Supervised statistical classification-Multinomial

Suraj Kumar

03/12/2021

Introduction

Background

Supervised multinomial classification is an essential aspect of Machine Learning. Sectors such as Business analysts, Data scientists, Financial analysts, and medicine are required to predict the data category based on feature instance vectors. The simple multinomial regression model is the most fundamental statistical method to serve the purpose; however, several advanced classification models such as RandomForest, and Neural Network have evolved over the years, which have brought a revolution in prediction accuracy. RandomForest has a very versatile application based on a decision tree hierarchy. On the other side, Neural network performance has improved over a few decades as more and more data are available for training. Performance comparison is a crucial aspect of choosing the best Statistical model for data. Overall classification accuracy and group-specific classification accuracy are the two most prominent used performance measures for supervised multinomial classification.

Aim of Analysis

The research project aims to evaluate the best multinomial classification models based on some performance measure criteria. Specifically, the main objectives are to design a study to access performance in classification, measure the performance of a simple multinomial regression model, and compare the performance to that of other statistical methods.

Workflow and Data Descriptions

The study goes through exploratory and formal data analysis of 4 datasets; Abalone, Car evaluation, Nursery school application, and Contraceptive method choice. All of the datasets have multi-class categorical output and a set of covariates. The aim of the analysis is specific for each dataset; predicting the age of abalone from physical measurements for Abalone dataset; predicting the car acceptability for given features of cars for Car dataset; predicting the contraceptive method choice of women based on several characteristics about women for Contraceptive method choice dataset; predicting success or failure of nursery class application provided socio-demographic information of the parents for Nursery dataset. The datasets have been sampled at different locations and sectors through a randomized method. The abalone dataset has been sampled from the Abalone population in Tasmania. I. Blacklip and Abalone found in North Coast and Islands of Bass Strait; the car evaluation dataset has been obtained from a simple hierarchical decision model previously developed for the demonstration of DEX; the contraceptive method choice dataset has been collected from a part of the 1987 National Indonesia Contraceptive Prevalence Survey; the nursery dataset has been obtained from the applications of nursery schools in Ljubljana, Slovenia. This report focuses on box plots, summaries, scatterplots, histograms, bar plots, and proportional tables to develop initial impressions about the data in section . While the study deepens on designing a performance measure, fitting multinomial regression,

Random forest, and Neural network to each dataset, and comparing the outcomes in section . describes all the methods used to achieve the goals. Eventually, the section details the remark extracted from the overall analysis, and section discusses the possible future extension in the work.

Description of the Methods

Multinomial Regression

Nominal logistic regression models for more than two categories can be used as a classification tool for observation to a class based on the highest predicted probability among groups. The statistical method considers one of the groups as a baseline and fits logistic regression to the ratio of each group member to the baseline. Eventually, the probability of observation falling into each class is evaluated, and a class is assigned based on the highest value.

Variable Selection

Wrapper methods are the greedy search approach that stepwise removes variables based on an evaluation criterion. A bi-directional elimination method works similar to forwarding selection but does extra backward elimination at each iteration of adding a variable. The process reaches optimal features until no new feature can be added or removed. Due to its greedy nature, different wrapper methods can give different results. Additionally, they are prone to over-fitting and have high computation time.

Principle Component Analysis

PCA is a robust unsupervised machine learning algorithm for feature extraction. The method is useful when a significant correlation is present among continuous covariates. Moreover, it holds no assumption for the distribution of data. PCA considers both covariance and correlation matrix depending upon the evenness of the order of data variance. The algorithm is badly affected by outliers; therefore, it's necessary to remove outliers firstly. Eventually, discretionary components are retained based on the proportion of variance(proportion desired), Cattell's scree plot method, or Kaiser's method.

RandomForest

RandomForest, likewise bagging, bootstrap samples from the training data with replacement and average across all out-of-bag errors to give overall miss-classification error rate. Additionally, the former method attempts to decorrelate decision trees by randomly selecting a subset of features, which helps reduce the variance. RandomForest beautifully handles missing data, mixed datasets, and multicollinearity.

Neural Network

The neural network has garnered the attention of the whole world in every sector as more and more data are available for training. The method is very effective in modeling the complex non-linear relationships among data features. The hyperparameter for the hidden layer needs to be carefully adjusted for every dataset to avoid overfitting and to model appropriately. Feedforward neural networks are most common in applications. It attempts to find a locally optimal solution based on initial reference using gradient descent and backpropagation.