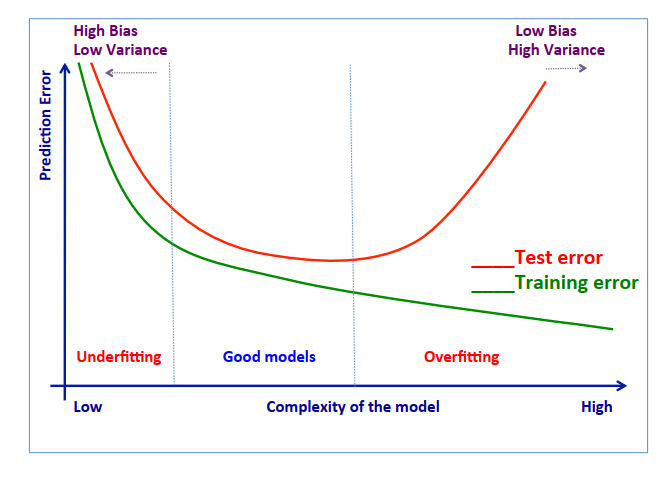
**AI Final Exam**

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

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

Phak Mengheang

Theory (10pts)

1. Explain in detail what is Training Set, Validation Set, and Testing Set?
2. Regarding the figure below, what is Underfitting, Good Models, Overfitting?
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 | 45 | 5 |
| Banana | 20 | 60 |

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

Answer

* Training Set : This is the largest portion of your data (usually 70-80%). It's used to train the model's parameters. The model "learns" the patterns and relationships within this data.
* Validation Set : This is a smaller portion of your data (usually 10-20%). It's used to fine-tune the model's hyperparameters (e.g., learning rate, number of hidden layers). These settings control how the model learns from the training data.
* Testing Set: This is the smallest portion of your data (usually 10-20%). It should be completely unseen by the model during training and validation. It's used to provide an unbiased assessment of the model's performance on new, unseen data. This is the most crucial measure of how well the model generalizes to real-world scenarios.

1. Underfitting, Good Models, and Overfitting: The x-axis represents the model complexity, which can be influenced by factors like the number of features used or the number of layers in a neural network. The y-axis shows the prediction error, measured by metrics like mean squared error or classification accuracy.

* Underfitting: This occurs when the model is too simple. It cannot capture the underlying patterns in the data, resulting in high training and test errors (represented by the leftmost part of the curve).
* Good Models: These models achieve a balance between complexity and accuracy. They fit the training data well without overfitting to noise or irrelevant patterns. This sweet spot is represented by the lowest point on the curve.
* Overfitting: This occurs when the model is too complex. It memorizes the training data too well, including noise and irrelevant details. This leads to a low training error but a high test error (represented by the rightmost part of the curve). Imagine a student who memorizes every detail of their practice tests without truly understanding the concepts, leading to good performance on practice tests but poor performance on the final exam with different questions. By using training, validation, and testing sets effectively, you can identify and avoid underfitting and overfitting, ultimately building better models that generalize well to unseen data.