SUPPORT:

VIRTUTE SPIRITUS

SCIENDI FUNDAMEN

DEPARTAMENTO DE ESTATÍSTICA UFBA

CAPES

CNPq

CONRE-5
CONSELHO REGIONAL DE ESTATÍSTICA DA 5ª REGIÃO
(AL - BA - PB - PE - RN - SE)

iasc isi
International Association for Statistical Computing

AMERICAN STATISTICAL ASSOCIATION FOUNDED 1839

Jusbrasil

avansys

BOOK OF ABSTRACTS

2nd Conference on Statistics and Data Science

Salvador-BA, Brazil, November 18-20, 2019

Book of abstracts of the 2nd Conference on Statistics and Data Science

November 18-20, 2019

Salvador - BA, Brazil

Edited by

Paulo Canas Rodrigues
Elisabete Batista Sampaio
Jalmar Carrasco
Lizandra Fabio
Federal University of Bahia, Salvador, Brazil

Web Design
Jalmar Carrasco
Federal University of Bahia, Salvador, Brazil

Design of the Cover
Samuel Barros
Federal University of Bahia, Salvador, Brazil

ISBN: 978-65-5102-008-7

Cite as:
Rodrigues, Paulo Canas; Fábio, Lisandra; Carrasco, Jalmar; Sampaio, Elisabete. Book of abstracts of the 2nd conference on statistics and data science. Florestal(MG): Sguerra Design, 2019.

Contents

Part I. Introduction

Part II. Scientific Program

Part III. Keynote Lectures

Part IV. Short Courses

Part V. Invited Paper Sessions

Part VI. Round Tables

Part VII. Tutorial

Part VIII. Contributed Papers

Part I

Introduction

Welcome to the CSDS 2018

On the behalf of the Scientific Program Committee and of the Local Organizing Committee we are pleased to welcome you to Salvador. This is the 2nd edition of the Conference on Statistics and Data Science (CSDS 2019) that we have the pleasure to organize in the beautiful city of Salvador.

The organization of this meeting has been carried out in collaboration with the Department of Statistics of the Federal University of Bahia, Brazil. The purpose of the CSDS 2019 is to bring together researchers and practitioners, from the academy and from the industry, that develop and apply statistical and computational methods for data science. This conference will provide a forum to share and discuss ways to improve the access to knowledge, and promote interdisciplinary collaborations.

The scientific program includes two keynote speakers who are national and international references in statistics and data science, six short courses, two invited paper sessions, three round tables, one tutorial, six contributed paper sessions, one general poster session, and one poster session where the students of the Specialization in Data Science and Big Data present their final projects, with a total of about 80 contributions.

The CSDS 2019 will offer opportunities to meet each other, to share scientific and professional experiences, and to promote new collaborations. This meeting will cover many of the research areas of Statistics and Data Science.

Students and young statisticians (up to 5 years after their last academic degree) attending this conference had the possibility to compete for two awards: “Best Paper Award on Statistics and Data Science”, and “Best Poster Award on Statistics and Data Science”.

The program also includes social events which will allow the participants to get to know each other and to experience the culture and history of Bahia, in addition to the taste of the well known Bahia’s cuisine and hospitality.

The organizers would like to thank to all institutions that have provided financial and other support to make this organization possible. Many thanks to the Coordination for the Improvement of Higher Education Personnel (CAPES), to the National Council for Scientific and Technological Development (CNPq), to the American Statistical Association, to the International Association for Statistical Computing, to the Federal University of Bahia, in particular the MSc program in Mathematics, to CONRE-5, to avansys, and to Jusbrasil, for the financial support which made possible the organization of the CSDS 2019. Last, but not the least, we thank the keynote speakers, the lecturers of the short courses, the speakers and discussants in the invited paper sessions and round tables, the speakers in the contributed sessions and the poster presenters, for their contribution to make a great scientific program. Thank you all for being here!

We wish you an enjoyable stay and a good time in Salvador!

On the behalf of the Scientific Program Committee and of the Local Organizing Committee,

Paulo Canas Rodrigues
Chair of the Scientific Program Committee of the CSDS 2019

Lizandra Fábio
Chair of the Local Organizing Committee of the CSDS 2019

Local Organizing Committee

- Lizandra Fabio (Chair), Federal University of Bahia, Brazil
- Paulo Canas Rodrigues (Chair), Federal University of Bahia, Brazil
- Elisabete Sampaio, Federal University of Bahia, Brasil
- Gabriel Ferreira dos Santos, Federal University of Bahia, Brazil
- Gilberto Pereira Sassi, Federal University of Bahia, Brazil
- Jalmar Carrasco, Federal University of Bahia, Brazil
- José Roberto Santos da Silva, Federal University of Bahia, Brazil
- Matheus Hofstede, Federal University of Bahia, Brazil
- Murilo Martins, Federal University of Bahia, Brazil
- Samuel Barros, Federal University of Bahia, Brazil
- Talita Nacimento dos Santos, Federal University of Bahia, Brazil

Scientific Program Committee

- Paulo Canas Rodrigues (Chair), Federal University of Bahia, Brazil
- David Banks, Duke University, USA
- Francisco Louzada, University of São Paulo, Brazil
- Hedibert Lopes, INSPER, Brazil
- Jalmar Carrasco, Federal University of Bahia, Brazil
- Katherine D. Ensor, Rice University, USA
- Lizandra Fábio, Federal University of Bahia, Brazil
- Luciano Rebouças de Oliveira, Federal University of Bahia, Brazil
- Moncef Gabbouj, Tampere University of Technology, Finland
- Narayanaswamy Balakrishnan, McMaster University, Canada
- Patrick Groenen, Erasmus University Rotterdam, The Netherlands
- Paula Brito, University of Porto, Portugal
- Wing K. Fung, University of Hong Kong, Hong Kong

Call for Papers

We are pleased to announce a **Special Issue of the journal Statistics, Optimization & Information Computing (SOIC)** devoted to papers presented at the 2nd Conference on Statistics and Data Science. This special issue will include selected papers strongly correlated to the talks of the conference and within the scope of the journal.

Guest Editors: Paulo Canas Rodrigues and Jalmar Carrasco

All papers submitted must meet the publication standards of the journal Statistics, Optimization & Information Computing (http://www.iapress.org/index.php/soic) and will be subject to normal refereeing procedure. Authors should send the papers should be sent to paulocanas@gmail.com, and use the SOIC template (available here: https://www.sugarsync.com/pf/D358196_07941120_0851496). Please be aware that a Special Issue is not a proceedings paper. Presenting a paper does not imply publication in a Special Issue. **The deadline for paper submission is January 31, 2020**.

Part II

Scientific Program

Scientific Program

2nd Conference on Statistics and Data Science
November 18-20, 2019, Salvador, Brazil

Time	**Monday, 18/11/2019**	**Tuesday, 19/11/2019**	**Wednesday, 20/11/2019**

08:00–08:30	**SC1**: {Pelourinho}
	SC2: {Mercado Modelo}
	Poster Session - ECD
		SC3: {Mercado Modelo}
		SC4: {Pelourinho}
			SC3: {Mercado Modelo}
			SC4: {Pelourinho}

08:30–09:00			
09:00–09:30			
09:30–10:00			
10:00–10:30	**Coffee Break**	**Coffee Break**	**Coffee Break**

10:30–11:00	**SC1**: {Pelourinho}
	SC2: {Mercado Modelo}
		W1: Richard De Veaux {Mercado Modelo}	**RT3**: {Mercado Modelo}

11:00–11:30			

11:30–12:00		**CPS3**: {Mercado Modelo}
		CPS4: {Pelourinho}
			KS3: Pedro Silva {Mercado Modelo}

12:00–12:30			
12:30–13:00	**Lunch**	**Lunch**	**Closing Ceremony**
13:00–13:30			**Lunch**
13:30–14:00			

14:00–14:30	**Opening Ceremony**	**RT2**: {Mercado Modelo}	**SC5**: {Pelourinho}
			SC6: {Mercado Modelo}

14:30–15:00	**KS1**: Richard De Veaux {Mercado Modelo}		
15:00–15:30			
15:30–16:00	**Coffee Break**	**Coffee Break**	

16:00–16:30	**IPS1**: {Mercado Modelo}
	IPS2: {Pelourinho}
		CPS5: {Mercado Modelo}
		avansys: {Pelourinho}
			Coffee Break

16:30–17:00			**SC5**: {Pelourinho}
			SC6: {Mercado Modelo}

17:00–17:30			

17:30–18:00	**CPS1**: {Mercado Modelo}
	CPS2: {Pelourinho}
		Poster Session	

18:00–18:30			
18:30–19:00	**RT1**: {Mercado Modelo}		
19:00–19:30			
19:45–	**Welcome Cocktail**		

Round Tables
RT1: How to get your paper published?
RT3: Data science and big data in Brazil
RT3: The role of statistics in the Era of big data

Workshop
W1: Some tips for effective presentations

Contributed Paper Sessions
CPS1: Machine learning methodology and applications
CPS2: Data science applications to the society
CPS3: Data science in practice
CPS4: Recent advances in regression and longitudinal models
CPS5: Recent advances in statistics and data science

Short Courses
SC1: Implementation of web crawling processes (Crysttian Arantes Paixão)
SC2: How to simplify your statistical reports using RMarkdown (Marcus Nunes)
SC3: My first dashboard with Shiny (Athos Petri Damiani)
SC4: Data literacy (Makson Reis)
SC5: An introduction to time series analysis with R (Everaldo Guedes)
SC6: Tidyverse: A data science introduction with R (Lucas Mascarenhas Almeida and Tarssio Barreto)

Invited Paper Sessions
IPS1: Data Science, Communication & Democracy: methodological approaches
IPS2: Statistical and Data Sciences and Recent Applications

Part III

Keynote Lectures

KS1: The seven deadly sins of big data – and how to avoid them

Richard D. De Veaux[1,2]

[1] Williams College, USA
[2] Vice-President of the American Statistical Association
Email: rdeveaux@williams.edu

Abstract

Organizations, from government to industry accumulate vast amounts of data from a variety of sources nearly continuously. Big data advocates promise the moon and the stars as you harvest the potential of all these data. There is certainly a lot of hype. There's no doubt that some savvy organizations are fueling their strategic decision making with insights from data mining, but what are the challenges? Much can wrong data analysis cycle, even for trained professionals. In this talk I'll discuss a wide variety of case studies from a range of industries to illustrate the potential dangers and mistakes that can frustrate problem solving and discovery – and that can unnecessarily waste resources. My goal is that by seeing some of the mistakes I have made, you will learn how to take advantage of data insights without committing the "Seven Deadly Sins."

KS2: Big data: potencial, paradoxos e a importância renovada do pensamento estatístico

Pedro Luis do Nascimento Silva[1]

[1]Brazilian Institute of Geography and Statistics, RJ, Brazil
Email: pedronsilva@gmail.com

Abstract

Vivemos numa era em que a disponibilidade e acessibilidade a dados não tem precedentes. 'Big data' é uma das tendências deste início do Milênio a confrontar o pensamento estatístico. Por um lado, há imenso potencial para aproveitar as novas fontes de informação que se tem tornado disponíveis, acessíveis e de baixo custo. Por outro lado, lacunas substanciais persistem e há imensos riscos de utilização inadequada dessas fontes pelos que desprezam as lições traduzidas nos principais fundamentos do pensamento e da metodologia estatística. Uma das falácias principais é a de que, com as imensas bases de dados disponíveis, não será mais preciso avaliar incerteza de estimativas, pois será possível 'conhecer' as quantidades de interesse a partir dos 'big data'. Apresentarei o conceito de 'Índice de defeito dos dados' proposto por Meng (2018), e usarei este conceito para mostrar que a qualidade de estimativas baseadas em pequenas amostras bem planejadas e executadas pode superar a de estimativas baseadas em conjuntos muito maiores provenientes de fontes orgânicas sujeitas a vieses de seleção. Penso que a metodologia estatística fornece a orientação essencial necessária para obter respostas atuais, relevantes, precisas e custo-efetivas às perguntas de interesse, mesmo na era do 'big data'. Apresentarei alguns exemplos para motivar a discussão dessas ideias e de caminhos para superar as limitações das novas fontes de informação.

