should be used when False Positives are costly.
2. It ensures our positive predictions are correct.

Example: Spam Detection

<u>Recall</u>: Measures ability of Model to identify True Positives out of all Actual Positives.

$$\boldsymbol{Recall = \frac{TP}{(TP + FN)}}$$

1. It should be used when False Negatives are costly.
2. It ensures to capture all the positive instances.

Example: Cancer Detection

<u>F1-Score</u>: Harmonic mean of Precision and Recall.

$$\boldsymbol{F1 - Score = \frac{2 \times precision \times recall}{(precision + recall)}}$$

1. Useful for imbalanced data.
2. Should use in applications where Precision and Recall both are important.

Example: Fraud Detection

<u>AUC – ROC</u> (Area Under Receiver Operating Characteristic Curve):

AUC - Degree or Measure of Separability.

ROC - Probability Curve.

1. Measure's ability of a model to distinguish between positive and Negative instances.
2. It is a plot between True Positive rate (TPR) and False Positive Rate (FPR)
3. Useful for imbalance dataset.

4. Values ranges from 0 to 1. Where 0 is worse and 1 is best.
5. 0.5 means Model has no class separation capability.

Example: Medical Diagnosis.

Regression:

Mean Squared Error (MSE): Average of squared differences between prediction and actual values.

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}(y - y\_pred)^2}$$

1. It is useful when we want to penalize large errors more heavily.
2. Smaller the MSE, Better the fit.

Example: Stock Price prediction

Root Mean Squared Error (RMSE): Square root of Average of squared differences between prediction and actual values.

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - y\_pred)^2}}$$

1. It is useful when we want to evaluate model performance in terms of units of Target variable.
2. Easy to interpret.

Example: Plant Height (It tells avg difference in cms between prediction and Actual heights.)

Mean Absolute Error (MAE): Average of absolute differences between prediction and actual values.

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}|y - y\_pred|}$$

1. It is useful when we want to know average error in terms of Target variable units.
2. Less sensitive to Outliers.

Example: House Price prediction.

R-Squared Score (R2): It measures the proportion of variance in the target variable that can be explained by the model.

$$\boldsymbol{R2 = 1 - \frac{ss\_\mathrm{res}}{ss\_tot}}$$

$ss\_res = \sum_{i=1}^{n}(y - y\_pred)^2$ $\quad$ $ss\_tot = \sum_{i=1}^{n}(y - y\_mean)^2$

1. It determines overall performance of the model which ranges from 0 to 1.
2. Near to 1, Better the Model.

Problems with R2:

1. R2 Score doesn't have to do with correlation between independent and dependent features. It simply increases whenever we add new feature to the Model.
2. Because, The Linear Regression tries to assign coefficients in such a way that ss_res always decreases.

Adjusted R2:

1. It is same as R2 Score, but it also considers number of independent variables are used to predict target variable.
2. By this we can determine whether adding new variables to the model actually increases the model fit or not.

$$\boldsymbol{Adjusted\ R2 = 1 - \frac{(1 - R2)(N - 1)}{(N - P - 1)}}$$

- N - Total number of datapoints.
- P - Number of independent features.
- R2 - R2_Score determined by the model.

It penalizes when features are not correlated to dependent variable.

$$\boldsymbol{Adjusted\_R2 \ \leq R2}$$

Clustering:

Silhouette Coefficient: It measures the similarity of an object to its own cluster compared to other clusters.

$$\boldsymbol{Silhouette\ Score} = \frac{(\boldsymbol{b} - \boldsymbol{a})}{\boldsymbol{max}(\boldsymbol{a}, \boldsymbol{b})}$$

- a – average intra-cluster distance (avg distance between each point within a cluster)
- b - average inter-cluster distance (avg distance between all the clusters)

1. It ranges from -1 to 1.
    - 1 - Clusters are well apart from each other.
    - 0 - Distance between clusters are insignificant.
    - -1 - Clusters assigned in wrong way.
2. It helps to evaluate quality of clusters and determine optimal number of clusters.

*** Thank You ***