ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

The impact of statistics on Data Science to quantify uncertainty

Bhavani Govardhan[1], Dr.Gandhavalli Sambasiva Rao[2], Palaparthi Seethalakshmi[3]

Abstract

In this paper, we substantiate our premise that statistics is one of the most important disciplines to provide tools and methods to find structure in and to give deeper insight into data, and the most important discipline to analyze and quantify uncertainty. We give an overview over different proposed structures of Data Science and address the impact of statistics on such steps as data acquisition and enrichment, data exploration, data analysis and modeling, validation and representation and reporting. Also, we indicate fallacies when neglecting statistical reasoning.

Keywords : Structures of data science · Impact of statistics on data science · Fallacies in data science.

Introduction and premise

Data Science, as a field, is strongly influenced by multiple scientific disciplines, including informatics, computer science, mathematics, operations research, and statistics, along with various applied sciences. The term "Data Science" first appeared in the title of a statistical conference in 1996, but its public perception often emphasizes the roles of computer science and business applications, especially in the era of Big Data.

The conceptual shift in statistics can be traced back to the 1970s when John Tukey's ideas redirected the focus from purely mathematical statistical testing to understanding data before hypothesis formation, i.e., exploring data. Another foundational aspect of Data Science is rooted in Knowledge Discovery in Databases (KDD) and its subfield Data Mining, which brings together diverse approaches such as inductive learning, Bayesian statistics, query optimization, expert systems, information theory, and fuzzy sets.

The contemporary notion of Data Science integrates these ideas, resulting in various definitions. Cao's comprehensive formulation for Data Science encompasses multiple disciplines including statistics, informatics, computing, communication, sociology, and management, all acting in conjunction with data, the environment, and a data-to-knowledge-to-wisdom thought process.

[1] Assistant Professor, Department of CSE, RISE Krishna Sai Gandhi Group of Institutions, Ongole,
[2] Professor, Department of CSE, RISE Krishna Sai Gandhi Group of Institutions, Ongole,
[3] Assistant Professor, Department of CSE, RISE Krishna Sai Gandhi Group of Institutions, Ongole.

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

A comprehensive overview of Data Science by Donoho in 2015 traces its evolution from statistics. There have been radical views suggesting a renaming of statistics to Data Science, highlighting the centrality of statistics and machine learning in this domain, as emphasized by ASA leaders in 2015.

In this context, statistical methods are essential in numerous key stages of Data Science. Therefore, our primary assertion is that statistics plays a pivotal role in providing tools and methodologies to uncover patterns and gain deeper insights from data, serving as the most crucial discipline in analyzing and quantifying uncertainty.

This paper intends to delve into the significant influence of statistics on the critical steps involved in the practice of Data Science.

Steps in Data Science

Two fundamental models have influenced the structural framework of Data Science: CRISP-DM (Cross Industry Standard Process for Data Mining) and an expanded perspective. CRISP-DM, organized into six key steps—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—is now integral to applied statistics.

In our conceptualization, inspired by CRISP-DM, the steps in Data Science have evolved. Our definition presents Data Science as a sequence of steps: Data Acquisition and Enrichment, Data Storage and Access, Data Exploration, Data Analysis and Modeling, Optimization of Algorithms, Model Validation and Selection, Representation and Reporting of Results, and Business Deployment of Results. Steps denoted in small capitals in Table 1 represent areas where statistics plays a relatively lesser role.

Steps	CRISP-DM	Cao's Definition	Proposed Framework
Business Understanding	Business Understanding	Domain-specific Data Applications and Problems	Business Deployment of Results
Data Understanding	Data Understanding	Data Storage and Management	Data Acquisition and Enrichment
Data Preparation	Data Preparation	Data Quality Enhancement	Data Storage and Access
Modeling	Modeling	Data Modeling and Representation	Data Exploration
Evaluation	Evaluation	Deep Analytics, Learning, and Discovery	Data Analysis and Modeling
Deployment	Deployment	Simulation and Experiment Design	Optimization of Algorithms
		High-performance Processing and Analytics	Model Validation and Selection
		Networking, Communication	Representation and Reporting of Results
		Data-to-Decision and Actions	

Typically, these steps are not linear but rather iterative in a cyclical loop. Furthermore, it is common to alternate between two or more steps. This especially applies to Data Acquisition and Enrichment, Data Exploration, and Statistical Data Analysis, as well as Statistical Data Analysis and Modeling, and Model Validation and Selection.

The table presents a comparison of different definitions of steps in Data Science. The relationships between terms are represented by horizontal blocks. The absence of the Data Acquisition and Enrichment step in CRISP-DM indicates that the scheme primarily deals with observational data. Additionally, we propose to expand CRISP-DM by including Data Storage and Access and Optimization of Algorithms, where statistics plays a relatively lesser role.

