Cross Validation

The confusion table is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

Cross Validation

The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications and often produces a poorer outcome. The cross validation is often termed a jack-knife classification, in that it successively classifies **all cases but one** to develop a discriminant function and then categorizes the case that was left out. This process is repeated with each case left out in turn.This is known as leave-1-out cross validation.

This cross validation produces a more reliable function. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

Validation and Testing

Validation

- When you have sufficient data, you can subdivide your data into three parts called the training, validation, and test data.
- Rather than estimating parameter values from the entire data set, the data set is broken into three distinct parts. During the **variable selection** process, models are fit on the training data, and the prediction error for the models so obtained is found by using the validation data.
- Validation is the process of using part of a data set to estimate model parameters, and using the other part to assess the predictive ability of the model.
- Validation can be used to assess whether or not overfitting has occurred.

This prediction error on the validation data can be used to decide when to terminate the selection process or to decide what effects to include as the variable selection process proceeds. Finally, once a selected model has been obtained, the test set can be used to assess how the selected model generalizes on data that played no role in selecting the model.

1. The training set is the part that estimates model parameters.
2. The validation set is the part that assesses or validates the predictive ability of the model.
3. The test set is a final, independent assessment of the models predictive ability.

Model Validation

- A validation set is a portion of a data set used to assess the performance of prediction or classification models that have been fit on a separate portion of the same data set (the training set).
- Typically both the training and validation set are randomly selected, and the validation set is used as a more objective measure of the performance of various models that have been fit to the the training data (and whose performance with the training set is therefore not likely to be a good guide to their performance with data that they were not fit to).

Model Validation

It is difficult to give a general rule on how many observations you should assign to each role. One important textbook recommended that a typical split might be 50% for training and 25% each for validation and testing.