Part IV

Short Courses

SC1: Implementation of web crawling processes

Crysttian Paixão[1]

[1]Federal University of Santa Catarina, SC, Brazil
Email: crysttian@gmail.com

Abstract: A big amount of the current knowledge is currently available on the internet. Since there is no standardization, it is necessary to develop techniques that must be customized so that data can be collected and processed. Among the numerous methodologies there is the web crawler. The web crawler is a software that performs a series of activities, which can be customized to perform the data collection. This short course aims to present some methodologies for implementing different types of web crawlers to collect different kinds of information and perform preprocessing. (slides in English, talk in Portuguese) **[4 hours]**

SC2: How to simplify your statistical reports using RMarkdown

Marcus Nunes[1]

[1]Federal University of Rio Grande do Norte, RN, Brazil
Email: marcus.nunes@gmail.com

Abstract: This course will be a practical introduction on how to create a reproducible workflow using RMarkdown. The participants will see how they can create tidy reports merging R code and their conclusions. There will be data analysis exercises where reports will be written by the participants, showing them how to create their own reports. (slides in English, talk in Portuguese/English) [**4 hours**]

SC3: My first dashboard with Shiny

Athos Damiani[1]

[1]IME and POLI, University of São Paulo, and Curso-R, SP, Brazil
Email: athos.damiani@gmail.com

Abstract: This hands on workshop will guide the student into their first steps onto creation of interactive dashboards using R and Shiny. An app for retrieve predictions from a machine learning model will be build throughout the class. (slides in English, talk in Portuguese/English) [**4 hours**]

SC4: Data literacy

Makson Reis[1]

[1]Federal University of Bahia, BA, Brazil
Email: maksonacademico@gmail.com

Abstract:Research Data Literacy aims to establish and disseminate among the event's participating community, training and literacy for data types, data management, data treatments and data curation forms. Research data or scientific data are records, files or content in print or digital format that contain results of research observations that may be shared among the academic community and may be contained in documents; laboratory minutes; questionnaires and reports; tapes; CDs; DVDs; photographs; Slides; statistical data files; database contents (video, audio, text, images); scripts and maps. With the ability to process, classify and filter large amounts of information, the professional should be able to search, filter and process, as well as produce and synthesize data. The short course will enable the participant to know and understand the importance of management and techniques used for their organization, with the aim of short course to improve skills and learning in the ecology of data science. (slides in English/Portuguese, talk in Portuguese) [**4 hours**]

SC5: An introduction to time series analysis with R

Everaldo Guedes[1]

[1]Federal University of Bahia, BA, Brazil
Email: efgestatistico@gmail.com

Abstract: In this course, we will approach the time series analysis with R, covering topics such as: time series plotting, decomposition, simulation and forecasting. (slides in Portuguese, talk in Portuguese) [**4 hours**]

SC6: Tidyverse: A data science introduction with R

Lucas Mascarenhas Almeida[1] and Tarssio Barreto[1]

[1]Federal University of Bahia, BA, Brazil
Email: lucasmascalmeida@gmail.com; tarssioesa@gmail.com

Abstract: Tidyverse is an ecosystem of R programming language packages designed for data science applications. The course is a brief dive into this precious ecosystem. The course consists of four steps: An overview about the R world, data import using the “readr” and “readxl” packages, data manipulation with pipe and the “dplyr” package and finally data visualization with the “ggplot2” package. (slides in English/Portuguese, talk in Portuguese) [**4 hours**]

Part V

Invited Paper Sessions

IPS1 – Data Science, Communication & Democracy: Methodological approaches

Samuel Barros[1] (Organizer) and Nina Santos[1] (Chair)

[1]Federal University of Bahia, BA, Brazil
Email: samuel.barros77@gmail.com

IPS1.1: Classification and opinion mining in digital environments: Analysing user perceptions of mobile government applications

Eurico Matos[1] and Alexandre Teles[1]

[1] Federal Institute of Bahia, BA Brazil

Abstract: This paper aims to discuss the use of semantic analysis techniques to classify and to measure the sentiment of opinions published in digital environments. More specifically, we focus on the analysis of mobile governmental apps users' comments published on Google Play. User experience is one of the most important aspects of assessment and ongoing development of digital platforms. The app store feedback section is a rich source for researchers interested in measuring user opinion, sentiment, and issues about apps. In addition to contributing to the development of research in the areas of Data Science and App Studies, the proposed methodology seeks to assist developers and public managers to understand the mobile government applications users' needs.

IPS1.2: Methodological proposal for agenda-setting research: Using topic modelling to understand big social data

Júnia Ortiz[1], André Rebouças[1], Gustavo Nunes[1] and Laion Boaventura[1]

[1]Federal University of Bahia, BA, Brazil

Abstract: The paper presents a methodological proposal for studies aimed at understanding the conversation in online environments. The goal is to understand the relationships between journalistic content and the topics circulating among users of social networking sites. For this, we start from the notions of Agenda-Setting (MCCOMBS; SHAW, 1992) and Newsmaking (BOYER, 2013) in order to build a theoretical foundation to the method. We collected and analyzed 572 thousand Twitter messages and 216 articles published in Folha de S.Paulo online newspaper about Jair Bolsonaro, current president of Brazil, between May 6 and 9, 2019. For tweet content analysis, we applied an unsupervised Topic Modeling algorithm, Structural Topic Models (STM). The classified themes were compared to the themes in the journalistic articles, identified from the frequency analysis of the terms in the headlines of the newspaper Folha de S. Paulo. The results indicate the relevance of using Text Mining and Machine Learning techniques to perform content analysis of big social data.

IPS1.3: Network analysis applied to viralization of false news: Applications of algorithms for classification of WhatsApp groups

João Guilherme Santos[1]

[1]Federal University of Bahia, BA, Brazil

Abstract: This paper brings a network analysis approach to problems encountered by researchers focused on viral fake news shared using mobile messaging applications. By understanding which network metrics are successful in predicting presidential paths for the circulation and viralization of misinformation, as well as its segmentation, we advance in the sense of fostering a dialogue between qualitative, quantitative and relational analyses around issues involving the internet and democracy. Our object of analysis is the collective behavior of 90 interconnected WhatsApp groups supporting six main presidential candidates, as well as the more than 500,000 texts and images sent by this application's users, during the five electoral campaign months. As one of this proposal results, we advance into tracing original sources of fake news during Brazilian electoral period. We identified that the extended scope of the application is a direct consequence of the structural interconnection among these groups and its topology. And that's what makes the viralization of fake news possible.

IPS1.4: Blockchain and government services: The state of the art of research and applications

Lucas Reis[1]

[1]Federal University of Bahia, BA, Brazil

Abstract: Distributed ledger technology has been used in many industries, especially in logistics and finance. Its potential has also been studied and applied in government initiatives to increase security, traceability, efficiency and scale of actions, both in providing public services to citizens, as well as in asset management and public budget. However, as presented by Carter and Ubacht (2018), their actual implementation raises questions about governance, privacy and inclusion of citizens. In this paper, we explore the blockchain impacts at all levels of government and introduce considerations regarding its effects on society, by analysing a corpus of 20 articles published last year about the topic.

IPS2 – Statistical and data sciences and recent applications

Francisco Louzada[1] (Organizer and Chair)

[1]Federal University of Bahia, BA, Brazil
Email: louzada@icmc.usp.br

IPS2.1: Bayesian networks: Methods and applications

Anderson Ara[1]

[1]Federal University of Bahia, BA, Brazil

Resumo: Bayesian networks, also known as causal networks, belief networks, or probabilistic dependency networks, emerged in the 1980s and have been applied to a wide variety of real-world activities. They are a machine learning model which has a graphical representation (acyclic and directed graph) of the variables and their relations to a specific problem. This presentation will expose some methods of network construction and parameter estimations as well as recent applications.

IPS2.2: A generalized Gamma zero-inflated cure-rate regression model applied on obstetric healthcare

Gleici da Silva Castro Perdoná[1]

[1]University of São Paulo, SP, Brazil

Abstract: In survival analysis there is a lack of proposed models which allow that time be equal to zero, although we have several examples in which these models would be necessary in practice. In this work, we propose a Generalized Gamma Zero-Inflated Cure-Rate (GG-ZICR) survival model, motivated by the pattern observed in the labour progression of African women, in which we have three groups: women who arrive at the hospital already having had a stillbirth (fetal death); women that may not undergo vaginal birth because of an intervention (caesarean section, for example) and women that undergo vaginal birth without fetal death or intervention. The study of childbirth times can support management in obstetrical healthcare, and it is important that the three groups be considered for more consistent results. From the time point of view, the first group present time equal to zero and the other one's present incomplete time greater than zero and complete times greater than zero, respectively. Besides proposing a more flexible ZICR model, we carried out a simulation study in order to evaluate the inference properties of maximum likelihood estimators, asymptotic confidence intervals and comparison methods and we present primary

results about residual analysis. Finally, we present an application of an OMS of data set. The work is co-authored by Hayala Cristina Cavenague de Souza and Francisco Louzada.

IPS2.3: Statistical and data sciences at CeMEAI

Francisco Louzada[1]

[1]ICMC, University of São Paulo, SP, Brazil

Abstract: Statistical and data science methodologies has been widely used in innovation processes, promoting interaction with professionals from the governmental and productive sectors, as well as with the community. efficiently directing the dialogue between academia and industry. This conference presents the main innovation projects in data science that we have developed within the CeMEAI (Center for Mathematics and Statistics Applied to Industry) of USP, in order to bring the academy, the productive sector and the community closer. Focus is given to reliability modeling for oil well construction equipment, financial transaction fraud detection classification, electroencephalography data, and communication for mobile phones and autonomous unmanned aerial vehicles. In addition, we present how we are dealing with the transfer of knowledge of technological developments obtained to our industrial partners.

Part VI

Round Tables

RT1: How to get your paper published?

Paulo Canas Rodrigues[1]

[1]Federal University of Bahia, Brazil
Email: paulocanas@gmail.com

Abstract: In this round-table, two widely experience scientists will share some of their views about publishing, while giving some important advices for younger researchers.

Participants:

- Francisco Louzada (ICMC, University of São Paulo, Brazil)
  Email: louzada@icmc.usp.br
- Richard D. De Veaux (Williams College, USA; Vice-President of the American Statistical Association)
  Email: rdeveaux@williams.edu

RT2: Data science and big data in Brazil

Paulo Canas Rodrigues[1]

[1]Federal University of Bahia, Brazil
Email: paulocanas@gmail.com

Abstract: In this round-table, four Brazilian statisticians will discuss the current state of statistics, data science and big data in Brazil.

Participants:

- Francisco Louzada (ICMC, University of São Paulo, SP, Brazil)
  Email: louzada@icmc.usp.br
- Jalmar M F Carrasco (Federal University of Bahia, BA, Brazil)
  Email: carrascojalmar@gmail.com
- Marcus Nunes (Federal University of Rio Grande do Norte, RN, Brazil)
  Email: marcus.nunes@gmail.com
- Pedro Luis do Nascimento Silva (ENCE/IBGE, RJ, Brazil)
  Email: pedronsilva@gmail.com

RT3: The role of statistics in the Era of big data

Paulo Canas Rodrigues[1]

[1]Federal University of Bahia, Brazil
Email: paulocanas@gmail.com

Abstract: In this round-table, three national and international references in statistics and data science will discuss the current state of statistical sciences and share their vision on the role of statistics in the Era of big data.

Participants:

- Richard D. De Veaux (Williams College, USA; Vice-President of the American Statistical Association)
  Email: rdeveaux@williams.edu
- Pedro Luis do Nascimento Silva (ENCE/IBGE, RJ, Brazil; former President of the International Statistical Institute)
  Email: pedronsilva@gmail.com
- Denise Britz de Nascimento Silva (IBGE, RJ, Brazil; President of the International Association of Survey Statisticians)
  Email: denisebritz@gmail.com

Part VII

Tutorial

Tutorial: Some tips for effective presentations

Richard D. De Veaux[1,2]

[1] Williams College, USA
[2] Vice-President of the American Statistical Association
Email: rdeveaux@williams.edu

Abstract

Have you ever looked out at your audience and realized that half of them are either asleep or looking at their phones?

It's happened to all of us. And it's not a great feeling. In this pair of short talks we'll show some of the best (and worst) practices for making the most of your presentations.

