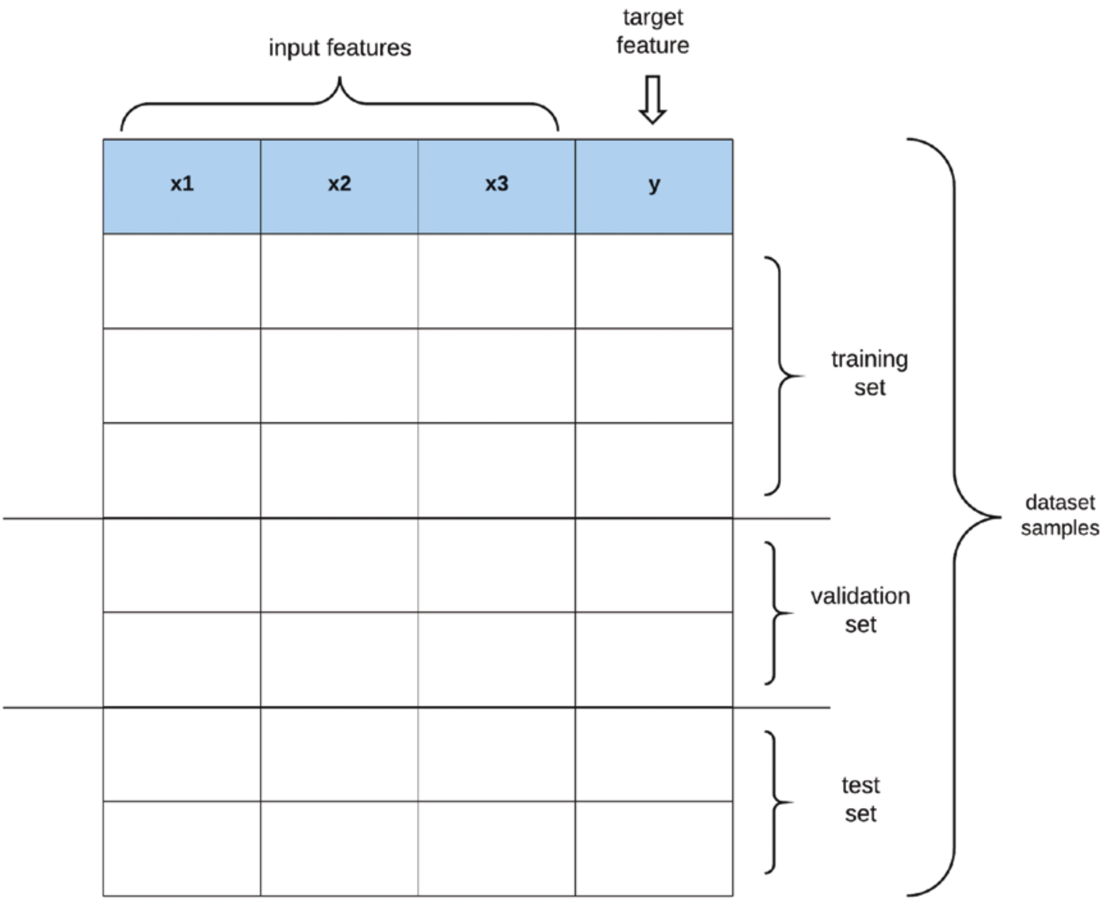

Figure 14-4. *Training, test, and validation set*

Bias vs. Variance Trade-Off

The concept of bias vs. variance is central to machine learning and is critical to understanding how the model is performing, as well as in suggesting the direction in which to improve the model.

A model is said to have high **bias** when it oversimplifies the learning problem or when the model fails to accurately capture the complex relationships that exist between the input features of the dataset. High bias makes the model unable to generalize to new examples.

High variance, on the other hand, is when the model learns too closely the intricate patterns of the dataset input features, and in the process, it *learns the irreducible noise of the dataset samples*. When the learning algorithm learns very closely the patterns of the training samples, including the noise, it will fail to generalize when exposed to previously unseen data.

Hence, we observe that there is a need to strike the right balance between bias and variance, and often it is down to the skill of the model builder to discover this middle ground. However, there exists practical rules of thumb for finding the right trade-off between bias and variance.

How Do We Recognize the Presence of Bias or Variance in the Results?

High bias is observed when the model performs poorly on the trained data. Of course, it will also perform poorly (or even worse) on the test data with high prediction errors. When high bias occurs, it can be said that the model has underfit the data. High variance is observed when the trained model learns the training data very well but performs poorly on unseen (test) data. In the event of high variance, we can say that the model has overfit the dataset.

The graph in Figure 14-5 illustrates the effect of bias and variance on the quality/performance of a machine learning model. In Figure 14-6, the reader will observe that there is a sweet spot somewhere in the middle where the model has good performances on both the training and the test datasets.

