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

Statistics in Data Science: A Profound Influence

Pulicherla Sushma [1], Chilakala Hari Krishna [2], Chejarla Ravi [3]

Abstract

This paper aims to support our claim that statistics holds a pivotal role among disciplines, offering essential tools and methodologies for uncovering patterns within data and gaining profound insights. Furthermore, we assert that statistics is paramount in the analysis and quantification of uncertainty. In this paper, we provide an extensive examination of various Data Science frameworks and emphasize the significant influence of statistics at each stage of the data analysis process, including data collection and enhancement, exploratory data analysis, data modeling, validation, and the presentation and reporting of findings. Additionally, we shed light on the potential pitfalls that may arise when statistical reasoning is disregarded.

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

Introduction

Data Science, as a scientific discipline, is a multifaceted field drawing influences from informatics, computer science, mathematics, operations research, statistics, and various applied sciences. In a significant milestone, the term "Data Science" was first integrated into the title of a statistical conference, the International Federation of Classification Societies (IFCS) conference in 1996, under the theme "Data Science, classification, and related methods" . Despite its origins in statistics, the public perception of Data Science often places more emphasis on computer science and practical business applications, especially in the age of Big Data.

The evolution of Data Science has been influenced by pivotal figures such as John Tukey, whose ideas in the 1970s transformed the perspective of statistics. Tukey shifted the focus from a purely mathematical context, centered on statistical testing, to an exploratory setting. This shift involved deriving hypotheses from data, striving to comprehend the data before formulating hypotheses.

Another foundational element of Data Science is the concept of Knowledge Discovery in Databases (KDD), with a particular emphasis on its subfield, Data Mining. KDD serves as a platform that unites diverse approaches for acquiring knowledge and insights from data.

edge discovery, including inductive learning, (Bayesian) statistics, query optimization, expert systems, information theory, and fuzzy sets. Thus, KDD is a big building block for fostering interaction between different fields for the overall goal of identifying knowledge in data.

Nowadays, these ideas are combined in the notion of Data Science, leading to different definitions. One of the most comprehensive definitions of Data Science was recently given by Cao as the formula

[1] Assistant Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions,
[2] Associate Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions,
[3] Assistant Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions

data science = (statistics + informatics + computing + communication + sociology + management) | (data + environment + thinking).

In this formula, sociology stands for the social aspects and | (data + environment + thinking) means that all the mentioned sciences act on the basis of data, the environment and the so- called data-to-knowledge-to-wisdom thinking.

A recent, comprehensive overview of Data Science provided by Donoho in 2015 focuses on the evolution of Data Science from statistics. Indeed, as early as 1997, there was an even more radical view suggesting to rename statis- tics to Data Science . And in 2015, a number of ASA leaders released a statement about the role of statistics in Data Science, saying that "statistics and machine learning play a central role in data science."

In this paper's perspective, the main steps in Data Science have drawn inspiration from CRISP-DM but have evolved into a sequence of the following 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. The paper acknowledges that statistics may have a reduced role in steps indicated in small capitals.

Typically, these steps are not executed in isolation but rather as part of an iterative and cyclic process. Moreover, it's common to alternate between multiple steps, notably in Data Acquisition and Enrichment, Data Exploration, Statistical Data Analysis, Data Analysis and Modeling, and Model Validation and Selection.

Table 1 in the paper compares different definitions of Data Science steps and illustrates the relationships between them. It highlights the absence of Data Acquisition and Enrichment in CRISP-DM, suggesting that CRISP-DM primarily deals with observational data. Additionally, the paper's proposal incorporates Data Storage and Access and Optimization of Algorithms, steps where statistics plays a less prominent role, into CRISP-DM.

The paper notes that the list of steps in Data Science can be further expanded, referencing Cao's proposal, which includes additional steps like Domain-specific Data Applications and Problems, Data Storage and Management, Data Quality Enhancement, and others. These align with the main steps in the paper's proposal, albeit with some variations in terminology and detail, depending on the background of the author (computer science or statistics).

In the subsequent sections of the paper, the role of statistics is discussed in detail, with a focus on the steps where it is significantly involved, encompassing all steps in the proposal from Table 1, except those indicated in small capitals.

In our view, statistical methods are crucial in most fundamental steps of Data Science. Hence, the premise of our contribution is:

Statistics is one of the most important disciplines to pro- vide tools and methods to find structure in and to give deeper

Methodology

This paper centers around the significance of statistics in the field of Data Science, emphasizing its pivotal role in providing insight into data and quantifying uncertainty. It delves into the major impact of statistics on the fundamental steps of Data Science, which have evolved from influential models like CRISP-DM (Cross Industry Standard Process for Data Mining).

CRISP-DM, organized into six primary steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, has played a foundational role in applied statistics.