Part VIII

Contributed Papers

CPS1: Machine learning methodology and applications

CPS1.1: Generalized model trees: A new algorithm proposal for regression tasks

Alberto Rodrigues Ferreira[1], Tibérius de Oliveira e Bonates[1] and Juvêncio Santos Nobre[1]

[1]Federal University of Ceará, CE, Brazil
Email: kobe.alberto7@gmail.com

Abstract

In many practical problems related to supervised machine learning, there is interest in predicting a continuous variable. In this work, we propose and evaluate the accuracy of a new algorithm called Generalized Model Tree (GMT), which is based on the Model Tree algorithm. A GMT addresses a different structure compared to the traditional model tree, both in predicting the variable response and with respect to the strategy used to try to prevent overfitting. In particular, a GMT can use, for the purpose of obtaining better accuracy than a conventional model tree, various regression models, such as generalized linear models, ridge regression and polynomial regression, adjusted on subsets of the training data. Tests were performed on publicly available datasets with the following algorithms: GMT, multiple linear regression, ridge regression, lasso regression, Bayesian ridge regression, random forests, regression tree, multilayer perceptron, and support vector regression. Fifty random partitions were performed, with 80% for training and 20% for testing, on three data sets obtained from the UCI Machine Learning Repository. For all algorithms, the mean absolute error metric was used. For the Iris dataset, we predicted the size of the sepal; for the Auto MPG dataset, we predicted the Weight variable; and for the Concrete Slump Test dataset, we predicted the Compressive Strength variable. The GMT algorithm obtained an average error close to the best average error obtained by the other algorithms tested.

CPS1.2: Clustering and elastic net logistic regression as support tools for Honeybee (Apis mellifera) colonies health diagnosis

Daniel de Amaral da Silva[1], Antonio Rafael Braga[1], Juvêncio Santos Nobre[1] and Danielo Gonçalves Gomes[1]

[1]Federal University of Ceará, CE, Brazil
Email: danielamaral@alu.ufc.br

Abstract

Bees are essential for food production for humans and for the maintenance of natural ecosystems. This paper presents a proposal to predict the health level of honeybee colonies using data from internal and external beehive sensors and from in-loco inspections by beekeepers. The data set was obtained by gathering inspection information and internal and external sensors measurements, based on the date of collection. However, obtaining inspection data frequently is not feasible due to the stress caused to the beehive, especially in periods such as winter, where the beehive becomes more sensitive. As a solution, the beehives health status was obtained through a partitioning clustering method and then validated by in-loco inspection data already obtained. We propose a logistic regression model with an elastic net penalty, which consists of a fusion of lasso (l1) and ridge (l2) methods. We obtained a flexible and robust model compared to the usual logistic regression and a diagnostic tool that can avoid unnecessary inspections and, consequently, reduce the stress of the beehives.

CPS1.3: Scikit-learn: Machine learning in Python

Robert Sebastian Castellanos Rodriguez[1]

[1]National university of Colombia, Colombia
Email: rscastellanosr@unal.edu.co

Abstract

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language. Emphasis is put on ease of use, performance, documentation, and API consistency. It has minimal dependencies and is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings.

CPS2: Data science applications to the society

CPS2.1: Indicators for public policy based on the use of text mining and machine learning: a case study of the Brazil's Lava Jato Operation

Douglas Farias Cordeiro[1], Kátia Kelvis Cassiano[1] and Núbia Rosa Da Silva[1]

[1]Federal University of Goiás, GO, Brazil
Email: cordeiro@ufg.br

Abstract

The development of indicators that allow the evaluation of the impacts of public actions and policies in reference to the benefits provided to the citizen is fundamental in the measurement of government success, and as a strategy for gathering demands. Although there are several channels of communication between government and society, citizen involvement is still a challenge to be overcome. On the other hand, there is a remarkable growth in access to information and communication technologies. According to data provided by Cetic.br, in 2017, 92% of Brazilian households had cell phones and 73% had internet access. According to the Statista Portal, in July 2019, Brazil presented more than 8.2 million active users in the social network Twitter. This scenario provides an opportunity to explore social networks as a source of strategic information for public administration. From this, this paper proposes to present a strategy based on the Knowledge Discovery in Databases (KDD) process to develop a solution focused on continuous monitoring of social networks, in this case specific to Twitter, focusing on public administration and determination of a data model oriented towards purposes of indicator generation. The proposal is based on the application of Natural Language Processing (NLP) methods to group the collected data, and on the application of sentiment analysis techniques based on the use of Naive Bayes statistical learning algorithms. The solution is applied to a case study focused on Operation Lava Jato, of the Brazilian Federal Police.

CPS2.2: Rise of the young women in Data Science: Data literacy in public schools

Karla Patricia Oliveira-Esquerre (coordinator)[1,11,12,13], Adriana Santana[1,12], Anna Cláudia Furtado[1,12], Ana Luiza Nogueira[1,12,13], Caroline Fernandes[1,12,13], Daniele Lima[1], Elaine Albuquerque[1,11], Gloria Meyberg[1,11], Graziela Santana[1,12,13], Herica Oliveira[1,11,12,13], Isabela Almeida[1,12,13], Jorge Ubirajara Pedreira Junior[1,12], Júlia Bijos[1,11,12,13], Julia Mosselman[1], Laís Bastos Pinheiro[1], Lília Meira[1], Lorenna Vilas Boas[1], Mariana Amorim[1], Rafaela Menezes[1,12,13], Raony Fontes[1,11], Rosana Fialho[1,11], Sandra Pinheiro[1,11,12], Silvia Miranda[1], Talita Costa[1], Thayne Sodré[1], Adonias Magdiel[1], Alana Almeida[1,10,11], Alice Rocha[1], Luciana Martinez[1], Márcio Martins[1,11], Leizer Schnitman[1], Reijane Gomes da Silva[1], Rejane Santos[1,11,12,13], Tatiana Dumêt[1], Roseline Oliveira[2], Livia Fialho Costa[3], Gabriela de Queiroz[4], Bruce Kent[5], Maíra Oliveira Esquerre[5], Luna Oliveira Esquerre[5], Allena Lyra Araújo[6], Ana Rosa Iberti[7], Alzira Nascimento Conceição de Melo[8], Daniela Borges Lira e Silva[7], Maysa Conceição Cavalcante Lima[9], Claudia Virginia de Santana Cajado[8], Maria Alice Rocha[6], Cecília Peixoto da Silva[14], Maria Lucia de Souza Oliveira[14], Gláucio André Barbosa Gaza[9]

[1] Federal University of Bahia, BA, Brazil
[2] Federal University of Alagoas, AL, Brazil
[3] Bahia State University, BA, Brazil
[4] 4 R-Ladies Global
[5] Torrey Pines Elementary School, United States
[6] Colégio Estadual Evaristo da Veiga, BA, Brazil
[7] Colégio Municipal Cidade de Jequié, BA, Brazil
[8] Colégio Estadual Henriqueta Martins Catharino, BA, Brazil
[9] Colégio Estadual Ypiranga, BA, Brazil
[10] Colégio Estadual Mário Costa Neto, BA, Brazil
[11] Programa de Pós-Graduação em Engenharia Industrial - Federal Univesity of Bahia, BA, Brazil
[12] Gamma - Federal Univesity of Bahia, BA, Brazil
[13] R-ladies Salvador, BA, Brazil
[14] Laboratório de Petróleo e Gás (LAPEG) - Federal University of Bahia, BA, Brazil
Email: karlaesquerre@ufba.br

Abstract

In Brazil, women participation in the production of knowledge, education and labor market related to fields of science, technology and innovation is inferior to the men one. Statistics have shown that the number of male and female undergraduate students in exact courses, like engineering, is sometimes the same, nevertheless, the second are still minorities in research and extension projects in university that require programming or computational skills. When speaking about basic education, in private schools there is still incipient learning in programming while in public schools this practice is practically nonexistent. Disconnected with school subjects, some increasing in computing literacy as informal education has been observed owing to the technological characteristic of the 21st Century, which has favored the formation

of digitally active children, highly seduced by technology. However, these children do not have computational and analytical thinking that may allow them to analyze and criticize daily life data. Consequently, that's a lack of ability to produce knowledge and technology from images, texts, data and sounds. This panorama is even worst to girls since they are seen as home caregivers or without technological skills and are often kept at home when they reach puberty. The research project entitled Gender Diversity in Data Science: Learning though Experimentation (popularly known as Girls on Data Science) was awarded by the CNPq/MCTIC Nº 31/2018 call (Girls in Exact Sciences, Engineering and Computing), in December 2018, with a maximum grade (10.0) in all items evaluated except for the one related with the proposal of the project continuity (wigrade 9.0). This project aims to introduce data science as the fourth pillar of the scientific method concepts by working on data analytics, statistical and computational thinking with about 500 girl students (12 to 17 years old) of five public schools located in low-income neighborhoods in Salvador city, Bahia State - Brazil. These girls, mostly black, face barriers to education caused by poverty, cultural norms and practices, inadequate infrastructure leading to poor learning environments, urban and domestic violence and many other fragilities, like hunger, racism and discrimination. The team responsible for carrying out the project is multidisciplinary and is composed of professors and professionals of Exact Sciences (Engineering, Statistics, Mathematics) and Human Sciences (Psychology, Pedagogy, Anthropology, Design and Architecture). The project is being carried out based on three lines of action: (a) Approaching the School; (b) Formative and Informative Process and (c) Connection between Theory and Practice. Data exploration and analysis are been used considering statistics and graphs in a creative way to the development of science projects and of a website about Salvador. The current activities also focus on the development of a multidimensional material for educators and students on data science and computational thinking education (computational, statistical and critical thinking, data-driven problem solving, creativity, and digital literacy). This project contributes to the approximation or reaffirmation of the partnership between public schools of Basic Education and Institutions of Higher Education; and to deepen students' knowledge about their city and community, which favors their social protagonism.

CPS2.3: Spatial variability of rainy season onset characteristics across the São Paulo state, Brazil

Aline Maia[1], Mári Firpo[2], Ana Ávila[3], Caio Coelho[2]

[1] Embrapa, SP, Brazil
[2] CPTEC/Inpe, SP, Brazil
[3] Cepagri/Unicamp, SP, Brazil
Email: aline.maia@embrapa.br

Abstract

Planning of economic activities, which are influenced by local pluviometry, depends on forecasts related to rainfall amount and its temporal distribution. In agriculture, for example, definition of crops calendar requires information on temporal patterns of the rainy season onset. In the SP state, Brazil, the onset occurs between late September and early October, with high variability among stations or years. Daily data (1961 - 2013) from 129 stations from São Paulo state were used in this study. The time to onset of rainy season (T, days after 01/07) was determined for each year and station, by Liebmann criteria. The T empirical distribution (n=53), for a particular station, represents the local onset interanual variability. Quartiles of such distributions were therefore used to perform cluster analysis, resulting in eight groups. The following spatial patterns were observed: two groups located at the inferior and superior pieces of the NW-SE diagonal of the state, respectively; another group, without spatial continuity, placed at parallel extreme strips, above, and below the NW-SE diagonal while two other groups, concentrate at extreme SE. Jaborandi and Iguape stations constitute isolated points, showing patterns, which are extremely different from all the others groups. Results show high onset space-temporal variability: median T varying from 100.5 (9/10) to 155.5 days (02/12) among stations and, in the stations with maximum interanual variability, the T range was 171 days. The similarity among patterns is related, in general, to stations spatial distribution, which allows delimitation of regions where similar patterns of onset variability are predominant.

CPS3: Data science in practice

A proposed method to objectively identify synchronic spatial domains of chlorophyll a in the marine region of the Amazon coast using remote sensing data: A multivariate approach.

Eduardo Paes[2], Nelson Gouveia[1], Elton Correa[2] and Jeandria Freire[2]

[1] INPE
[2] Federal Rural University of Amazonia, AM, Brazil
Email: etpaes@gmail.com

Abstract

Here we present an original proposal to objectively identify the synchronic spatial domains (SSD) of chlorophyll-a for the coastal and Oceanic region of the Amazon coast. SSD is defined as the spatially coherent regions where seasonal variations of chlorophyll-a present similar patterns. The variability of chlorophyll-a concentration over a year is an estimator of oceanic primary productivity and is modulated by several highly complex biogeophysical processes. Also, play a key role in important mechanisms, for example, in the control and limits of fishing productivity, as well as for estimates of carbon sequestration rates. All these processes respond to climatic variations. We used chlorophyll-a concentration data estimated from Moderate Resolution Imaging Spectroradiometer aboard the Aqua satellite for the years that occurred canonic El Nino(2016), La Nina (2008), El Nino Modoki in the positive(2010) and negative phases (2011) and a Neutral year (2013). The worked matrix had 7500 lines per 12 columns, which was submitted to a protocol that combined empirical orthogonal function and K-means classification analysis. The spatialization results showed 17 groups of SSD. Details and applications of this methodology will be discussed during the presentation.