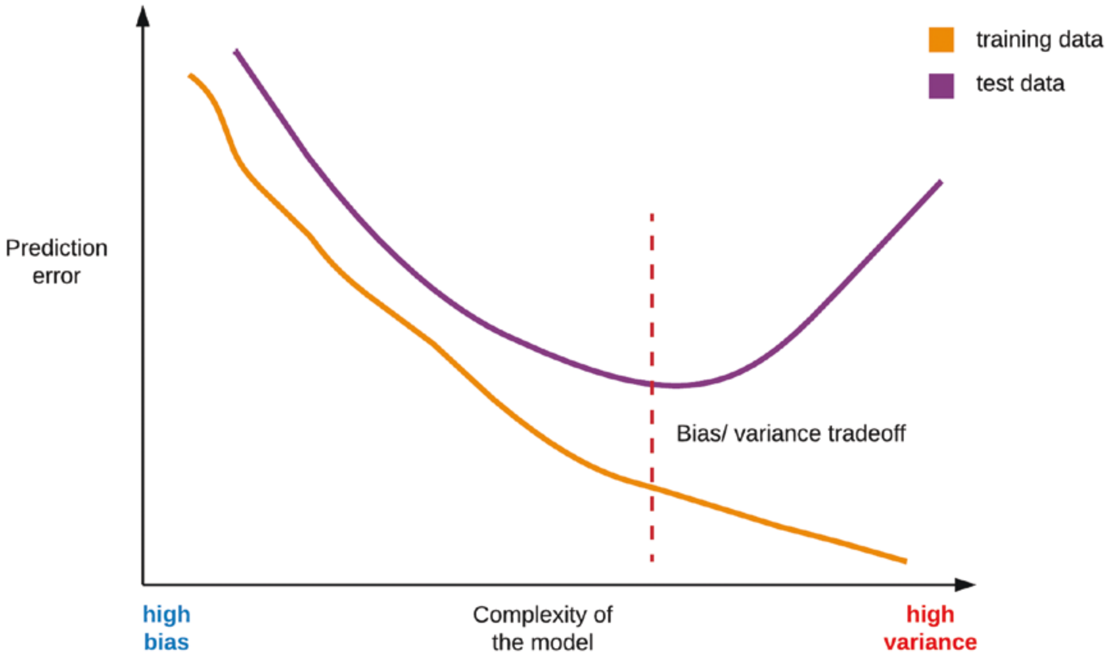

Figure 14-5. *Bias and variance*

To recap, our goal is to have a model that strikes a balance between high bias and high variance. Figure 14-6 provides further illustration on the effects of models with high bias and variance on a dataset. As seen in the image to the left of Figure 14-6, we want to have a model that can generalize to previously unseen example, such a model should have good prediction accuracy.

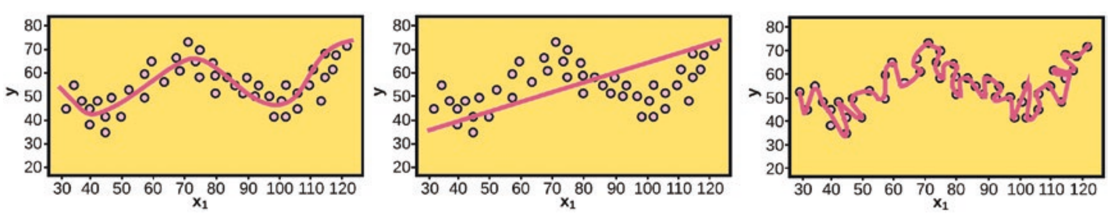

Figure 14-6. ***Left:*** *Good fit.* ***Center:*** *Underfit (high bias).* ***Right:*** *Overfit (high variance)*

Evaluating Model Quality

Evaluation metrics give us a way to quantitatively evaluate how well our model is performing. The model's performance on the training data is evaluated to get the training set accuracy, while its performance on the test data is evaluated to get the test data accuracy when the model predicts the targets of previously unseen examples. Evaluation on test data helps us to know the true performance measure of our model.

The learning problem determines the type of evaluation metric to use. As an example, for regression prediction problems, it is common to use the root mean squared error (RMSE) to evaluate the magnitude of the error made by the model. For classification problems, one of the common evaluation metrics is to use a confusion matrix to get a picture of how many samples are correctly classified or misclassified. From the confusion matrix, it is possible to derive other useful metrics for evaluating classification problems such as accuracy, precision, and recall.

The following are the evaluation metrics for machine learning that we will consider in this text:

Classification

- Confusion matrix
- Area under ROC curve (AUC-ROC)

Regression

- Root mean squared error (RMSE)
- R-squared (R^2)

Let's go through each.

Classification Evaluation Metrics

In this section, we'll briefly explain performance metrics for classification machine learning tasks.

Confusion Matrix

The confusion matrix is a popular evaluation metric for gleaning insights into the performance of a classification supervised machine learning model. It is represented as a table with grid-like cells. In the case of a two-class classification problem, the columns