#meetingdata

18 a 20 de Outubro, 2018 - UFSCar, São Carlos-SP

Livro de Programa, Resumos e Anotações

Organização e Apoio:

#meetingdata

Ciência de Dados é um campo multidisciplinar que engloba métodos e processos para obtenção de conhecimento ou compreensão de premissas a partir de dados (informações).

As soluções apresentadas em Ciências de Dados empregam métodos de diferentes áreas da Matemática, Estatística, Ciência da Informação, Computação Científica e especialmente nas subáreas de aprendizado de máquina, modelagem estatística, classificação, mineração de dados, inteligência artificial, métodos preditivos e modelos probabilísticos.

Conceitos estatísticos e análise de dados são empregados para entender e analisar fenômenos atuais. Grandes quantidades de dados estão disponíveis e sua análise depende não apenas dos métodos estatísticos usuais, mas também do uso de técnicas computacionais para resolver o problema de *Big Data*. Dessa forma, entende-se que Ciências de Dados une naturalmente as ciências Estatística e Computação, desenvolvendo soluções para os desafios enfrentados nas áreas de Astronomia, Biologia, Epidemiologia, Linguística, Medicina, entre outras.

O #meetingdata é um encontro para discutir/apresentar o desenvolvimento de Ciências de Dados sob o domínio de multidisciplinaridades fundamentais para a pesquisa. Os tópicos cobrirão desde a análise estatística até métodos de aprendizado de máquina, promovendo uma busca por soluções atu-

ais para os desafios e necessidades da sociedade.

O site do evento é:
http://www.datascience.ufscar.br/meetingdata

São Carlos, 18 de Outubro de 2018

Comissão Organizadora

ORGANIZAÇÃO

Comitê de Organização

- Adriano Polpo (DEs-UFSCar)
- Agatha Sacramento Rodrigues (IME-USP/FMUSP)
- Danilo Lourenço Lopes (DEs-UFSCar)
- Diego Furtado da Silva (DC-UFSCar)
- Guilherme Barreto Fernandes (Serasa Experian)
- Heloisa de Arruda Camargo (DC-UFSCar)
- Murilo Cantoni (DEs-UFSCar)
- Rafael Bassi Stern (DEs-UFSCar)
- Ricardo Ciferri (DC-UFSCar)
- Teresa Cristina Martins Dias (DEs-UFSCar)

Comitê Científico

- Adriano Polpo (DEs-UFSCar)
- Estevam Rafael Hruschka Júnior (DC-UFSCar)
- Hermes Senger (DC-UFSCar)
- Marcio Alves Diniz (DEs-UFSCar)
- Rafael Izbicki (DEs-UFSCar)
- Ricardo Cerri (DC-UFSCar)

Contato:

- datascience@ufscar.br

	Quinta-feira (18/10)	**Sexta-feira (19/10)**	**Sábado (20/10)**
09:00-09:25		**Sessão Oral O3** João Carlos P. Ferreira	**Sessão Oral O7** Ana Carolina Simionato
09:25-10:15		**Conferência C2** Estevão Vieira	**Conferência C6** Alexandre Chiavegatto
10:15-10:35		**Coffee Break**	**Coffee Break**
10:35-11:00		**Sessão Oral O4** Marcio A. Diniz	**Sessão Oral O8** Osvaldo A. Júnior
11:00-11:50		**Conferência C3** Igor Braga	**Conferência de Encerramento** Rafael Izbicki
11:50-12:00		**Horário de Almoço**	**Encerramento**
12:00-13:50			
13:50-14:00	**Abertura**		
14:00-14:50	**Conferência de Abertura** André de Carvalho	**Conferência C4** Rafael Monteiro	
14:50-15:15	**Sessão Oral O1** Guaraci Requena	**Sessão Oral O5** Adriano B. Morales	
15:15-15:40	**Sessão Oral O2** Eduardo K. Nakao	**Sessão Oral O6** Carlos Roberto S. Júnior	
15:40-16:30	**Sessão Pôster** **com café**	**Sessão Pôster** **com café**	
16:30-17:20	**Conferência C1** Otávio Vasques	**Conferência C5** Florencia Leonardi	