CPS3.2: Forecasts, Brazilian GDP projections and time series inference in the high-dimension and irregular-time span context

André Maranhão[1]

[1]Catholic University of Brasília, DF, Brazil
Email: andrenmaranhao@gmail.com

Abstract

The issue of High Dimension is the situation where the number of time series is greater than the time interval. This context is characterized by the new availability of large databases, or Big Data. Data science has developed strongly to address the cross-section approach to data, yet temporal data is still evolving. Linear predictions (predictions), however best the model may have, have presented in many practical applications large predictive ranges, to address this issue the literature has used the combination of linear and nonlinear predictions. The nonlinear prediction consists of estimating the cyclic component, in High-Dimension this challenge becomes even greater given that the time series has interval-irregular time. In the present article, we present a solution for estimating the High-Dimension and irregular-interval cyclic components for the Brazilian macroeconomic data, allowing to generate projections of the High-Dimension Brazilian GDP growth.

CPS3.3: Machine learning for health management: A breast cancer case study in Brazil

Matheus Carvalho[1], Katia Cassiano[2] and Barbosa Talles[1]

[1] Pontifical Catholic University of Goias, GO, Brazil
[2] Federal University of Goiás, GO, Brazil
Email: maheuscarv@gmail.com

Abstract

Data science is shown as the field of study with the potential for discovering useful information from a large volume of data that, once interpreted, generates knowledge for decision making. In health, it was used in the automation of data storage processes, collection and extraction of characteristic patterns that allow analysis. Automated health data analysis supports decision-making and can assist in diagnosing and implementing public health policies for continuous improvement. Therefore, the development of health information management solutions is feasible, as the information extracted is necessary and may include inputs for continuous observation of scenarios and analysis of effective policies and actions. According to data from the Global Cancer Observatory (GCO), breast cancer had a higher incidence in 2018 in Brazil. The Department of Informatics of the Unified Health System (DATASUS) is responsible for maintaining the collection of SUS databases. Among these bases it is possible to find several data related to the health area. This scenario ultimately provides an opportunity for data exploration as a source of strategic information for health management. From this, this paper presents a strategy based on the Knowledge Discovery in Databases (KDD) method to develop an analysis solution aimed at understanding the characteristic patterns of breast cancer in Brazil through the application of exploratory descriptive analysis and predicting morbidity, a solution contemplated in the implementation risk classification models. The solution is applied to a case study based on the diagnosis of breast cancer provided by DATASUS.

CPS4: Recent advances in regression and longitudinal models

CPS4.1: Influence diagnostics in mixed effects logistic regression models

Alejandra Andrea Tapia Silva[1]

[1]Universidad Católica del Maule, Chile
Email: alejandraandreatapiasilva@gmail.com

Abstract

Correlated binary responses are commonly described by mixed effects logistic regression models. This article derives a diagnostic methodology based on the Q- displacement function to investigate local influence of the responses in the maximum likelihood estimates of the parameters and in the predictive performance of the mixed effects logistic regression model. An appropriate perturbation strategy of the proba- bility of success is established, as a form of assessing the perturbation in the response. The diagnostic methodology is evaluated with Monte Carlo simulations. Illustrations with two real-world data sets (balanced and unbalanced) are conducted to show the potential of the proposed methodology.

CPS4.2: Piecewise-linear regression with smooth phase-transitional functions by the assumption of random thresholds

Iuri Ferreira[1] and Silvio Zocchi[2]

[1] Federal University of São Carlos, SP, Brazil
[2] University of São Paulo, SP, Brazil
Email: ferreira.iep@gmail.com

Abstract

In this study, multi-phase processes with random thresholds were modeled by piecewise-linear regression with smooth phase-transitional functions. A new parametrization of max-min segmented models was proposed, extending the bent-cable models of Chiu et al. (2006) for situations with more than two linear phases. Also, based on the assumption of a mixture of processes and random thresholds, two new families of phase-transitional functions were derived to replace the original 'bent-cable'. The modeling approach was flexible enough to describe the complex behavior shown by multi-phase data sets. However, has been verified that the gathering of information about the random thresholds is a complicated task which depends on a large sample effort and designs with many points of support. Our modeling approach was applied in a practical problem of agriculture, to provide a complete description of the mineral deficiency curve. This curve is generally described as three linear nutritional phases (the mineral deficiency, luxury consumption, and toxicity) joined by smooth transitional sections (sufficiency and toxicity ranges). By fitting the proposed models to the agronomic data, they were obtained estimates for the thresholds (critical levels for deficiency and toxicity) and their variances. Thus, it was possible to make inferences about the transition ranges (for sufficiency and toxicity).

CPS4.3: Joint models of longitudinal data with informative time measurements

Ines Sousa[1]

[1]University of Minho, Portugal
Email: isousa@math.uminho.pt

Abstract

In longitudinal studies individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However, in medical context it is common that patients in worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristic should allow for an association between longitudinal and time measurements processes. We consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. Estimation of model parameters is through maximum likelihood. We conducted a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set when monitoring for biomarkers CEA and CA15.3 on breast cancer progression. In this case the follow-up time process should be considered dependent on the longitudinal outcome process.

CPS4: Recent advances in statistics and data science

CPS5.1: A computational model for after-sales services

Henrique Zaidan[1], Bruno Guedes[1] and Claudio Cristino[2]

[1] PPGBEA, Federal Rural University of Pernambuco, PE, Brazil
[2] PPGBEA, Federal University of Pernambuco, PE, Brazil
Email: santos.henrique624@gmail.com

Abstract

This paper aims to explain a quantitative model of after-sales services through a framework based on game theory and discrete events simulation. It shows a new approach to Murthy and Asgharizadeh's model (1998). The authors originally developed a theoretical decision problem between a service agent and a customer, where the buyer is the owner of the equipment and outsources maintenance to a service agent that presents two maintenance options, a service contract, and services on demand (no contract), one of which must be selected by the customer if he decides to buy the equipment. The main assumptions are: the product fails at a constant rate, constant repair costs, the customer is risk-averse, and the parties have perfect information about the model parameters. Once the model was set up as a Sequential Stackelberg game, its elements were defined: number of players, set of strategies, the order of moves and payoff functions. The optimal solutions for each player were obtained analytically, determining the player's optimal decisions and the equilibrium. The main contribution of this new approach will be to use computational techniques to simulate the random variables via the Monte Carlo method. It is important to emphasize that this kind of decision problem holds random features since the equipment is an unreliable system. Thus, equipment failures happen randomly and it is necessary to compute reliability-related performance measures that affect the players' decisions and expected payoffs. Finally, this study also presents a sensitivity analysis of the random variables and model parameters.

CPS5.2: Finding optimal classification rules using decision diagrams

Lucas Braga de Albuquerque[1] and Tiberius Oliveira Bonates[1]

[1]Federal University of Ceará, CE, Brazil
Email: lucas9ba@gmail.com

Abstract

In rule-based classifiers, a central task is finding rules that are descriptive of large subsets of the training data. Let D be a dataset consisting of binary data, and having positive and negative observations. A positive rule is a subcube having a nonempty intersection with the positive part of D, and an empty intersection with the negative part of D. A negative rule is defined analogously. An observation is covered by a rule if it belongs to the corresponding subcube, and the coverage of a rule is simply the number of observations in D covered by it. The maximum x-rule problem consists in finding a rule whose coverage is maximum among those that cover a given observation x in D, which amounts to maximizing a nonlinear function subject to set covering constraints. We model and solve the maximum x-rule problem using a recently-developed optimization methodology based on decision diagrams (DD). We compare the performance of our DD-based solver with that of two mixed integer linear programming (MILP) approaches from the literature. Our results indicate that a straightforward DD-based branch-and-bound implementation provides higher quality solutions to medium-sized datasets than a commercial MILP software within a common time limit. Furthermore, a DD-based approach provides a large pool of feasible rules (as opposed to a single rule, as in a MILP-based approach), which is an advantageous feature when building rule-based classifiers.

CPS5.3: Joint modelling of survival and multivariate longitudinal data in clinical research

Denisa Mendonça[1], Anabela Rodrigues[2], Inês Sousa[3] and Laetitia Teixeira[4]

[1] Abel Salazar Biomedical Sciences Institute (ICBAS) - University of Porto & Institute of Public Health (ISPUP), Portugal
[2] Abel Salazar Biomedical Sciences Institute - University of Porto & Departament of Nephrology, Centro Hospitalar do Porto, Portugal
Email: dvmendon@icbas.up.pt

Abstract

In many clinical studies, several information about the patients are collected, namely on baseline characteristics, longitudinal repeatedly registered biomarker and the time to a specific outcome event. In these situations, it is relevant to simultaneously analysis such information and a joint modelling approach needs to be considered. One of the main recent extension of the classical approach is the inclusion of multiple longitudinal outcomes, allowing the benefit of harnessing different and relevant information in a single model. The main objective of this work is to present and applied a joint modeling approach for time-to-event and multivariate longitudinal data to evaluate a peritoneal dialysis program. We fit a joint model with a bivariate longitudinal submodel (albumin and log transformation of creatinine as longitudinal outcomes) and a time-to-event submodel (death as event of interest). For albumin, time, sex, age groups and provenience were considered as fixed effect and for log-creatinine, time, sex and age groups. For both, random intercept and random slope were also considered. Considering time-to-event submodel, sex, age and diabetes are considered as covariates. The fitted model indicated that an increase in the subject-specific random deviation from the population trajectory of log-creatinine was significantly associated with increased hazard of death. Additionally, a significant association was also observed for subject-specific decreases in albumin from the population mean trajectory. In conclusion, joint modelling for multivariable longitudinal and time-to-event outcomes is useful in different areas of applications when the interest is the evaluation of the relationship between these two types of outcomes.

Part IX

Contributed Posters

Poster Session

CP1: Machine learning applied to the study of tuberculosis in Brazil

Adelmo Inácio Bertolde[1], Luiz Henrique Quinelato[1], Carolina Martins Sales[2] and Ethel Leonor Maciel[2]

[1] Department of Statistics, Federal University of Espirito Santo, ES, Brazil
[2] Nursing Department, Federal University of Espirito Santo, ES, Brazil
Email: adelmoib@gmail.com

Abstract

In this paper we address the issue of tuberculosis in Brazil. More specifically, we discuss cases that result in death or treatment dropout (“unfavorable cases”) and those that result in cure (“favorable”). In order to better understand which factors impact such outcomes, the objective of this work is to find a classification model that can discriminate the class of unfavorable. Firstly, a case database of tuberculosis patients in Brazil was organized, from 2013 to 2016, of about 350,000 notified cases, which includes a set of characteristics of each case. The characteristics and assumptions of each model (algorithm) were also analyzed in order to understand advantages and disadvantages in using such technique. After applying the classification models, the best results for Brazil were the Logistic Regression and Naive Bayes models. In the case of the first, the results were 0.746 for sensitivity, 0.703 for specificity and Accuracy of 0.738. In particular for the state of Espírito Santo, the results for this same technique were 0.746 for sensitivity, 0.710 for specificity and 0.739 for accuracy. Given the results, we can conclude that we found a satisfactory model for predicting the outcome of patients with tuberculosis under treatment, considering a binary outcome model.

CP2: Correção de viés das estimativas dos parâmetros de uma nova extensão da distribuição Nadarajah-Haghighi

Alice Buarque Vieira de Mello[1], Maria Lima[1] and Tatiane Ribeiro[1]

[1]Federal University of Pernambuco, PE, Brazil
Email: alicebuarque31@gmail.com

Abstract

O trabalho apresenta uma nova extensão da distribuição Nadarajah-Haghighi (NH) através da família Marshall Olkin extended-G (MOE-G), proposta por Cordeiro et al. (2019). A nova família de distribuições apresenta quatro parâmetros, dois de forma pertencentes a família MOE-G e dois oriundos da distribuição base, sendo de escala e forma. Todos os parâmetros positivos e suporte da distribuição nos reais positivos.