Table 1: Comparison of Steps in Data Science - CRISP-DM, Cao's Definition, and Our Proposal

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

The entries "Data Storage and Access" and "Optimization of Algorithms" are primarily within the domain of informatics and computer science, where expertise in data management and algorithm optimization plays a crucial role. On the other hand, "Business Deployment of Results" is a step that typically falls under the purview of Business Management, involving decision-making, strategy, and practical implementation of data-driven insights within an organization.

2.1 Data acquisition and enrichment

Design of experiments (DOE) is essential for a systematic generation of data when the effect of noisy factors has to be identified. Controlled experiments are fundamental for robust process engineering to produce reliable products despite variation in the process variables. On the one hand, even con- trollable factors contain a certain amount of uncontrollable variation that affects the response. On the other hand, some factors, like environmental factors, cannot be controlled at all. Nevertheless, at least the effect of such noisy influencing factors should be controlled by, e.g., DOE.
DOE can be utilized, e.g.,

– to systematically generate new data (data acquisition) ,
– for systematically reducing data bases , and
– for tuning (i.e., optimizing) parameters of algorithms , i.e., for improving the data analysis methods themselves.

Simulations may also be used to generate new data. A tool for the enrichment of data bases to fill data gaps is the imputation of missing data .
Such statistical methods for data generation and enrichment need to be part of the backbone of Data Science. The exclusive use of observational data without any noise control distinctly diminishes the quality of data analysis results and may even lead to wrong result interpretation. The hope for "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" [4] appears to be wrong due to noise in the data.
Thus, experimental design is crucial for the reliability, validity, and replicability of our results.

2.2 Data exploration

Exploratory statistics is essential for data preprocessing to learn about the contents of a data base. Exploration and visualization of observed data was, in a way, initiated by John Tukey. Since that time, the most laborious part of data analysis, namely data understanding and transformation, became an important part in statistical science.
Data exploration or data mining is fundamental for the proper usage of analytical methods in Data Science. The most important contribution of statistics is the notion of distribu- tion. It allows us to represent variability in the data as well as (a-priori) knowledge of parameters, the concept underlying Bayesian statistics. Distributions also enable us to choose adequate subsequent analytic models and methods.

2.3 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.

a) Hypothesis testing is one of the pillars of statistical anal- ysis. Questions arising in data driven problems can often be translated to hypotheses. Also, hypotheses are the natural links between underlying theory and statistics. Since statistical hypotheses are related to statistical tests, questions and theory can be tested for the available data. Multiple usage of the same data in different tests often leads to the necessity to correct significance levels. In applied statistics, correct multiple testing is one of the most important problems, e.g., in pharmaceutical studies. Ignoring such techniques would lead to many more significant results than justified.

b) Classification methods are basic for finding and predict- ing subpopulations from data. In the so-called unsuper- vised case, such subpopulations are to be found from a data set without a-priori knowledge of any cases of such subpopulations. This is often called clustering.

In the so-called supervised case, classification rules should be found from a labeled data set for the prediction of unknown labels when only influential factors are available.

Nowadays, there is a plethora of methods for the unsupervised as well for the supervised case

In the age of Big Data, a new look at the classical methods appears to be necessary, though, since most of the time the calculation effort of complex analysis methods grows stronger than linear with the number of observations n or the number of features p. In the case of Big Data, i.e., if n or p is large, this leads to too high calcula- tion times and to numerical problems. This results both, in the comeback of simpler optimization algorithms with low time-complexity and in re-examining the tradi- tional methods in statistics and machine learning for Big Data .

c) Regression methods are the main tool to find global and local relationships between features when the tar- get variable is measured. Depending on the distributional assumption for the underlying data, different approaches may be applied. Under the normality assumption, linear regression is the most common method, while gener- alized linear regression is usually employed for other distributions from the exponential family . More

This section provides an overview of various advanced statistical methods and their applications in the field of Data Science:

1. Functional Regression for Functional Data: This method, such as the one described in reference is used to analyze data where the variables are functions. It can be applied to a wide range of fields, including economics, natural sciences, and engineering.

2. Quantile Regression: Quantile regression, as referenced in , is a technique used to model conditional quantiles of a response variable. It is valuable for understanding the relationships between variables in different quantiles.

3. Regression Based on Loss Functions: Regression methods, like Lasso regression mentioned in , consider loss functions other than the traditional squared error loss. These approaches are employed to address specific data characteristics and modeling requirements.

4. Challenges in Big Data: Handling large datasets in the context of Big Data poses challenges, particularly due to the high volume of observations (n) and features (p). Techniques such as compressed sensing, random projection, and sampling-based procedures are used to reduce computation time (n reduction). Additionally, variable selection and shrinkage methods like Lasso help reduce the number of features (p reduction) while preserving their interpretability.

