CHAPTER 1

The Training/Validation/Testing Process

In machine learning, it's essential to assess the performance of a model accurately. This assessment helps us choose the best model and tune its parameters. The data used to develop machine learning models is typically divided into three sets: training, validation, and testing.

1.0.1 Training Set

The training set is used to train the model, i.e., to adjust the model's weights and biases in the case of neural networks, or to determine the best split in decision trees, among other things. The model learns from this data, which is why it's called the "training" set.

1.0.2 Validation Set

The validation set is used to tune model parameters (hyperparameters), to choose the model architecture (for example, the number of hidden layers in a neural network), or to determine the degree of the polynomial in polynomial regression, among other uses. This set provides an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.

1.0.3 Testing Set

The testing set is used to provide an unbiased evaluation of the final model fit on the training dataset. The test set serves as a proxy for real-world data that the model has not seen before. It's important to only use the test set once, after all training and validation is complete, to avoid "leaking" information from the test set into the model.

The key to this process is that each set of data is separate and independent. Mixing data between the sets can lead to overly optimistic or pessimistic assessments of a model's performance.

In practice, the division of data into these three sets can be done randomly (often with 70%

for training, 15% for validation, and 15% for testing), or using more structured methods like cross-validation, depending on the amount and nature of the available data.

Appendix A

Answers to Exercises