Boffin Access

Open Access Journal | Volume: 1.3

Advances in Computer Sciences

ISSN 2517-5718

Computer-Intensive Statistics: A Promising Interplay between Statistics and Computer Science

Sunil K Sapra | California State University, Los Angeles, California, United States

Editorial

Statistics and computer science have grown as separate disciplines with little interaction for the past several decades. This however, has changed radically in recent years with the availability of massive and complex datasets in medicine, social media, and physical sciences. The statistical techniques developed for regular datasets simply cannot be scaled to meet the challenges of big data, notably the computational and statistical curses of dimensionality. The dire need to meet the challenges of big data has led to the development of statistical learning, machine learning and deep learning techniques. Rapid improvements in the speed and lower costs of statistical computation in recent years have freed statistical theory from its two serious limitations: the widespread assumption that the data follow the bell-shaped curve and exclusive focus on measures, such as mean, standard deviation, and correlation whose properties could be analyzed mathematically [1]. Computer-intensive statistical techniques have freed practical applications from the constraints of mathematical tractability and today can deal with most problems without the restrictive assumption of Gaussian distribution. These methods can be classified into frequentist and Bayesian methods. The former methods utilize the sample information only while the latter methods utilize both the sample and prior information.

Frequentist statistical methods have benefitted enormously from the interaction of statistics with computer science. A very popular computer-intensive method is the bootstrap for estimating the statistical accuracy of a measure, such as correlation in a single sample. The procedure involves generating a very large number of samples with replacement from the original sample. Bootstrap as a measure of statistical accuracy has been shown to be extremely reliable in theoretical research [2,3]. Another widely used computer-intensive method for measuring the accuracy of statistical methods is cross validation. It works non-parametrically without the need for probabilistic modelling and measures the mean-squared-error for the test sample using the training sample to evaluate the performance of various machine learning methods for selecting the best method. Other frequentist statistical methods that rely on a powerful computing environment include jackknife for estimating bias and variance of an estimator, classification and regression trees for prediction, generalized linear models for parametric modelling with continuous, discrete or count response [4], generalized additive models for flexible semi-parametric regression modeling [5], the LASSO method for Cox proportional hazard regression in high dimensional settings [6], and EM algorithm [7] for finding iteratively the maximum likelihood or maximum a posteriori (MAP) estimates of parameters in complex statistical models with latent variables, alternating between performing an expectation (E) step, which evaluates the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. Bagging, random forests, and boosting [8,9] are some relatively recent developments in machine learning which use large amounts of data to fit a very rich class of functions to the data almost automatically. These methods represent a fitted model as a sum of regression trees. A regression tree by itself is a fairly weak prediction model, so these methods greatly improve prediction performance by constructing ensembles of either deep trees under random forests or shallow trees under boosting. Support vector machine (SVM) [10], an approach for linear and nonlinear classification developed in computer science, has been found to perform very well and is widely used by statisticians and data scientists. Neural networks are a class of learning methods developed separately in statistics and artificial intelligence, which use a computer-based model of human brain to perform complex tasks. These methods have found applications across several disciplines, including medicine, geosciences, hydrology, engineering, business, and economics. Some common statistical models, such as multiple linear regression, logistic regression, and linear discriminant analysis for classifying binary response are akin to neural networks. The main idea underlying these methods is to extract linear combinations of inputs as derived features and model the output as a nonlinear function of these features called the activation function.

