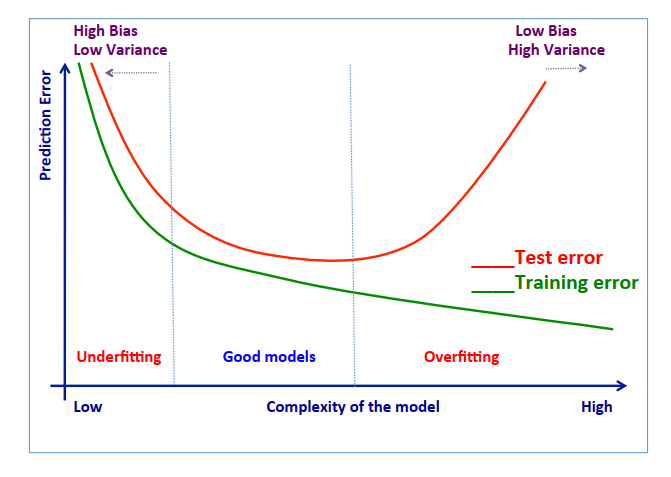
**AI Final Exam**

**Exam Date: 25-Feb-2024 9:00am**

**Submitted Date: 25-Feb-2024 12:00pm**

Theory (10pts)

1. Regarding the figure below, what is Underfitting, Good Models, Overfitting
2. Explain in detail what is Training Set, Validation Set, and Testing Set?
3. In case of Overfitting existing, how to fix it and please explain in details?
4. Below Confusion Matrix, please calculate the value of these term and explain:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | Orange | Banana |
| Prediction | Orange | 45TN | 5FP |
| Banana | 20FN | 60TP |

* 1. Accuracy
  2. Precision
  3. Recall
  4. specificity
  5. FI measure

**Answer**

**Name: Prum Savat**

**15E Weekend 29/32**

**1.** Regarding the figure below, what is Underfitting, Good Models, Overfitting

* **Underfitting**:
  + Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in the data.
  + Models that underfit perform poorly not only on the training data but also on unseen data because they fail to learn the relationships between the features and the target variable.
  + In terms of performance metrics, underfit models often have high bias and low variance.
  + Underfitting can be addressed by using more complex models, increasing the number of features, or reducing regularization.
* **Good Models**:
  + Good models strike a balance between underfitting and overfitting, capturing the underlying patterns in the data without memorizing noise.
  + Good models typically exhibit low bias and moderate variance.
  + Achieving a good model often involves careful selection of model architecture, appropriate feature engineering, and tuning hyperparameters.
* **Overfitting**:
  + Overfitting occurs when a machine learning model learns the noise and random fluctuations in the training data rather than the underlying patterns.
  + Models that overfit perform well on the training data but generalize poorly to unseen data.
  + Overfit models often have low bias and high variance.

**2.** Explain in detail what is Training Set, Validation Set, and Testing Set?

* **Training Set** is a set of examples used for learning a model (e.g., a classification model).
* **Validation Set** is a set of examples that cannot be used for learning the model but can help tune model parameters (e.g., selecting K in K-NN). Validation helps control overfitting.
* **Testing Set** is used to assess the performance of the final model and provide an estimation of the test error.

**3.** In case of Overfitting existing, how to fix it and please explain in detail?

* **Cross-Validation**:
  + Cross-validation is a technique used to assess how well a model will generalize to an independent dataset. It involves splitting the dataset into multiple subsets. Cross-validation helps in detecting overfitting by providing a more robust estimate of the model's performance.
* **Regularization**:
  + Regularization is a technique used to prevent overfitting by adding a penalty term to the model's loss function. Two common types of regularization are:
    - L1 Regularization (Lasso): Adds the absolute values of the coefficients to the loss function, leading to sparsity in the model.
    - L2 Regularization (Ridge): Adds the squared values of the coefficients to the loss function, penalizing large coefficients.
* **Feature Selection**:
  + Feature selection involves selecting a subset of the most informative features from the dataset while discarding irrelevant or redundant features.
  + Techniques for feature selection include univariate feature selection, recursive feature elimination, and feature importance ranking based on model coefficients or tree-based algorithms.
* **Early Stopping**:
  + Early stopping is a technique used during the training process to prevent the model from overfitting by stopping the training process when the performance on a validation set starts to deteriorate. It involves monitoring the model's performance on a separate validation set and halting training when the performance stops improving or begins to worsen.
* **Ensemble Methods**:
  + Ensemble methods combine multiple base models to make predictions, thereby reducing the risk of overfitting and improving generalization. Popular ensemble methods include:
    - Bagging (Bootstrap Aggregating): Constructs multiple models using bootstrap samples of the training data and averages their predictions.
    - Boosting: Builds a sequence of models, each focusing on correcting the errors of its predecessor.
    - Random Forest: A type of ensemble method based on decision trees, where multiple trees are trained on random subsets of the data and features.
    - Gradient Boosting Machines (GBM): A boosting technique that builds an ensemble of weak learners, such as decision trees, in a sequential manner.

**4.** Below Confusion Matrix, please calculate the value of these term and explain:

**A. Accuracy**

* Accuracy measures the proportion of correctly classified samples out of the total number of samples.
* It is calculated as: Accuracy = TP+TN / FP+FN+TP+TN​

We have TP = 60, TN = 45, FP = 5, FN = 20

Accuracy = 60 + 45 / 5 + 20 + 60 + 45 = 105 / 130 = 0.8076 or 80%

**B. Precision**

* Precision measures the proportion of correctly predicted positive cases out of all cases predicted as positive.
* It is calculated as: Precision= TP / TP + FP
* Precision = 60 / 60 + 5 = 0.9230 or 92%

**C. Recall (Sensitivity)**:

* Recall measures the proportion of correctly predicted positive cases out of all actual positive cases.
* It is calculated as: Recall = TP / TP + FN = 60 / 60+20 = 0.75 OR 75%

**D. Specificity**:

* Specificity measures the proportion of correctly predicted negative cases out of all actual negative cases.
* It is calculated as: Specificity = TN / TN +FP = 45 / 45 + 5 = 0.90 OR 90%

**E. F1 measure**:

* The F1 score is the harmonic means of precision and recall, providing a balance between the two metrics.
* It is calculated as:

F1 measure = 2 \* Precision \* Recall / Precision + Recall

F1 measure = 2 \* 0.9230 \* 0.75 / 0.9230 + 0.75 = 1.3845 / 0.6922 = 2.0001