Uma das vantagens deste novo modelo probabilístico é que sua função distribuição acumulada (fda) possui forma fechada e independente de funções especiais. Este fato viabiliza a geração de ocorrências pseudo-aleatórias com esta distribuição pelo método da inversão, uma vez que a função quantílica possui forma explícita. Consequentemente, usamos este método para obter amostras pseudo-aleatórias e realizar a simulação de Monte Carlo. Os parâmetros foram estimados via aplicação do método da máxima verossimilhança (MV) em uma rotina computacional elaborada em R. Os estimadores de MV possuem excelentes propriedades assintóticas, mas, em geral, para pequenas amostras os vieses podem ser consideráveis. Neste caso, para reduzir o viés das estimativas de MV obtidas para os quatro parâmetros do modelo proposto realizou-se a aplicação do método bootstrap. Embora o custo computacional da aplicação deste método tenha sido elevado, obtemos boas estimativas corrigidas para todos os parâmetros e cenários considerados.

CP3: Regulatory milestones facing “The Evitable Conflict” - human noises in Data Science from Isaac Asimov

Bruna de Alencar Carvalho[1]

[1]Bahia State University, BA, Brazil
Email: bruna_gcc@hotmail.com

Abstract

This paper aims to present the perspectives of data science in view of the regulatory function of national and international law. Therefore, it is proposed an overview of historical-current understanding from global, national and local levels regarding the flow of data from geopolitical concepts of center and periphery, developed countries and developing countries; and for the field of prospecting, open science is presented from the social-communicational complexity of Niklas Luhmann and Thomas Kuhn’s “The Structure of Scientific Revolutions”. Thus, for contextual assimilation of data science as a field of knowledge based on the human-machine interface, Isaac Asimov’s tale entitled “The Evitable Conflict” is used to understand the articulation of human knowledge in data management and conceive the possible directions for the area. The results, in turn, show the human preponderance in the interpretation and use of data and, consequently, of the noise inherent to individual interests that demand ethical control and collective supervision.

CP4: A machine learning based approach to classify real-time optimization executions

Caroline Fernandes[1] and Karla Patrícia Santos Oliveira Rodriguez Esquerre[1]

[1]Federal University of Bahia, BA, Brazil
Email: carolinefernandes.eq@gmail.com

Abstract

Real-Time Optimization (Real-Ttime Optimization – RTO) is a powerful tool from research optimal set point for continuous processes, that results in big economic benefits. However, there are little explored questions about its performance, that require the application of more sophisticated techniques, such as Machine Learning, for information mapping and solution identification. This paper aims to compare the performance of five Machine Learning algorithms in the classification of convergence and non-convergence scenarios of RTO runs, from the process variables of a typical petroleum destilation unit. For this, the multiple imputation technique with regularized PCA was applied to missing data and K-means and descriptive statistics during the exploratory data analysis. Classification algorithms such as Random Forest, Bagging, Boosting, kNN and SVM were used to classify observations. The statistical techniques allowed to explore the behavior of the process variables, as well as their possible groups, through K-means. Finally, through the confusion matrix it was possible to verify that the classification model that presented the best performance was the Random Forest, and from this, it was identified the most important variables that contributed to the best performance of the classification model.

CP5: Early monthly estimation of manufacturing activity level using electric energy consumption data

Daniel Alba-Cuellar[1] and Hugo Hernandez-Ramos[1]

[1]National Institute of Statistics and Geography, Mexico
Email: daniel.alba@inegi.org.mx

Abstract

Directly measuring the monthly Manufacturing Production Level in Mexico via national accounting methods is an elaborate process, yielding a preliminary figure approximately 40 days after the end of the reference month. A separate analysis conducted by INEGI (Mexico's National Statistical Office) showed that in Mexico's manufacturing sector, in principle there exists a significant linear relationship between electric energy consumption and Production Level. Currently, electric energy consumption data from the Federal Electricity Commission (CFE) are made available to INEGI approximately 15 days after the end of the reference month; this timeliness in the availability of CFE data, combined with the observed relationship, allowed INEGI to build an econometric model which produces early estimates for the Manufacturing Production Level Index just 20 days after the end of the reference period. In this work we describe the initial analysis conducted by INEGI to determine the relationship between both variables; then we describe the characteristics and evolution of the established econometric model, and compare early estimates, computed in real time, against officially published values, as an empirical means for evaluating early estimation accuracy. We observed that 93% of the time, the official value is located inside the prediction interval, which was computed with a 95% confidence level; this means that, in this case, observed empirical accuracy approached the theoretical confidence level. Finally, in this work we comment about INEGI's data sharing experience with CFE, and talk about future steps to improve this nowcasting process.

CP6: Nonlinear time series analysis: Unfolding complexity in neuroscience

Diego Nascimento[1,3], Francisco Louzada[1,3] and Taiza Santos[2,3]

[1] Institute of Mathematics and Computer Sciences - University of Sao Paulo, SP, Brazil
[2] Medical School of Ribeirão Preto - University of Sao Paulo, SP, Brazil
Email: dnstata@gmail.com

Abstract

In the medical field it is common to use the statistical mean as a summary statistics for the observed time series and then perform standard statistical test. However, this procedure is only valid relying on the assumption that the time series is stationary, which often is not the case. Taking this into account, this work discusses the use of entropy as a measurement of time series complexity, in the context of Neuroscience, due to the non-stationary character of the data. Additionally, others Machine Learning techniques were also adopted to rehearse the isometric space between the proposed treatment (tDCS) against the resting-state (basal reference). We also elucidate our discussion regarding the usage of entropy to analyze recorded data from EEG signals, targeting to test its electrical accommodation tDCS dose-response, and brought elements to establish a safety protocol test to address human verticality (as a human treatment post-stroke) through the application of electrical stimulation of the brain.

CP7: Goodness-of-fit using nonparametric full Bayesian significance test

Djidenou Montcho[1], Rafael Izbicki[1] and Luis Salazar[1]

[1]Federal University of São Carlos, SP, Brazil
Email: hansamos@usp.br

Abstract

Many statistical analysis require the ability to test whether a sample comes from a given distribution, a goodness-of-t test. Although much work exists on performing this test under a frequentist framework, there are almost no attempts for tackling this problem under a Bayesian perspective. The reason is that the null hypothesis is precise (sharp), which created difficulties for most Bayesian methods. In this work we show how the Full Bayesian Signicance Test, a framework for testing parametric sharp hypothesis, can be extended to nonparametric models by using pseudo-densities. We compare our method to the classical t-test, Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von Mises tests in simulated datasets, and show that our method performs better and equally well to some of them.

CP8: Text analytics on open ended questions data: A new analysis alternative in opinion surveys

Felipe González[1]

[1]University of Costa Rica, Costa Rica
Email: felipe.2408@hotmail.com

Abstract

The process of manually coding open-ended questions is laborious and involves a loss of information, text mining applications offer an alternative that facilitates the analysis of the data extracted from the answers in the case of open-ended questions in opinion surveys. The purpose of this study was to provide a description of the use of text mining for exploratory and predictive analysis. Twelve open-ended questions from the 2019 National Transparency Perception Survey were used. The methods to analyze the text were shown in this work, such as the necessary steps to clean the text, as well as to explore the text data through an analysis of frequencies, networks, and clusters. It was intended to show how the task of automatic coding of open-ended questions can be posed as a problem of multi-class categorization through supervised machine learning. Also, it was shown in this work how to remove words that add noise in the models and how to select the most appropriate model using cross-validation. The algorithms used were support vector machines, naive Bayes classifier, random forest, XGBoost and k nearest neighbor. It was found that the results obtained through exploratory techniques from text mining are like the ones that were manually coded. About the text classification, the precision of the models for each of

the questions lies between 48% and 76%. Likewise, it was shown that the categories predicted by the models chosen for each question allow to establish similar results compared to those obtained with the pre-established categories.

CP9: On extended negative binomial distribution for count data with inflated frequencies

Ian Jay Serra[1] and Daisy Lou Polestico[2]

[1] University of the Philippines Cebu, Philippines
[2] MSU-Iligan Institute of Technology, Philippines
Email: iaserra@up.edu.ph

Abstract

This study extends the existing zero-inflated distributions through the flexibility of peaks in the data with excessive counts other than zeros and handles an overdispersion in the data. Moreover, this study formulates a proposed zero and k inflated negative binomial (ZkINB) distribution which is a mixture of a multinomial logistic and negative binomial distribution. The multinomial logistic component captures the occurrence of excessive counts, zero and $k \geq 1$, while the negative binomial component captures the counts that are assumed to follow a negative binomial distribution. Furthermore, this study derived the moment generating function of the distribution in order to solve some structural properties of the ZkINB, including the mean, variance, and the skewness and kurtosis. Two real datasets were used to analyze the characteristics of the ZkINB. In both examples, we used zero and k inflated negative binomial (ZkINB) distribution and compare it to the zero and k inflated Poisson (ZkIP) and zero-inflated count distributions. The first example illustrated a ZkINB distribution with inflations at 0 and k = 3, while the second example has inflations at 0 and k = 1. As a result, the zero and k inflated negative binomial distribution seems to exhibit a better fit than the inflated NB and POI count data distributions.

CP10: New statistical process control chart for overdispersed count data based on the Bell distribution

Laion Lima Boaventura[1], Paulo Henrique Ferreira da Silva[1], Rosemeire Leovigildo Fiaccone[1], Pedro Ramos[2] and Francisco Louzada[2]

[1] Federal University of Bahia, BA, Brazil
[2] CeMEAI, São Paulo University, SP, Brazil
Email: englimaboaventura@gmail.com

Abstract

Poisson distribution is a popular discrete model used to describe counting information, from which traditional control charts involving count data, such as the c and u charts, have been established in the literature. However, several studies recognize the need for an alternative control chart that allows for data overdispersion, which can be encountered in many fields including ecology, healthcare, industry and others. The Bell distribution, recently proposed by Castellares *et al.* (2018), is a particular solution of a multiple Poisson process able to accommodate overdispersed data, and thus can be used as alternative to the usual Poisson (which, although not nested in the Bell family, is approached for small values of the Bell distribution) and negative binomial distributions for modeling count data in several areas. Therefore, in this paper we consider the Bell distribution to introduce a new, interesting and useful statistical control chart for counting processes, which is capable of monitoring count data with overdispersion. The performance of the so-called Bell chart is evaluated by the average run length in numerical simulation. Some real-life and simulated processes/data sets are used to illustrate the application of the proposed chart.

CP11: Using random forest to predict the BOD5 of an aerated lagoon at a pulp and paper mill

Lucas Mascarenhas Almeida[1], Brenner Biasi Souza Silva[1] and Karla Patrícia Santos Oliveira Rodriguez Esquerre[1]

[1]Federal University of Bahia, BA, Brazil
Email: lucasmascalmeida@gmail.com

Abstract

The treatment and discharge process of effluent at an industry is a process very important to attend the environmental requirements. The aim of wastewater treatment systems is reduce the amount of pollutants that the effluent has for the wastewater be able to discharge adequately. The monitoring of water quality parameters is very important for treatment processes, because they translate the principal proprieties of effluent. One of the main effluent quality parameters in the pulp and paper industry is the biochemical oxygen demand, however the analysis to obtain BOD usually takes five days (BOD5). The aim of this work is to propose a monitoring alternative for BOD5, for this a predictive model is produce using Random Forest. The aim of this work is to propose a monitoring alternative for BOD5, so a predictive model is produced using Random Forest as an alternative. In this work the data are from monitoring a wastewater treatment plant in a pulp and paper industry. The database contains the main water quality parameters such as biochemical oxygen demand, chemical oxygen demand, pH and etc. The data are diaries. The BOD5 model presents a $RMSE = 31.09$ and $R^2 = 0.514$. The mainly predictor to model develop is the COD. Although of the difficulties in relation the data missing problems and the complexity of modeling a bio process vulnerable to external factors, as rains and temperature variation, applications of machine learning algorithms, such as Random Forest, can be good tools for monitoring important variables in industrial processes.