Tabela 2: Contribuições

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
André de Carvalho	AutoML: Automated Machine Learning	12	18/Out 14:00-14:50	A
Otávio Vasques	Reconstruindo modelos de crédito e segmentação com dados de celulares	13	18/Out 16:30-17:20	C1
Estevão Vieira	Como nosso cérebro representa o tempo	14	19/Out 9:25-10:15	C2
Igor Braga	Aprendizado de máquina sob a influência de covariate-shift	15	19/Out 11:00-11:50	C3
Rafael Monteiro	Data Science & Data Engineering aplicados ao mercado de capitais Brasileiro	16	19/Out 14:00-14:50	C4

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Florência Leonardi	Estimadores regularizados para problemas de alta dimensão	17	19/Out 16:30-17:20	C5
Alexandre Chiavegatto	Aplicações de machine learning em saúde	18	20/Out 9:25-10:15	C6
Rafael Izbicki	FlexCode: modelando incertezas em problemas de predição	19	20/Out 11:00-11:50	E
Guaraci Requena	Multinomial regression via binomial regressions	20	18/Out 14:50-15:15	O1
Eduardo Kazuo Nakao	Manifold learning for non-linear dimensionality reduction in hyperspectral image unsupervised classification	21	18/Out 15:15-15:40	O2
João Carlos Poloniato Ferreira	Aplicação do FBST em modelos bayesianos de alta dimensão	23	19/Out 09:00-09:25	O3

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Marcio Alves Diniz	Um experimento sobre previsões probabilísticas	24	19/Out 10:35-11:00	O4
Adriano Barasal Morales	Firm location: an approach using spatial point process	25	19/Out 14:50-15:15	O5
Carlos Roberto Silveira Junior	Mineração de regras de associação espaço-temporais temáticas aplicada a imagens de explosões solares	26	19/Out 15:15-15:40	O6
Ana Carolina Simionato	Sustentabilidade e curadoria digital para coleções de patrimônio cultural	28	20/Out 09:00-09:25	O7
Osvaldo Anacleto Júnior	Generalising dynamic Bayesian networks to accommodate causal and symmetric signals in high-dimensional time series	30	20/Out 10:35-11:00	O8

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Agatha Sacramento Rodrigues	Curva de crescimento fetal personalizada	31	18-19/Out 15:40-16:30	P1
Alejandra Estefanía Patiño Hoyos	Adaptative significance levels in normal mean hypothesis testing	33	18-19/Out 15:40-16:30	P2
Alisson Hayasi da Costa	Performance analysis of deep neural networks in piRNAs classification	34	18-19/Out 15:40-16:30	P3
Amanda Azevedo dos Santos	Recursos sonoros e linked open data: MusicBrainz	35	18-18/Out 15:40-16:30	P4
Aulida Berenice Moretti dos Santos	Dados abertos: criação de novos negócios	37	18-19/Out 15:40-16:30	P5
Brendon Gouveia Cambuí	Feature extraction for multi-target learning	39	18-19/Out 15:40-16:30	P6
Bruna Zamith Santos	Predicting protein functions via interaction prediction	40	18-19/Out 15:40-16:30	P7

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Camila Lorencetti Brolo	Comparação de testes de hipóteses para duas médias em variáveis do tipo proporção	42	18-19/Out 15:40-16:30	P8
Camila Sgarioni Ozelame	Redes Bayesianas: uma comparação entre métodos de estimação de estrutura	43	18-19/Out 15:40-16:30	P9
Dhiogo José Corrêa de Sá	Momentos de vida: mineração de textos e Big Data para oferecimento de serviços personalizados	45	18-19/Out 15:40-16:30	P10
Diogo Barboza Moreira	Obtenção de curvas de confiabilidade em testes de vida acelerados	47	18-19/Out 15:40-16:30	P11
Elaine Cecília Gatto	Proposta de um modelo de aprendizado competitivo para classificação hierárquica multirrótulo	48	18-19/Out 15:40-16:30	P12

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Estela Maris Pereira Bereta	Estimadores, pontual e intervalar, para dados com censuras intervalar	50	18-19/Out 15:40-16:30	P13
Julio M. Stern	Haphazard intentional sampling techniques in network design of monitoring stations	51	18-19/Out 15:40-16:30	P14
Leandro Augusto Ferreira	Random Bernstein polynomials – a nonparametric bayesian estimation of densities via ABC	53	18-19/Out 15:40-16:30	P15
Leonardo Utida Alcântara	Semi-supervised predictive clustering tree for protein subcellular localization	54	18-19/Out 15:40-16:30	P16
Lucas Eduardo de Moraes	Implementação de inferência Bayesiana não paramétrica para processos pontuais espaciais	56	18-19/Out 15:40-16:30	P17

