Trends Over Time In Common Statistical Methods

Sophie Berube

9/28/2017

Abstract

Introduction

Over the past decade, machine learning algorithms have become increasingly accessible to the scientific community as a means of data analysis. Through the use of various coding languages such as R and Python, scientists have sucessfully implemented machine learning algorithms in a variety of applied fields like neuroscience, genetics, proteomics and cell and molecular biology. Under the umbrella of machine learning advances, various other tools have been developed by statisticians, mathematicians and computer scientist and are beginning to make a significant impact in the larger scientific community. Coined by Hinton in 2006, deep learning has begun to emerge as one of these modern methods of analysis, meeting the scientific community's ever growing needs demands for tools that can handle increasingly complex and large datasets. Similarly, artifical neural networks seek to harness the properties of the nervous system to analyze a variety of data and rely on similar machine learning related concepts to attack various problems.

The following analysis uses the articles in the various journals of the Public Libray of Science to assess whether modern statistical methods like machine learning, deep learning and neural networks are surpassing traditional statistical methods like ANOVA, linear regression, bootstraps and bayesian methods in popularity across applied scientific fields.

The Public Library of Science or PLOS began publishing articles in 2003 in PLOS Biology and has since added eight journals to the repertoire in a variety of fields including medicine, computational biology, genetics, clinical trials, pathogens and tropical diseases. PLOS ONE, the largest of the PLOS journals, publishes more articles annually than any other current scientific journal. An open access, pay to publih model ensures that articles are peer rieviewed to verify that scientific standards are met but submissions are not turned down based on percieved lack of importance or impact to a particular field's advancement the way they would be in many prestigious mainstream jourals like Nature or Cell. This makes PLOS journals a fairly adequate representation of average scientific research over the past decade, and thus an appropriate database to assess statistical methodologies used by the scientific community over the past decade.

The R package `rplos` which accesses articles from all nine currently published PLOS journals since their respective inceptions, was used in this analysis.

Methods

First, a list of ten key words representing both modern and traditional data analysis techniques was assembled by looking through a variety of plos articles and based on fundamental tecnhniques discussed in most basic statistics textbooks. The five words meant to represent traditional statistical analysis methods were ANOVA, bayesian, linear regression, mixed effect model, generalized estimating equation and bootstrap. Similarly four words meant to represent modern statistical methods are machine learning, deep learning, support vector machine, artifical neural network.

In order to visualize their distribution over all publihed articles in PLOS over time, the function `plot_overtime` from the `rplos` package was used. The limit for each word was set to 10,000 articles which means that 10,000 articles mentioning each of the key words were sampled and were attributed to a particular month based on their date of publication. Counts for various months were plotted from the period 2003 to October of 2017 and a loess curve was fit to these plots.

Quasipoisson models were fit for each key word with a count of articles mentioning a particular key word as the outcome and time points as the regressor, in order to identify possible differences in trends over time. More specifically, using visual inspection from the loess plots, the data were split up into a timeperiod of increasing counts for all terms and a timeperiod of decreasing or stable counts for all terms. For each key word, a different quasipoisson model was fit for both increasing and decreasing counts periods and the coefficients for the time points were compared both within key words and across key words.

Since the `plot_throughttime` function inclues any articles that mention the key words in the counts for a particular timepoint, a false discovery rate was estimated to give an idea of how many articles might mention a key word without it being used as a statistical analysis method in the paper. This rate was estimated using the `highplos` function which returns the API (or unique identifier) of articles that contain a particular key word. For each key word or phrase, ten of these articles were visually inspected to determine whether or not that particular statistical method was used in the analysis.

Results