CP12: Alcohol consumption, smoking and psychiatric comorbidity’ s relationship study

Marcela Portela Santos de Figueiredo[1]

[1]Federal Rural University of Pernambuco, PE, Brazil
Email: portela.marcela.producao@gmail.com

Abstract

Alcohol and tobacco are commercially released in Brazil and their consumption represents risk factors associated with several diseases.This study aimed: 1. Analyze the relationship between alcohol consumption and tobacco consumption; 2. Analyze the relationship between depression, and the consumption of alcohol and/or tobacco. Then, we analyzed microdata from the National Health Survey- 2013 (most recent edition).We analyze a representative sample for adults with a size of 60,202 individuals older than 18 years.Nonlinear Factorial Analysis was used, and two main components were formed: mental health and consumption of licit drugs. The choice of non-linear analysis is justified by the fact that the variables investigated are qualitative. The mental health component is compounded by ordinal variables: feeling bad about yourself or feeling like you’ve failed or feeling you’ve disappointed your family, feeling depressed or out of perspective, and thinking about getting hurt or being dead. The second component is composed of alcohol consumption and tobacco consumption.As a result of the categories’ analysis, there is a relationship between these symptoms: Brazilians who think they are a failure, or feel bad about themselves or who have disappointed their family, usually think about hurting themselves or prefer to be dead and feel depressed or without perspective. We also found the relationship between alcohol and tobacco consumption, so that Brazilians who smoke, usually drink, that is, the consumption of these two drugs are associated in Brazil population.There was no association between psychiatric symptomatology and alcohol or tobacco consumption in the sample studied through this type of analysis.

CP13: Analysis of sunflower data from a multi-attribute genotype x environment trial in brazil

Marisol García Peña[1], Sergio Arciniegas Alarcón[2] and Kaye Basford[3]

[1] Pontificia Universidad Javeriana, Colombia
[2] Universidad de La Sabana, Colombia
[3] School of Agriculture and Food Sciences, The University of Queensland, Australia
Email: luzmara@gmail.com

Abstract

In multi-environment trials it is common to measure several response variables or attributes to determine the genotypes with the best characteristics. Thus it is important to have techniques to analyse multivariate multi-environment trial data. The main objective is to present two multivariate techniques: the mixture maximum likelihood method of clustering and three-mode principal component analysis, to analyse jointly genotypes, environments and attributes. In this way, both global and detailed statements about the performance of the genotypes can be made, highlighting the utility of using three-way data in a direct way and providing an alternative analysis for researchers. We illustrate using sunflower data with twenty genotypes, eight environments and three attributes. The procedures provide an analytical procedure which is relatively easy to apply and interpret in order to describe the patterns of performance and associations in multivariate multi-environment trials.

CP14: Produção científica, tecnológica e acadêmica dos programas de pós-graduação da Universidade Federal da Bahia representada na plataforma Lattes do CNPq

Natanael Vitor Sobral[1], Brenda Barbara Costa Ribeiro[1] and Valdinei Silva de Souza[1]

[1]Federal University of Bahia, BA, Brazil
Email: natanvsobral@gmail.com

Abstract

O presente trabalho vale-se de técnicas relacionadas à Ciência de Dados para a coleta, produção e análise de dados sobre a produção científica, tecnológica e acadêmica da Universidade Federal da Bahia (Ufba), especificamente, os Programas de Pós-Graduação Stricto Sensu. Para isto, fez-se uso da Plataforma Sucupira da Capes, identificando 76 programas com 1663 docentes vinculados. Em seguida, obtiveram-se dados relativos à produtividade científica, tecnológica e acadêmica destes pesquisadores a partir da ferramenta ScriptLattes, que extrai e compila dados da Plataforma Lattes do CNPq. Adiante, os dados foram estruturados em formato de texto padronizado, denominado “arquivo bibliométrico”, para que se

realizasse a exploração deste conjunto de registros. As técnicas aplicadas envolveram o software “The Vantage Point”, que possibilita atividades de mineração de textos e dados, favorecendo a criação de matrizes matemáticas, rankings e processamento de linguagem natural; VOSViewer, Netdraw, Ucinet e Gephi com o propósito de representar redes com os fundamentos da teoria dos grafos e Planilha de cálculo para a produção de estatísticas gerais . Com isto, obtiveram-se dados da produção de artigos, livros, trabalhos em evento, orientações, redes de colaboração científica, produções técnicas, bolsas do CNPq, organização e participação em eventos, entre outros. Estes indicadores estão sendo reunidos para a publicação em um site que pretende divulgar anuários estatísticos da produção científica da Ufba, dando conhecimento destes dados à comunidade acadêmica, aos tomadores de decisão e a toda sociedade que se interessar por este conjunto de informações, inclusive, cientistas de dados que trabalhem com reuso.

CP15: Predictive modeling of economic and financial monitoring of the Brazilian electricity regulatory agency (ANEEL)

Renato Panaro[1], Silvio Patricio[1], Vinícius Mayrink[2] and Marcelo Costa[3]

[1] Federal University of Minas Gerais, MG, Brazil
[2] Department of Statistics - Federal University of Minas Gerais, MG, Brazil
[3] Department of Production engineering - Federal University of Minas Gerais, MG, Brazil
Email: renatovp@id.uff.br

Abstract

According to the The Brazilian Electricity Regulatory Agency (ANEEL), the economic-financial indicators of the electricity distribution companies are fundamental to the work of supervising the management of electricity distribution in Brazil. The continuous monitoring is performed by ANEEL in accordance with Technical Note 111/2016 (June 29, 2016) in order to prevent the degradation of the regulated service and identify any issues in energy distribution administration. In this sense, a quarterly report namely Relatório de Indicadores de Sustentabilidade Econômico-Financeira das Distribuidoras comprises 11 indicators divided into 6 subareas: debt, efficiency, investments, profitability, shareholder return and operating from 2011 to 2017. The objective of this paper is to quantify the indebtedness level of companies in the tensor structured data (company x indicator x year) provided by those reports. For this, tree-based, linear and polynomial regression models were fitted in which feature variables originated from dimensionality reduction methods, such as Principal Component Analysis and Autoencoder. Also, a Bayesian Structural Equation Model simultaneously promoting Confirmatory Factor Analysis (outer model) and incorporating linear relationships (inner model) between latent variables was fitted. The performance comparison of these methods was made from the predictive power in the validation set.

CP16: Multiple imputation procedures using the GabrielEigen algorithm

Sergio Arciniegas Alarcón[1], Marisol García Peña[2], Wojtek Krzanowski[3] and Decio Barbin[4]

[1] Universidad de La Sabana, Colombia
[2] Pontificia Universidad Javeriana, Colombia
[3] University of Exeter, United Kingdom
[4] University of São Paulo, SP, Brazil
Email: sergio.arciniegas@gmail.com

Abstract

GabrielEigen is a simple deterministic imputation system without structural or distributional assumptions, which uses a mixture of regression and lower-rank approximation of a matrix based on its singular value decomposition. We provide multiple imputation alternatives (MI) based on this system, by adding random quantities and generating approximate confidence intervals with different widths to the imputations using cross-validation (CV). These methods are assessed by a simulation study using real data matrices in which values are deleted randomly at different rates, and also in a case where the missing observations have a systematic pattern. The quality of the imputations is evaluated by combining the variance between imputations (Vb) and their mean squared deviations from the deleted values (B) into an overall measure (Tacc). It is shown that the best performance occurs when the interval width matches the imputation error associated with GabrielEigen.

CP17: On similarity and edge detection in images processing

Silvia Ojeda[1] and Grisel Maribel Britos[2]

[1]Facultad de Matemática Astronomía Física y Computación - National University of Córdoba, Argentina
Email: ojeda@famaf.unc.edu.ar

Abstract

One of the most interesting topics in the field of the study of problems involving digital images, is without a doubt the comparison and evaluation of the similarity between two images. Among the index that have been developed so far for this purpose, excels a recent proposal called index CQ, which is gaining notoriety against others of its type. This index is based on the coefficient of codispersion and it has established itself as an attractive development, based on its ability to capture the hidden similarity in a preset direction, between two images. CQ also has interesting mathematical properties that makes it an useful tool to solve optimization problems in image restoration. In this paper we address the study of the conditions under which two images are considered equal by the index CQ. From a standard image X we define the class $F(X,h)$ constituted by all the images Y that are equal to X according to the evaluation of the index. We prove that the change of any element of the class $F(X,h)$ in the direction of the definition of the index, is a multiple of the change of X in that same direction. From this result, we relate the index CQ with the process of detection of edges in digital images. Specifically, we show that the images of the class $F(X,h)$, under certain conditions, preserve the edges of the pattern image.

CP18: Trade-off between bias and variance: a simulation study with some machine learning models

Tatiana Felix da Matta[1], Anderson Ara[1], Lucas Eber Floriano de Oliveira[1] and Laion Lima Boaventura[1]

[1]Federal University of Bahia, BA, Brazil
Email: lucaseber@hotmail.com

Abstract

In statistical modeling, there is a direct relationship among the proposed model, data quality and the response obtained. An efficient way to evaluate the quality of the selected model is by checking the trade-off between bias and variance over the observed test error in the light over number of parameters or model complexity. However, there are several ways to measure the test error thought different data splitting procedures. This work considers a simulation study over data splitting procedures as hold out, repeated hold out and k-fold cross validation in order to verify the trade-off between bias and variance in common machine learning models.

CP19: Tuning of convolutional neural networks parameters applied to traffic sign recognition

Yan Andrade Neves[1] and André Luiz Carvalho Ottoni[1]

[1]Federal University of Bahia, BA, Brazil
Email: yan.andrade@hotmail.com.br

Abstract

Artificial Neural Networks (ANN) have been increasingly present in people's lives. Applications range from personal assistants that uses voice recognition to autonomous vehicles. A special type of ANN is Convolutional Neural Networks (CNN), mainly used for image recognition and video processing. In this sense, the present study aimed to analyze the influence of parameters (learning rate and number of filters in the convolutional layers) in the CNN perfomance applied to the traffic sign recognition. For

this, 25 combinations of the analyzed parameters were simulated. Experiments were performed with the python language and the keras library. The database comprises 51839 images divided into 43 traffic signs classes. Learning evaluation criteria were the loss and the accuracy of the network. The results indicated that learning rates of greater than or equal to 0.1 presented low accuracy and high loss. It is also possible to notice worse performance when using a very small rate (0.0001). As for the increasing number of filters, from 30 filters the improvement of the indicators becomes small compared to the increase in computational cost. Thus, between simulated values, learning rate of 0.001 and 30 filters in the convolutional layers proved to be the most optimized parameters for the problem.

Part X

TCC - Specialization in Data Science and Big Data

TCC - Specialization in Data Science and Big Data

ECD.1: Técnica de visualização de dados aplicada em planilhas em uma empresa de automação bancaria

Adalberto Nuno Souza da Conceição[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: adalbertonuno@hotmail.com
Orientador: Anderson Ara

Abstract

Hoje na sociedade existe o crescimento do volume e dos tipos de dados a necessidade de analisar, entender e representa esses relacionamentos. Portanto, técnicas de visualizações baseadas nos mais diversos tipos de variações ganha espaço, interesse e importância como uma possível ferramenta para esse problema, proporcionando uma forma simples e rápida em identificar padrão, tendências e extrair novos conhecimentos. Card e outros (CARD et al., 1999) definem Visualização de Informações como sendo "o uso de representações visuais de dados abstratos suportados por computador e interativas para ampliar a cognição".

Para o rápido e fácil conhecimento através da analise, interpretação e visualização da planilha PEÇA EM TESTE foram realizadas o desenvolvimento de várias visualizações amigáveis, simples e didáticas utilizando Shiny e Linguagem R.

Neste contexto, este projeto tem como objetivo principal criar metáforas visuais denotativas baseadas em atributos quantitativos e qualitativos, através das medidas estatísticas e artefatos para a visualização dos dados, para auxiliar na exploração e análise.

Outrossim, propõe-se disponibilizar em ambiente dinâmico, as técnicas de visualização de dados desenvolvidas na conclusão do curso. O desenvolvimento do projeto com visualização dinâmica inspira-se na dificuldade de criar visualização limpa, simples e de fácil entendimento. Além disso, o fato do projeto apresentar visualizações especifica sobre as operações em regiões, usuário e registros tem o propósito de facilitar o acesso e a divulgação das informações para o publico interno.

