Preface

Data preparation may be the most important part of a machine learning project. It is the most time consuming part, although it seems to be the least discussed topic. Data preparation, sometimes referred to as data preprocessing, is the act of transforming raw data into a form that is appropriate for modeling. Machine learning algorithms require input data to be numbers, and most algorithm implementations maintain this expectation. As such, if your data contains data types and values that are not numbers, such as labels, you will need to change the data into numbers. Further, specific machine learning algorithms have expectations regarding the data types, scale, probability distribution, and relationships between input variables, and you may need to change the data to meet these expectations.

The philosophy of data preparation is to discover how to best expose the unknown underlying structure of the problem to the learning algorithms. This often requires an iterative path of experimentation through a suite of different data preparation techniques in order to discover what works well or best. The vast majority of the machine learning algorithms you may use on a project are years to decades old. The implementation and application of the algorithms are well understood. So much so that they are routine, with amazing fully featured open-source machine learning libraries like scikit-learn in Python. The thing that is different from project to project is the data. You may be the first person (ever!) to use a specific dataset as the basis for a predictive modeling project. As such, the preparation of the data in order to best present it to the problem of the learning algorithms is the primary task of any modern machine learning project.

The challenge of data preparation is that each dataset is unique and different. Datasets differ in the number of variables (tens, hundreds, thousands, or more), the types of the variables (numeric, nominal, ordinal, boolean), the scale of the variables, the drift in the values over time, and more. As such, this makes discussing data preparation a challenge. Either specific case studies are used, or focus is put on the general methods that can be used across projects. The result is that neither approach is explored. I wrote this book to address the lack of solid advice on data preparation for predictive modeling machine learning projects. I structured the book around the main data preparation activities and designed the tutorials around the most important and widely used data preparation techniques, with a focus on how to use them in the general case so that you can directly copy and paste the code examples into your own projects and get started.

Data preparation is important to machine learning, and I believe that if it is taught at the right level for practitioners, it can be a fascinating, fun, directly applicable, and immeasurably useful toolbox of techniques. I hope that you agree.

Jason Brownlee
2020