OVERVIEW OF DATA SAMPLING AND QUANTIZATION

After adding a data source to an Azure ML experiment, you generally need to work with the data fields that that the data source contains. Data flow in an Azure ML experiment involves passing a table (encapsulated in a data frame object in R and Python) between modules, and working with the columns in that table.

Many data columns contain numeric values that indicate some sort of quantifiable metric, such as a count, size, weight, or other measurement. Measurements such as these can be discrete values or values within a continuous scale. When creating a predictive model, it can be useful to quantize continuous numeric variables into discrete values, for example grouping values for a continuous variable like temperature into three discrete values that represent *cold*, *moderate*, and *hot*.

Occasionally, columns contain values that are not measurements, but representations of categories. For example, a *Direction* column might contain the values 1, 2, 3, and 4; representing North, East, South, and West. Machine learning models work best when the metadata for your dataset indicates that these columns should be treated as categorical values and not quantities.

This chapter explores some techniques for quantizing continuous variables and managing categorical values.