ECD.2: Data warehouse e data mining, extraindo conhecimento de dados públicos federais

Adriana Ferreira Lacerda[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: flacerda.adriana@gmail.com
Orientador: Juracy Araujo de Almeida Junior

Abstract

No cenário atual, no qual a produção de dados está cada vez maior, nos deparamos com a crescente necessidade da utilização de soluções tecnológicas, que atuam como agentes facilitadores para a extração de conhecimento através da exploração de dados. Diante desta perspectiva, a proposta deste artigo é constatar que através da utilização de Data Warehouse e Data Mining é possível extrair informações, que por sua vez tornam-se insumos para a obtenção do conhecimento e consequentemente constituir uma base para corroborar com a tomada de decisões. Logo, para isto, aplicaremos tais soluções na base de dados de Despesas Executadas, disponibilizada no Portal Transparência do Governo Federal, para o período de janeiro de 2014 até outubro de 2019.

ECD.3: Object tracking aplicado em contagem de pessoas

Alã de Cerqueira Damasceno[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: alacerdan@hotmail.com
Orientador: Luciano Oliveira Rebouças

Abstract

É uma tarefa árdua, se não impossível, considerando as limitações humanas, contar, qualquer que seja, o objeto de interesse em casos de aglomeração, onde a sobreposição de alvos e o tempo para a execução dessa tarefa são fatores impeditivos para a formação de decisão, com dados provenientes do videomonitoramento. O presente trabalho destina-se a utilizar técnicas de detecção e rastreamento de objetos (object detection and traking) com redes neurais convolucionais (CNN) para estimar quantidade de pessoas em locais de formação de multidões, mas precisamente na área do metrô da Estação da Lapa, na cidade de Salvador, estado da Bahia, onde se pode estimar multidões afim de subsidiar o planejamento na segurança pública, mobilizando agentes conforme a demanda de público. Assim, monitoramento através de vídeo de vigilância, posicionados estrategicamente, permite a mobilização de recursos públicos para melhor atendimento da sociedade, baseando-se na ideia de concentração de pessoas através do processo de contagem, direcionando policiamento em eventos festivos, a instalação de prepostos policiais, dentre outros.

ECD4: DAT - Un dashboard de ativos de TI

Alex Sandro Brandão de Almeida[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: rdalias189@gmail.com
Orientador: Wecsley Otero Prates

Abstract

O trabalho tem por objetivo apresentar um modelo de dashboard – painel interativo de visualização de dados – construído a partir de uma base de dados de ativos tecnológicos onde se prioriza visões das relações entre diversas variáveis que se associam. Unimos à esta proposta um painel gerencial de ativos cujo propósito é permitir que o gestor dos dados possa realizar diversas análises a respeito do comportamento dos ativos de tecnologia ao longo do ciclo de vida do ativo.

ECD.5: Aplicação de modelos de redes neurais artificiais com retroalimentação para previsão do índice de produção industrial do Brasil

André Gama Rebouças[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: andre.reboucas@bcb.gov.br
Orientador: Eduardo Furtado de Simas Filho

Abstract

Este trabalho estuda uma estratégia de aplicação de redes neurais artificiais com retroalimentação dos tipos Echo State Networks (ESN) e Long Short-Term Memory Networks (LSTM) na previsão de valores futuros das séries de índices de produção da indústria extrativa (PIM IE) e de transformação (PIM IT) brasileiras, divulgadas pelo Instituto Brasileiro de Geografia e Estatística. Para previsão desta última série, é necessário o auxílio de séries complementares correlacionadas. Foi implementada uma estratégia de variação de parâmetros para os quais foram executadas trinta diferentes inicializações de modelos realizando previsões em um conjunto de validação. Para avaliação da melhor arquitetura para as redes neurais, foi realizado o cálculo do erro médio e do desvio padrão dos resultados. A melhor configuração para cada série e tipo de rede neural foi então executada em trinta diferentes inicializações para prever no conjunto de teste formado pelos últimos doze pontos da série. A média das trinta predições de cada ponto foi considerada como a predição final da estratégia proposta. Foi calculado também um intervalo de confiança robusto. Os resultados obtidos indicam que os modelos propostos são adequados para a previsão das séries PIM IE e PIM IT.

ECD.6: Técnicas de machine learning aplicadas a predição da quantidade de leitos hospitalares nos municípios do Brasil

Bruno Almeida de Carvalho[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: bruno.carvalho@live.com
Orientador: Ricardo Rocha
Co-Orientador: Anderson Ara

Abstract

Utilizando diversas bases públicas foram levantadas importantes variáveis que podem contribuir no modelo, como: tipos de unidades básicas de saúde, projeção populacional do município, área do município e sua repetitiva densidade populacional. Uma vez trabalhados, esses dados servirão pra que seja possível prever a quantidade de leitos em municípios que não os possuem, mas que deveriam estar à disposição da população. Para tanto, serão empregadas técnicas de aprendizado de máquina pertinentes ao contexto do problema.

ECD.7: Construção de um data warehouse sob a luz da Lei Geral de Proteção de Dados (LGPD)

Carla Nascimento Caldeira da Costa[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: carlacaldeira@hotmail.com.br
Orientador: Juracy Almeida

Abstract

O estudo refere-se a aplicação da Lei Geral de Proteção de dados - LGPD (Lei n. 13.7092018) sancionada no Brasil em agosto de 2018 e entrará em vigor em agosto de 2020, aborda a problemática sob a ótica dos novos desafios a serem enfrentados para adequação a nova legislação e os impactos que sua implementação trará. A LGPD teve uma forte inspiração no Regulamento Geral de Proteção aos dados (GDPR - General Data Protection Regulation) pois o regimento Europeu obriga que só pode haver fluxo internacional de dados com países que obtenham leis semelhantes a GDPR. Assim, o presente trabalho acadêmico tem como escopo promover análise dos efeitos dessa mudança em um Data Warehouse - banco de dados organizado para dar suporte à tomada de decisões estratégicas da organização, que frequentemente encontram-se espalhados em diversos sistemas - e a nova sistemática de proteção de dados no Brasil.

ECD.8: Aplicações dos métodos de classificação da causa da mortalidade de empregados de uma companhia hidrelétrica brasileira no período de 1985 a 2011

Ednai Batista Alves[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: ednaialves@hotmail.com
Orientador: Gecynalda Soares da Silva Gomes

Abstract

No mundo contemporâneo seria praticamente impossível viver sem a energia elétrica devido a ampla funcionalidade dela para a vida humana. Uma das formas de obter esta energia é por meio das Usinas Hidrelétricas, que funcionam convertendo energia potencial hidráulica em elétrica. No Brasil, essa matriz, corresponde a 66% da energia produzida, segundo a ANEEL, 2016. Devido às usinas hidrelétricas serem a maior fonte de energia no país, para o pleno funcionamento das mesmas é necessário um grande número de trabalhadores, os quais, muitas vezes, estão expostos a situações de periculosidade. Para este estudo, utilizou-se uma base de dados coletada nas declarações de óbito e nas fichas cadastrais da companhia em estudo que corresponde à população de trabalhadores de uma companhia hidrelétrica brasileira, no período de 1985 a 2011. O objetivo é aplicar técnicas de Aprendizagem de Máquina (ou Machine Learning), para identificar o melhor classificador binário dentre os estudados para este conjunto de dados. De modo a obter o modelo mais adequado para classificar a causa do óbito do indivíduo (empregados da hidrelétrica), em causa devido à periculosidade (classe 1) e causa devido a outros fatores (classe 2). Foram utilizados quatro métodos de classificação e os modelos ajustados foram avaliados por meio da validação holdout repetida, adotando-se como medida de desempenho a acurácia.

ECD.9: Engenharia de features linguísticas para classificação de triplas relacionais em Galego, Português e Espanhol

Elian Conceição Luz[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: elianconceicaoluz@gmail.com
Orientador: Daniela Barreiro Claro

Abstract

O exponencial crescimento do volume de documentos digitais escritos em linguagem natural impulsiona a demanda por modelos automáticos capazes de extrair informação de dados não-estruturados. Nessa perspectiva, a Extração de Informação Aberta (EIA) possibilita a estruturação da linguagem natural em triplas relacionais ao processar as sentenças de um texto, obtendo uma estrutura composta por três elementos (arg1, relação, arg2). (CLARO et al, 2019) Tradicionalmente, a EIA inclui métodos de extração e validação de triplas relacionais que, em sua maioria, foram desenvolvidos com base em corpus de língua inglesa, considerando suas especificidades linguísticas, como o preenchimento do sujeito e a ordem sujeito-verbo, características pouco expressivas nas línguas românicas. (BARBOSA, 2018) Destarte, neste estudo, objetivou-se elencar features linguísticas com base em triplas relacionais extraídas de corpora em quatro línguas românicas (galego, português europeu, português brasileiro e espanhol europeu), explorando as características linguísticas genéricas por meio de uma revisão de estudos da Linguística Formal, (OLIVEIRA, 2004) em destaque, para os de base gerativista. (KATO, 2002) Assim, as features elencadas foram utilizadas para a classificação das triplas em validas (1) e inválidas (0) por meio de métodos de aprendizagem de máquina, identificando sentenças que oferecem maior dificuldade para extrair triplas válidas, como as que apresentam objeto/sujeito nulo e inversão verbo-sujeito; bem como triplas relacionais que formam subsentenças agramaticais. Dessa forma, os experimentos realizados relacionam estudos da Linguística Formal à engenharia de features, colaborando para a Extração de Informação Aberta em perspectiva multilingual.

ECD.10: Governança de dados em instituições de ensino superior apoiado por um data warehouse para estudo do planejamento de turmas

Fábio Roberto dos Anjos Santos[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: bocamaster@hotmail.com
Orientador: Juracy Almeida

Abstract

O presente trabalho visa apresentar uma arquitetura genérica de um Data Warehouse(DW) que possa auxiliar uma Instituição de Ensino Superior a analisar como foi a ocupação de turmas de disciplinas que aconteceram em períodos passados, para que a partir desta informação seja possível realizar inferências para a tomada de decisão sobre planejamentos de disciplinas em períodos futuros. Serão vistos como os conceitos de gestão e governança de dados podem auxiliar uma instituição para a definição de melhores formas de coletar e tratar dados a fim de auxiliar na construção do Data Warehouse. A finalidade da execução dessas ações será permitir que os gestores tenham acesso a informação já tratada e de acordo com as necessidades de análise que o negócio carece.

ECD.11: Minerando as interações de características em smart-homes

George Dantas Cardozo[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: george.dantas.c@gmail.com
Orientador: Daniela Claro

Abstract

Os sistemas de casas inteligentes (Smart-Homes) têm se tornado uma tecnologia cada vez mais importante na vida moderna, através da Internet das Coisas, sensores, redes de comunicação e atuadores uma ampla variedade de serviços domésticos inteligentes são fornecidos, facilitando os cuidados domésticos e melhorando o estilo de vida das pessoas. No entanto, a introdução de tantos dispositivos que estão constantemente interagindo, pode resultar em comportamentos indesejados, esse efeito é conhecido como interações de características. Estes dispositivos conectados dentro da casa inteligente produzem uma quantidade significativa de dados, o que revela comportamentos e padrões de suas interações. Este trabalho propõe a detecção dessas interações de características, através de técnicas de classificação em aprendizado de máquina por meio dos dados fornecidos pelos aparelhos e dispositivos conectados. É proposta uma abordagem que consiste em três etapas. Na primeira etapa, um modelo é desenvolvido para capturar a ocorrência das interações de características em uma base de dados de casa inteligentes. Na segunda etapa, o modelo é analisado através de técnicas de validação de modelo. Na terceira etapa são apresentados os resultados obtidos da utilidade da abordagem proposta na detecção das interações de características.