Bayesian statistical methods have also benefited greatly from computer-intensive methods, notably the Markov Chain Monte Carlo (MCMC) approach [11], which is a class

Article Information

DOI: 10.31021/acs.20181113
Article Type: Editorial Article
Journal Type: Open Access
Volume: 1 **Issue:** 3
Manuscript ID: ACS-1-113
Publisher: Boffin Access Limited

Received Date: 04 September 2018
Accepted Date: 01 October 2018
Published Date: 04 October 2018

***Corresponding author:**

Sapra Sk
California State University, Los Angeles, CA
90032, United States
Tel: +1- 323 343 2941
E-mail: ssapra@calstatela.edu

Citation: Sapra SK. Computer-Intensive Statistics: A Promising Interplay between Statistics and Computer Science. Adv Comput Sci. 2018 Oct;1(3):113.

of methods for sampling from a probability distribution. Chief among these methods are the Gibbs sampling procedure designed for sampling from the conditional distribution and the Metropolis-Hastings algorithm designed for sampling from the posterior distribution as well as for multidimensional integration problems [12,13]. These methods construct a Markov chain that has the desired distribution as its equilibrium distribution and obtain a sample of the desired distribution by observing the chain after several steps. Development of MCMC methods has been a key step in making it possible to compute large hierarchical models that require integrations over hundreds or even thousands of unknown parameters.

Over the past six decades, the center of statistics has steadily moved away from its traditional foothold in mathematics and logic towards a more computational focus [14]. Some may argue that with the popularity of the new field of data science, emphasis has shifted from statistical inference to algorithmic thinking. Impressive, ambitious predictive algorithms, such as random forests, boosting, support vector machines, and neural networks all illustrate algorithmic thinking of data scientists as opposed to traditional statistical thinking. Surprisingly, probability models, the main building blocks of statistical inference are largely missing from the development of these algorithms. Development of these predictive algorithms without parametric probability models has been made possible by computer-intensive non-parametric techniques, such as bootstrap, cross-validation, and permutation. Regrettably, related fields of Econometrics, Sociometry, and Psychometrics have been too slow to embrace these developments and remain reluctant to use these successful algorithmic methods absent statistical inference justification for them.

While interaction between computer science and statistics is growing rapidly with the development of algorithmic models, such as bagging, boosting, random forests, and support vector machines, we need to understand why these techniques work by developing theoretical justification and inference for these methods. The field of statistics has been energized by computer science with the rise of data science as a discipline. It is fair to expect that the field of computer science is likely to be enriched by statistics as well through the development of techniques for analysis of massive data sets and probability models to understand why black box algorithmic models, such as neural networks work for pattern recognition and other predictive tasks. All scientific and business disciplines are likely to benefit from the interplay between computer science and statistics as demonstrated by the development of fast, powerful and versatile applications, such as CART® and Bayesia Lab for advanced analytics employed by applied researchers across many different fields.

References

1. Persi D, Efron B. Computer-intensive Methods in Statistics. Scientific American. 1983;248(3):116-131.
2. Efron B. The jackknife, the bootstrap, and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics. 1982.
3. Efron B. Bootstrap methods: Another look at the jackknife. Annals of Statistics. 1979;7(1):1-26.
4. McCullagh P, Nelder J. Generalized Linear Models. London: Chapman and Hall; 1989.
5. Hastie T, Tibshirani R. Generalized Additive Models. London: Chapman and Hall; 1990.
6. Tibshirani R. The Lasso method for variable selection in the Cox model. Statistics in Medicine. 1997;16(4):385-395.
7. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data *via* the EM algorithm. J R Stat Soc. Series B. 1977;39(1):1-38.
8. Breiman L. Bagging Predictors. Mach Learn. 1996;24(2):123-140.
9. Breiman L. Random Forests. Mach Learn. 2001;45(1):5-32.
10. Cortes, Corinna, Vapnik, Vladimir N. Support-vector networks. Mach Learn. 1995:20(3):273-297.
11. Markov Chain Monte Carlo, Wikipedia.
12. George C, Edward IG. Explaining the Gibbs sampler. Am Stat. 1992;46(3):167-174.
13. Gelfand AE, Smith AFM. Sampling-Based Approaches to Calculating Marginal Densities. J Am Stat Assoc. 1990;85:398-409.
14. Efron B, Tibshirani R. Encyclopedia of Statistical Sciences. New York, John Wiley & Sons Inc; 2004. Computer-Intensive Statistical Methods; p. 1-9.