The steps for Data Science can be expanded further, as seen in Cao's work, which adds elements such as Domain-specific Data Applications and Problems, Data Storage and Management, Data Quality Enhancement, and more.

While Cao’s formulation aligns with our proposal in principle, it provides more detailed descriptions. For instance, our step 'Data Analysis and Modeling' corresponds to 'Data Modeling and Representation, Deep Analytics, Learning, and Discovery' in Cao’s work. It is important to note that the vocabulary might slightly differ depending on whether the background is in computer science or statistics. For instance, 'Experiment Design' in Cao’s definition refers to the design of simulation experiments.

In the subsequent sections, we will emphasize the role of statistics in all the steps where it is significantly involved. This corresponds to all steps in our proposal in Table 1, except for the steps denoted in small capitals.

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

Entries Data Storage and Access and Optimization of Algorithms are mainly covered by informatics and computer science.

1.Data acquisition and enrichment

Experimental design, commonly known as Design of Experiments (DOE), is a critical component for systematically generating data, especially when the impact of extraneous factors must be identified. Controlled experiments serve as the bedrock of robust process engineering, ensuring the production of dependable products despite variations in process variables. While controllable factors inherently carry some uncontrollable variations influencing the response, certain factors, such as environmental influences, remain beyond control. Employing DOE is essential for managing the impact of such noisy factors.

DOE finds application in various areas such as the systematic generation of new data for data acquisition, the systematic reduction of databases, and the optimization of algorithm parameters to enhance data analysis methods. Simulations also play a role in generating new data, while imputation of missing data serves as a valuable tool for enriching databases and filling data gaps.

These statistical techniques for data generation and enrichment should form an integral part of the foundation of Data Science. Relying solely on observational data without controlling for noise significantly undermines the quality of data analysis results and may lead to erroneous interpretations. The notion of "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" seems flawed, as the presence of noise in the data necessitates the continued importance of experimental design.

Therefore, the integration of experimental design is indispensable for ensuring the reliability, validity, and replicability of our research findings.

2 Data exploration

Exploratory statistics plays a crucial role in data preprocessing by enabling an understanding of the underlying content within a database. The practice of exploring and visualizing observed data finds its roots in the pioneering work of John Tukey. Over time, the process of comprehending and transforming data, which constitutes the most labor-intensive aspect of data analysis, has emerged as a significant component of statistical science.

In the realm of Data Science, data exploration and data mining are pivotal for ensuring the appropriate application of analytical methods. Among the most significant contributions of statistics is the concept of distribution. This concept not only facilitates the representation of variability within the data but also aids in the incorporation of a priori knowledge of parameters. This foundational idea serves as a critical underpinning for various statistical analyses.

3 Statistical data analysis

Finding structure in data and making predictions are the most important steps in Data Science. Here, in particular, statistical methods are essential since they are able to handle many different analytical tasks. Important examples of statistical data analysis methods are the following.

Hypothesis testing

Hypothesis formulation stands as a fundamental pillar of statistical analysis. Often, queries arising in data-driven predicaments can be translated into hypotheses, serving as the bridge connecting underlying theories to statistical analyses. Statistical hypotheses, closely intertwined with statistical tests, enable the scrutiny of questions and theories based on available data. However, the recurrent use of the same data in various tests necessitates the adjustment of significance levels. In the realm of applied statistics, the proper handling of multiple testing assumes paramount importance. This is particularly evident in domains such as pharmaceutical studies. Disregarding such corrective techniques would likely result in an inflated number of statistically significant findings, surpassing what is truly warranted.

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

Classification

In data analysis, methods for identifying and predicting subpopulations play a fundamental role. In the unsupervised scenario, these subpopulations are identified from a dataset without any prior knowledge about their existence, often referred to as clustering. Conversely, in the supervised setting, the objective is to derive classification rules from labeled datasets for predicting unknown labels based on influential factors.

In contemporary times, a wide array of methods exists for both unsupervised and supervised cases. However, the era of Big Data demands a fresh perspective on classical techniques, as the computational effort of complex analysis methods often grows more than linearly with the number of observations (n) or features (p). This situation becomes challenging when dealing with extensive datasets, resulting in prolonged computation times and numerical challenges. Consequently, there is a renewed focus on simpler optimization algorithms with lower time-complexity, alongside a reevaluation of traditional statistical and machine learning methods to address the challenges posed by Big Data.

Regression

Methods serve as the primary tool for uncovering global and local relationships between features when the target variable is under measurement. The choice of approach typically hinges on the distributional assumptions concerning the underlying data. In instances where the data adhere to the normality assumption, linear regression stands as the most prevalent method, while distributions from the exponential family often necessitate the application of generalized linear regression. More sophisticated techniques encompass functional regression for functional data , quantile regression, and regression based on alternative loss functions, such as Lasso regression, which deviates from the typical squared error loss.