ECD.12: Previsão do índice da indústria de transformação Brasileiro com o uso de redes neurais artificiais

Gustavo Loula Castro Nunes[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: gustavo.nunes@bcb.gov.br
Orientador: Eduardo Furtado de Simas Filho
Co-Orientador: Gecynalda Soares da Silva Gomes

Abstract

O objetivo deste trabalho é avaliar o uso de métodos de previsão de valores de séries temporais, usando redes neurais artificiais (RNA) dos tipos Multilayer Perceptron (MLP) e Extreme Learning Machine (ELM) aplicados à previsão de um valor futuro da série do índice de produção da indústria de transformação brasileira (PIM IT), divulgada pelo Instituto Brasileiro de Geografia e Estatística. As RNA têm sido usadas em diferentes áreas do conhecimento, considerando sua capacidade de, como o cérebro humano, reconhecer padrões e regularidades, e generalizar com base no conhecimento acumulado. O método utilizado envolve treinar as redes com um grupo de valores anteriores da série, além de variáveis exógenas correlacionadas com capacidade de explicar o comportamento da série alvo, adaptando o conjunto de aprendizado para um modelo autorregressivo, em que valores atrasados da própria série alvo, além das variáveis exógenas, são usados para previsão de um único valor à frente. As redes treinadas são posteriormente avaliadas num outro conjunto de dados denominado conjunto de validação e, por fim, é selecionada a de menor erro médio medido por medida estatística apropriada para a previsão de um ponto à frente num conjunto separado previamente para o teste final. Os resultados obtidos revelam o bom desempenho dos métodos escolhidos e a viabilidade do uso de RNA para previsão do índice.

ECD.13: Técnicas de machine learning aplicadas a classificação de clientes de uma distribuidora de derivados de petróleo do Brasil

Hamilton Batista Lima Sobrinho[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: hamilton21@gmail.com
Orientador: Ricardo Rocha

Abstract

A partir da base de dados de compras de clientes, levantou-se importantes variáveis que através da aplicação de algoritmos de machine learning podem classificar os clientes existentes em categorias de modo a possibilitar diferentes ações comerciais, se acordo com a estratégia da organização. Uma vez processados, os clientes podem ser categorizados de acordo com a probabilidade de recompra, faixa de margem de lucro, e faixa de volume de vendas.

ECD.14: Aplicação de CB-SVM em bases de dados grandes

Ivalbert dos Santos Pereira[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: ivalbert.pereira@gmail.com
Orientador: Anderson Ara
Co-Orientador: Ricardo Rocha

Abstract

O objetivo desse trabalho é aplicar uma variação do método de classificação SVM chamado Clustering-Based SVM (YU, 2005). Essa variante propõe-se a tornar possível a utilização de máquinas de vetor suporte em grandes bases de dados. Assim, são aplicados o CB-SVM e SVM, de duas diferentes implementações de pacotes do R-Studio, em uma base de dados grande e comparados os resultados em termos de tempo de processamento e poder de predição (acurácia).

ECD.15: Churn modelling: aplicando machine learning na predição das saídas de clientes da base de dados de instituições financeiras

Ive de Oliveira Gavazza[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: gavazza.ive@gmail.com
Orientador: Ricardo Rocha
Co-Orientador: Anderson Ara

Abstract

O churn pode ser definido como a saída de um usuário da base de clientes de uma instituição financeira, por exemplo. O entendimento do comportamento do cliente que deixará a base de usuários é importante para as organizações, pois ajudam a tomada de decisão a fim de minimizá-lo, reduzindo custos de angariar novos clientes ou de manter antigos. Com isso, este trabalho propõe a utilização de aprendizado de máquina na análise do comportamento do usuário que deixará a base de clientes de uma instituição financeira, utilizando uma base de dados contendo informações sobre faixa etária, faixa salarial, credit score, localidade, entre outras. A análise destas variáveis serve de embasamento para a identificação de correlações entre variáveis, determinação de perfis de clientes e aplicação de estratégias de retenção de clientes.

ECD.16: Youtube, redes de influência e polarização: avaliando o posicionamento político de influenciadores digitais

Júnia Ortiz[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: junia.ortiz@gmail.com
Orientador: Wecsley Prates
Co-Orientador: Crysttian Arantes Paixão

Abstract

O trabalho apresenta uma análise dos principais canais de influenciadores digitais no Youtube com o objetivo de identificar o perfil político e as redes de influência destes atores. Para tanto, foram analisados 2513 canais. A composição da amostra foi realizada a partir de uma base de 58 canais de influenciadores brasileiros, identificados por matriz política (esquerda, centro e direita), a fim de garantir diversidade na amostra. A partir dessa base, foram coletadas as redes conectadas a estes canais com a utilização da ferramenta Channel Network Module disponibilizada pela Digital Methods Initiative. Foram coletados, então, os dados referentes aos canais (vídeos, comentários e metadados). Para a análise do perfil político dos influenciadores foram empregadas técnicas de Mineração de Texto e Machine Learning. O estudo é

uma importante fonte para a compreensão das dinâmicas que envolvem a circulação de conteúdo político em ambientes online no contexto brasileiro.

ECD.17: Visualização dinâmica sobre as empresas Brasileiras

Laion Boaventura[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: laion@limaboaventura.com.br
Orientador: Anderson Ara

Abstract

Neste trabalho, utilizamos técnicas de análise e tratamento em dados massivos para extrair informações a respeito da base de dados CNPJ, disponível, conforme determina a Constituição Federal de 1988 cc a Lei de Acesso à Informação e cc o Decreto Federal n. 8.7772016, no site da Receita Federal. Os dados representam informações a cerca de todas as empresas cadastradas no Brasil, sendo portanto, uma das fontes governamentais de informação pública mais relevantes do país. Contudo, apesar do banco de dados CNPJ estar disponível para download, a Receita Federal publicou-o em um formato do tipo fixed width - arquivo de largura fixa, o que impossibilita o uso direto de sofwares como R e Python para análise de dados. Além disso, os arquivos descompactados tem mais de 85Gb, o que limita mais ainda o amplo acesso ao público, em geral. Posto isso, neste trabalho, utilizamos o Sotware R Core Team 2019, através da biblioteca qsacnpj, desenvolvido pelo Observatório Social do Brasil, localizado no Município de Santo Antônio de Jesus, no Estado da Bahia, para transformar os dados CNPJ do formato fixed width - arquivo de largura fixa em csv. Após isso, foi possível, por meio das bibliotecas sparklyr e ggplot2, também do Sotware R, analisar os dados, e extrair de forma visual, informações úteis para sociedade. Por fim, através do Software Tableau, montamos um painel visual para ilustrar, de maneira dinâmica, as informações extraídas no tratamento dos dados CNPJ.

ECD.18: Desenvolvimento de um classificador usando relatos dos assistidos da defensoria pública da Bahia

Lucas Pereira da Silva Souza[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: lucas.siva@hotmail.com
Orientador: Paulo Canas Rodriges
Co-Orientador: Crysttian Arantes Paixão

Abstract

A defensoria pública da Bahia possui um sistema de chamado agendamento online para realizar a solicitação de serviços. Por meio desse sistema, os cidadãos solicitam serviços da defensoria. Esse agendamento direciona o atendimento para defensores específicos. Hoje esse direcionamento é realizado por um servidor, que analisa o pedido e toma a decisão de qual defensoria deve avaliá-lo. O objetivo desse trabalho é desenvolver um classificador, baseado em aprendizado de máquina, que realize o direcionamento dos pedidos, considerando o texto da descrição do pedido. Com isso, espera-se gerar uma otimização nos agendamentos, aumento na eficiência nos serviços prestados na defensoria pública do Estado da Bahia.

ECD.19: Métodos quantitativos de previsão de vendas: aplicado a uma companhia de insumos agrícolas

Murilo Martins de Souza[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: murilomartinz@gmail.com
Orientador: Gecynalda Soares da Silva Gomes

Abstract

O Brasil é um grande produtor agrícola e consumidor de insumos que possibilitam sustentar suas altas produtividades. As vendas de produtos para a agricultura no país sofrem forte influência de fatores externos e internos, que podem causar variações significativas no consumo e no processo produtivo desses insumos. Garantir previsibilidade nas vendas, traz para a companhia uma melhor capacidade de planejamento e otimização dos recursos necessários para produção. Partindo dessa necessidade, aplicaremos métodos quantitativos de séries temporais, com o intuito de realizar previsões de vendas através de modelos que produzam maior precisão possível.

ECD.20: Uma arquitetura big data analytics para integração de sistemas com fontes heterogêneos

Nelci Gomes Lima[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: nelcisgomes@gmail.com
Orientador: Luciano Rebouças de Oliveira

Abstract

A integração e uniformidade de dados com fontes heterogêneas são tarefas complexas que, devido ao grande volume de dados produzidos diariamente por empresas e organizações, requerem muito estudo. (RAJABIFARD, ano) Assim, neste trabalho, apresentaram-se a concepção, o processo de desenvolvimento e os resultados operacionais iniciais de uma solução para integrar dados oriundos de fontes distintas. Essa pesquisa teve como objetivo geral validar uma aplicação onde os dados extraídos de diversas tecnologias são direcionados a um ponto central de armazenamento e processamento de informações de forma homogênea, o que permite a apresentação dos dados de maneira assertiva e padronizada. Nessa perspectiva, a ferramenta propósta realizou em tempo real as principais etapas no processo de normalização de dados, contribuindo para o avanço de estudos dessa área ao disponibilizar uma solução que promove integração de dados com intuito de obter uma gestão operacional de ativos e pessoas focalizadas em georreferenciamento com eficiência e rapidez, possibilitando o surgimento de novas variáveis que ganham maior potencial de aplicação quando atendem à analise de tendências futuras em uma arquitetura big data analytcs.

ECD.21: Aperfeiçoando a coleta de dados disponibilizados por um site do governo dos Estados Unidos através de web scraping: uma aplicação com as taxas de incidência de lesões ou doenças relacionadas ao trabalho

Thaline Ferreira Silva[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: thalinefs@gmail.com
Orientador: Gecynalda Soares da Silva Gomes

Abstract

O web scraping é uma técnica que envolve a coleta automatizada para extração de informação de páginas web. Suas ferramentas possibilitam poupar tempo na etapa de coleta dos dados. É capaz de transformar informação não estruturada em informação estruturada que pode depois ser armazenada e analisada.

O objetivo do artigo foi utilizar a ferramenta de extração de dados em ambiente web para auxiliar no estudo espacial da taxa de incidência de lesões ou doenças ocupacionais. Inicialmente, realizou-se uma fundamentação teórica sobre as ferramentas de extração de informações via web. Assim, foi utilizada uma biblioteca do software R para extração de dados do site da secretária de estatísticas trabalhistas dos Estados Unidos, buscando a taxa de incidência de lesões ou doenças relacionadas ao trabalho por estado, transformando-as em um banco de dados estruturado. Após essa etapa, foi possível realizar uma análise espacial a fim de verificar se há relação de dependência entre essas taxas de lesões ou doenças ocupacionais. A metodologia apresentada poderá auxiliar as esferas públicas em extrair informações estratégicas que estão disponibilizadas na web, com baixo custo, otimizando ações e garantindo uma melhoria no uso de recursos.

ECD.22: Modelo para previsão de chance de um paciente retornar para uma Unidade de Terapia Intensiva (UTI) durante a mesma hospitalização após receber alta de uma UTI

Yandreson Carvalho Cavalcante[1]

[1]Universidade Federal da Bahia, BA, Brasil
Email: yan.ecomp@gmail.com
Orientador: Lizandra Castilho Fabio
Co-Orientador: Luciano Oliveira

Abstract

Os pacientes internados em Unidades de Terapia Intensiva (UTI) frequentemente são avaliados pela equipe médica e multidisciplinar a fim de verificar se eles apresentam condições para alta. Geralmente define-se o momento ideal para saída do paciente da UTI com base em evidências clínicas, características individuais e subjetivas. No entanto os critérios empregados são amplos e bastante subjetivos, contribuindo para indicações indevidas de alta, o que pode provocar o retorno dele para a UTI durante a mesma hospitalização ou expô-lo a níveis inadequados de cuidados, podendo acarretar no agravamento de seu quadro clínico de saúde.

Para que as altas realizadas nas UTI sejam realizadas de forma mais assertiva, tem se utilizado o score SWIFT (Stability and Workload Index for Transfer) para medir as condições adequadas para alta dos pacientes. A pontuação do score possui escala de 0 a 64, sendo que estudos indicam que quanto maior a pontuação, maior é o risco de reinternação na UTI.

O objetivo deste projeto foi construir um modelo baseado no score SWIFT para identificar os pacientes com risco de retornar à unidade de terapia intensiva. O resultado obtido pela aplicação do modelo será o percentual que representa a chance de o paciente ser reinternado na UTI caso ele receba alta de uma UTI em determinado instante.

Index