Chapter 7

Modern Statistical Modeling in Machine Learning and Big Data Analytics:

Statistical Models for Continuous and Categorical Variables

Niloofar Ramezani
George Mason University, USA

ABSTRACT

Machine learning, big data, and high dimensional data are the topics we hear about frequently these days, and some even call them the wave of the future. Therefore, it is important to use appropriate statistical models, which have been established for many years, and their efficiency has already been evaluated to contribute into advancing machine learning, which is a relatively newer field of study. Different algorithms that can be used within machine learning, depending on the nature of the variables, are discussed, and appropriate statistical techniques for modeling them are presented in this chapter.

INTRODUCTION

Machine learning is an important topic these days as it involves a set of many different methods and algorithms that are suited to answer diverse questions about a business or problem. Therefore, choosing an algorithm is a critical step in the machine learning process to ensure it truly fits the solution proposed in answering a problem at hand (Segal, 2004). To better understand machine learning algorithms and when each algorithm needs to be used, it is helpful to understand them within the framework of statistics and separate them into two main groups based on the data and the format of their outcomes. These two types of machine learning methods are classification and regression for categorical and continuous response variables, respectively. Within this book chapter, we will differentiate these two types and mention related algorithms and statistical techniques that can be used to answer real world problems.

DOI: 10.4018/978-1-6684-6291-1.ch007

Then the concept of high-dimensional data and some of the methods that are appropriate for handling such data will be discussed.

An outline of the sections and subsections within this chapter are discussed below. First, the continuous approaches are discussed and regression algorithm and regression algorithm, as the most commonly used approach for such scenarios, is explained. Such methods help modeling the continuous response variables. On the other hand, generalized linear models and classification techniques assist researchers to model discrete, binary, and categorical responses. While discussing the statistical models, which are appropriate in predicting the categorical response variables, binary logistic regression as well as multinomial and ordinal logistic regression models are introduced and discussed here. These models can assist researchers and applied practitioners in modeling binary and categorical response variables. Within the section that discusses ordinal logistic models, different logit models are briefly discussed. These logit functions are cumulative logit, adjacent-categories logit, and continuation-ratio logit.

The next section of this chapter is dedicated to introducing and discussing dimension reduction methods. Dimension reduction techniques applied within the two classification and regression algorithms, paired with the analytically powerful computers, will assist researchers in handling data sets with higher dimensions of variables and observations efficiently within a reasonable timeline. Different models to address each set of techniques are introduced in this chapter to provide different statistical and research tools that can be used by researchers and practitioners in machine learning and data analysis. Selected dimension reduction methods are discussed in this chapter. These subsections are principal components analysis, decision trees and their use in building random forest, regression decision trees and random forest for continuous responses, classification decision trees and random forest for categorical responses, LASSO and ridge regression, and finally cluster analysis.

This chapter adopts an introductory and educational approach, rather than rushing into the programming or data analysis details, to familiarize the readers with different data scenarios and the proper statistical and machine learning algorithms to appropriately model the data. The author believes the first step in ensuring the quality, correctness, and reliability of any quantitative analysis is to learn about the characteristics of the data, by exploring them, and then choosing the proper statistical and machine learning approach to model the data. Years of teaching statistics, data analytics, and quantitative methodology, as well as providing statistical consultation, have taught the author the importance of strengthening the foundations of statistical knowledge of students, clients, or data analysts before diving into running codes and printing output to ensure the accuracy of the results. This chapter is meant to educate the readers about the correct modeling options to minimize the chances of choosing the wrong methods of data analysis while exploring and modeling real-world data.

CONTINUOUS DATA

When dealing with continuous data and predicting such outcome measures, regression approaches and algorithms can be used. Linear regression is by far the simplest and most popular example of a regression algorithm used in many fields. Though it is often underrated due to its simplicity, it is a flexible prediction tool. Regression trees and support vector regressions are more advanced algorithms within the framework of regression that can be used for high dimensions of data.