When confronted with the challenges posed by Big Data, similar considerations to those in classification methods arise, primarily due to the substantial number of observations (n), as seen in data streams, and/or the presence of a high number of features (p). To tackle the reduction of n, expedited computation is made possible through data reduction techniques like compressed sensing, random projection methods , or sampling-based procedures. Likewise, for the reduction of p to the most influential features, variable selection or shrinkage approaches such as the Lasso can be utilized, thereby preserving the interpretability of the features. Additionally, (sparse) principal component analysis serves as another viable option.

Time series analysis

Forecasting and understanding temporal structures constitute essential tasks in data analysis. Time series data are prevalent in various studies involving observational data, with prediction emerging as a critical challenge in this context. Fields such as behavioral sciences, economics, natural sciences, and engineering commonly employ time series analysis. For instance, in signal analysis, such as the study of speech or music data, statistical methods encompass the examination of models within the time and frequency domains. The primary objective often revolves around predicting future values of the time series itself or its properties. For example, modeling the vibrato of an audio time series allows for a realistic prediction of future tones , while rules learned from past time periods can predict the fundamental frequency of a musical tone .

In econometrics, the analysis often involves multiple time series and their co-integration. Time series analysis also finds common application in technical domains, particularly in process control.

4 Statistical modeling.

a) Complex interactions between factors can be modeled by graphs or networks. Here, an interaction between two factors is modeled by a connection in the graph or network. The graphs can be undirected as, e.g., in Gaussian graphical models, or directed as, e.g., in Bayesian networks. The main goal in network analysis is deriving the network structure. Sometimes, it is necessary to separate (unmix) subpopulation specific network topologies .

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

(b) Stochastic differential and difference equations can represent models from the natural and engineering sciences. The finding of approximate statistical models solving such equations can lead to valuable insights for, e.g., the statistical control of such processes, e.g., in mechanical engineering . Such methods can build bridge between the applied sciences and Data Science.

(c) Local models and globalization Typically, statistical models are only valid in sub-regions of the domain of the involved variables. Then, local models can be used. The analysis of structural breaks can be basic to identify the regions for local modeling in time series. Also, the analysis of concept drifts can be used to investigate model changes over time. In time series, there are often hierarchies of more and more global structures. For example, in music, a basic local structure is given by the notes and more and more global ones by bars, motifs, phrases, parts etc. In order to find global properties of a time series, properties of the local models can be combined to more global characteristics. Mixture models can also be used for the generalization of local to global models. Model combination is essential for the characterization of real relationships since standard mathematical models are often much too simple to be valid for heterogeneous data or bigger regions of interest.

5 Model validation and model selection

In scenarios where multiple models are proposed, statistical tests for model comparison serve a crucial role in evaluating and structuring these models, particularly with respect to their predictive capabilities. The assessment of predictive power commonly relies on resampling methods, wherein the distribution of power characteristics is studied by systematically altering the subpopulation used for model learning. Characteristics derived from these distributions are then employed for model selection.

Perturbation experiments offer an alternative means of evaluating model performance by assessing the stability of different models in the face of noise. Additionally, meta-analysis and model averaging represent methodologies for assessing combined models.

Over the years, model selection has gained increasing importance, especially considering the rapid proliferation of classification and regression models within the academic literature.

6 Representation and reporting

In statistical analyses, visualization plays a crucial role in interpreting discovered structures, while storing models in a readily updatable format is essential for effective communication of results and ensuring the safe deployment of data analysis. The deployment phase holds significant importance in Data Science as it serves as the final step in the CRISP-DM framework and underlies the crucial data-to-decision and action step outlined by Cao.

In addition to visualization and appropriate model storage, a primary task for statistics involves the reporting of uncertainties and undergoing thorough reviews.

Fallacies

The statistical methodologies delineated in Section 2 play a foundational role in identifying underlying structures within data, facilitating a more comprehensive understanding, and thus contributing to the efficacy of data analysis. Disregarding contemporary statistical practices or resorting to simplistic data analytics/statistical methods may lead to avoidable inaccuracies, especially when dealing with complex and voluminous datasets.

As highlighted towards the end of Section 2.2, the concept of distribution stands as a pivotal contribution of statistics. Overlooking the role of distributions in data exploration and modeling confines us to reporting mere values and parameter estimates without accounting for their inherent variability. It is only through a grasp of distributions that we can make predictions along with their corresponding error margins.

Furthermore, distributions form the bedrock of model-based data analytics. For instance, unsupervised learning aids in the discovery of data clusters. When additional structural dependencies such as space or time come into play, inferring

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

parameters like cluster radii and their spatio-temporal evolution becomes crucial. Such model-based analyses heavily rely on an understanding of distributions, as demonstrated in applications such as the study of protein clusters.

