Intermediate Machine Learning

# Dealing with Missing Values

Most machine learning libraries and models will produce an error if data containing missing values is used within them. As such, it is important to consider how missing values should be dealt with.

## Drop Columns with Missing Values

The simplest approach is to simply drop columns containing missing values. However, if only a small proportion of rows in a given column contain missing values, then a lot of information is lost with this approach.

## Imputation

Imputation fills in missing values with some specified value such as the mean value of the column. This won’t be exactly right in most cases but will generally lead to more accurate models than simply dropping columns containing null values.

## Extension to Imputation

In some cases, it can be beneficial to record whether a column value was missing and include this within the model.

# Categorical Variables

A categorical variable is discrete in that it can only take on a finite number of values. Inputting these variables into machine learning models within Python without pre-processing will result in an error, and so these must be dealt with prior to being put into a model.

## Drop Categorical Variables

The easiest approach is to simply not include categorical variables within Machine Learning models. If this approach is to be used, then the categorical columns should not include any useful information.

## Label Encoding

Label encoding can be used to assign each unique value to an integer, which effectively converts categorical variables to a numeric datatype. This type of approach works best with ordinal variables, in that the categories represents some attribute that has a defined order (for example frequency, or class of ticket).

## One-hot Encoding

One-hot encoding creates new columns within the data which indicate the presence or absence of each unique value within the categorical column. This does not assume that the categorical data within the original column has any kind of order. Categorical variables that do not have an intrinsic ranking are known as nominal variables. This approach works best when the categorical data does not take on a large number of unique values.