5. Time Series Analysis: Time series analysis, as discussed in , focuses on understanding and predicting temporal patterns. This analysis is essential in various fields, including behavioral sciences, economics, and natural sciences. Statistical methods are applied to model and predict future values or properties of time series data.

6. Statistical Modeling:

Complex Interactions: Complex interactions between factors can be modeled using graphs or networks. These interactions can be directed or undirected and are valuable in various fields.

Stochastic Differential and Difference Equations: These equations are used to represent models in natural and engineering sciences. Approximate statistical models derived from such equations can provide insights into statistical control of processes.

Local Models and Globalization: Statistical models are often valid only in sub-regions of variable domains. Local models, structural breaks, and concept drift analysis are employed to adapt models to local regions or time periods.

Mixture Models: Mixture models are used to generalize from local to global models, accounting for heterogeneity in data.

7. Model Validation and Model Selection: In cases where multiple models are proposed for prediction or analysis, statistical tests, resampling methods, and perturbation experiments are used to assess and compare model performance. Model selection techniques help choose the most suitable model for a given task.

8. Representation and Reporting: Visualization and model storage in an easily updatable format are essential for interpreting results and deploying data analysis. Reporting of uncertainties and review are crucial for communicating results effectively and ensuring the accuracy of data analysis.

This section highlights the diverse range of statistical methods and their applications in Data Science, demonstrating the field's versatility and significance in various domains and data-driven decision-making.

In the realm of data analysis, there are potential fallacies that can arise when statistical methods are not adequately considered or when overly simplistic data analytics and statistical techniques are employed. These fallacies are especially pertinent in the analysis of large and complex datasets. Here are some key fallacies that can be encountered:

1. Neglecting the Role of Distributions: Distributions, as emphasized in Section 2.2, are a cornerstone of statistical thinking. Failing to take distributions into account during data exploration and modeling can limit the reporting of values and parameter estimates without considering their associated variability. Distributions are crucial for making predictions with corresponding error bands.

2. Importance of Model-Based Data Analytics: Model-based data analytics often require an understanding of distributions. For instance, in unsupervised learning to find clusters in data, incorporating additional structural information, such as spatial or temporal dependencies, may necessitate the inference of parameters like cluster radii and their spatio-temporal evolution. This type of analysis heavily relies on the concept of distributions.

3. Comparing Multivariate Hypothesis Testing: When multiple parameters are of interest, it is advisable to compare univariate hypothesis testing approaches to multiple procedures, such as in multiple regression. Restricting analysis to univariate testing may lead to overlooking relationships between variables.

4. Complex Models for Deeper Insights: Gaining deeper insights into data may require more complex models, such as mixture models for detecting heterogeneous groups within the data. Ignoring the presence of mixtures can result in meaningless averages, necessitating the identification of subgroups through unmixing components. In a Bayesian framework, latent allocation variables in models like the Dirichlet mixture model can facilitate this process. For example, such models can be applied to decompose a mixture of different networks within a heterogeneous cell population in molecular biology.

5. Handling Mixtures with Unequal Component Sizes: In some scenarios, mixture models may represent mixtures of components with highly unequal sizes, where small components (outliers) are of particular importance. In the context of Big Data, naïve sampling procedures are often used for model estimation. However, these procedures carry the risk of missing small mixture components. Therefore, model validation and the use of more appropriate distribution choices, as well as resampling methods to assess predictive power, are crucial to avoid these pitfalls.

In summary, understanding the role of distributions, employing model-based data analytics, considering multivariate approaches, using complex models for deeper insights, and being cautious with respect to mixtures of unequal sizes are all essential aspects of robust and effective data analysis, especially in the context of large and complex datasets. Ignoring these principles can lead to avoidable fallacies in data analysis.

Conclusion

In conclusion, the assessment of the capabilities and impacts of statistics in Data Science leads us to a fundamental observation:

The role of statistics in the field of Data Science is often underestimated when compared to the prominence of computer science. This underestimation is particularly evident in the areas of data acquisition and enrichment, as well as advanced modeling essential for prediction.

This conclusion underscores the need for statisticians to take a more proactive and assertive role in the evolving and highly respected domain of Data Science. It is crucial for statisticians to recognize and assert their contribution to this field.

Furthermore, it's essential to emphasize that, especially in the context of Big Data, the integration and synergy of mathematical methods, computational algorithms, and statistical reasoning are necessary to yield scientifically sound results based on appropriate approaches. Achieving success in Data Science necessitates a harmonious collaboration between all scientific disciplines involved.

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 Appli- cations. 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)
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 evo- lutionary computation. Evol. Comput. 20(2), 249–275 (2012)
8. Bischl, B., Schiffner, J., Weihs, C.: Benchmarking local classifica- tion methods. Comput. Stat. 28(6), 2599–2619 (2013)
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