When multiple parameters warrant attention, it is advisable to compare univariate hypothesis testing approaches with multiple procedures, particularly in scenarios like multiple regression, and select the most appropriate model through variable selection. Confining analyses to univariate testing would overlook the relationships that exist between variables.

Gaining deeper insights into data often necessitates the utilization of more complex models, such as mixture models, for identifying heterogeneous groups within the data. Disregarding such complexities often results in a meaningless average, calling for the need to discern subgroups through the separation of components. Within a Bayesian framework, this is facilitated by latent allocation variables in a Dirichlet mixture model, as demonstrated in applications involving the decomposition of a mixture of various networks in a heterogeneous cell population within molecular biology.

A mixture model can represent combinations of components with significantly unequal sizes, where small components (outliers) hold particular significance. In the context of Big Data, simplistic sampling procedures are frequently utilized for model estimation, despite the inherent risk of overlooking small mixture components. Consequently, model validation, the adoption of more suitable distribution-based sampling, and the use of resampling methods for predictive power assume critical importance.

Conclusion

Following the above assessment of the capabilities and impacts of statistics our conclusion is: The role of statistics in Data Science is under-estimated as, e.g., compared to computer science. This yields, in particular, for the areas of data acquisition and enrichment as well as for advanced modeling needed for prediction. Stimulated by this conclusion, statisticians are welladvised to more offensively play their role in this modern and well accepted field of Data Science. Only complementing and/or combining mathematical methods and computational algorithms with statistical reasoning, particularly for Big Data, will lead to scientific results based on suitable approaches. Ultimately, only a balanced interplay of all sciences involved will lead to successful solutions in Data Science. Acknowledgements The authors would like to thank the editor, the guest editors and all reviewers for valuable comments on an earlier version of the manuscript. They also thank LeoGeppert for fruitful discussions.

References

1.Adenso-Diaz, B., Laguna, M.: Fine-tuning of algorithms using fractional experimental designs and local search. Oper. Res. 54(1), 99–114 (2006)

2. Aggarwal, C.C. (ed.): Data Classification: Algorithms and Applications. CRC Press, Boca Raton (2014)

3. Allen, E., Allen, L., Arciniega, A., Greenwood, P.: Construction of equivalent stochastic differential equation models. Stoch. Anal. Appl. 26, 274–297 (2008)

4. Anderson, C.: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine https://www.wired.com/2008/06/pb-theory/ (2008)

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

5. Aue, A., Horváth, L.: Structural breaks in time series. J. Time Ser. Anal. 34(1), 1–16 (2013)

6. Berger, R.E.: A scientific approach to writing for engineers and scientists. IEEE PCS Professional Engineering Communication Series IEEE Press, Wiley (2014)

7. Bischl, B., Mersmann, O., Trautmann, H., Weihs, C.: Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol. Comput. 20(2), 249–275 (2012)

8. Bischl, B., Schiffner, J., Weihs, C.: Benchmarking local classification methods. Comput. Stat. 28(6), 2599–2619 (2013) 123 International Journal of Data Science and Analytics

9. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. arXiv preprint arXiv:1606.04838 (2016) 10. Brown, M.S.: Data Mining for Dummies. Wiley, London (2014)

11. Bühlmann, P., Van De Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin (2011)

12. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (2017). https://doi.org/10.1145/3076253

13. Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging. Cambridge University Press, Cambridge (2008)

14. Cooper, H., Hedges, L.V., Valentine, J.C.: The Handbook of Research Synthesis and Meta-analysis. Russell Sage Foundation, New York City (2009)

15. Dmitrienko, A., Tamhane, A.C., Bretz, F.: Multiple Testing Problems in Pharmaceutical Statistics. Chapman and Hall/CRC, London (2009)

16. Donoho, D.: 50 Years of Data Science. http://courses.csail.mit.edu/ 18.337/2015/docs/50YearsDataScience.pdf (2015)

17. Dyk, D.V., Fuentes, M., Jordan, M.I., Newton, M., Ray, B.K., Lang, D.T., Wickham, H.: ASA Statement on the Role of Statistics in Data Science. http://magazine.amstat.org/blog/2015/10/01/asastatement-on-the-role-of-statistics-in-data-science/ (2015)

18. Fahrmeir, L., Kneib, T., Lang, S., Marx, B.: Regression: Models, Methods and Applications. Springer, Berlin (2013) 19. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, Berlin (2006)

20. Geppert, L., Ickstadt, K., Munteanu, A., Quedenfeld, J., Sohler, C.: Random projections for Bayesian regression. Stat. Comput. 27(1), 79–101 (2017). https://doi.org/10.1007/s11222-015-9608-z

21. Hastie, T., Tibshirani, R.,Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)