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Marcelo de Souza Lauretto	Alocação intencional fortuita: um estudo de caso em avaliação de software	57	18-19/Out 15:40-16:30	P18
Márcio Luis Lanfredi Viola	Using Markov chain in data science	58	18-19/Out 15:40-16:30	P19
Maykon Rocha Santana	Uma abordagem distribuida para sistemas fuzzy evolutivos multiobjetivos em problemas de Big Data	59	18-19/Out 15:40-16:30	P20
Pedro Luiz Paolino Chaim	Análise estatística do desempenho do PT nas eleições proporcionais entre 2000 e 2016, e previsões para 2018	61	18-19/Out 15:40-16:30	P21
Rafael Stoffalette João	Mineração de regras de associação temporais envolvendo dados quantitativos contínuos	62	18-19/Out 15:40-16:30	P22

Continua na próxina página

Tabela 2 – *Contribuições*

Apresentador(a)	**Título**	**Pág.**	**Data**	**Sessão**
Suzane Carol de Lima	Aplicação de aprendizado ativo na tarefa de classificação de textos em fluxo de dados	64	18-19/Out 15:40-16:30	P23
Taís Roberta Ribeiro	O modelo de cópula de Frank para dados de sobrevivência bivariados: modelagem, estimação Bayesiana e pontos influentes	65	18-19/Out 15:40-16:30	P24
Thiago Zafalon Miranda	Proposta de geração de regras de classificação multirrótulo simultaneamente eficazes e interpretáveis via otimização multiobjetivo com algoritmos genéticos	66	18-19/Out 15:40-16:30	P25
Victor Azevedo Coscrato	Agnostic tests can control the type I and type II errors simultaneously	67	18-19/Out 15:40-16:30	P26

AutoML: Automated Machine Learning

André Carlos Ponce de Leon Ferreira de Carvalho
ICMC–USP

Resumo

As the number of successful applications of Machine Learning algorithms grows, there is also an increase in the need to make these algorithms easily accessible by users without Machine Learning expertise. There have been several efforts in this direction, involving not only the recommendation of the most suitable algorithm, but also their most appropriate hyper-parameter values. These several efforts started a new research area, named Automated Machine Learning, AutoML, which has attracted the attention of researchers and practitioners not only from the academia, but also from several companies working with data science. This talk will present the main approaches and recent advances in this area, covering also works carried out in the Analytics Laboratory, at USP São Carlos.

Reconstruindo modelos de crédito e segmentação com dados de celulares

Otávio Vasques
Serasa Experian - IF-USP

Resumo

Como dados provenientes de aparelhos celulares auxiliam a desenvolver modelos de crédito e segmentação de marketing? Discutimos os desafios e soluções para coletar e produzir modelos a partir de variáveis provenientes de aparelhos de celular, aplicativos, localização, etc.

Como nosso cérebro representa o tempo

Estevão Vieira
Serasa Experian – UFABC

Resumo

Um problema discutido em neurociência é como os nossos cérebros representam o mundo externo (ou conceitos internos). Apresentamos algumas técnicas de aprendizado de máquina para medir a representação de tempo no cérebro, a partir da atividade de neurônios durante a realização de uma tarefa dependente do tempo, encontrando que regiões do cérebro mudam sua representação através de repetições da tarefa.

Aprendizado de máquina sob a influência de covariate-shift

Igor Braga
Big Data

Resumo

Covariate-shift acontece quando as bases de treinamento e de teste não compartilham a mesma distribuição das variáveis de entrada. Esse fenômeno está presente em diversos problemas de mundo real e, potencialmente, quando a distribuição de teste está sob o controle de terceiros. Nesta palestra, mostramos como covariate-shift prejudica o desempenho dos modelos aprendidos, e abordamos técnicas para detectar e tratar esse problema.

Data Science & Data Engineering aplicados ao mercado de capitais Brasileiro

Rafael Monteiro
Serasa Experian – Mackenzie

Resumo

Apresentamos como tratar as dezenas de gigabytes de dados gerados por dia, provenientes da bolsa de valores (em tempo real), utilizando o algoritmo de Long Term Short Memory Networks (LSTM) para identificação de padrões e análises.

Estimadores regularizados para problemas de alta dimensão

Florencia Leonardi
IME–USP

Resumo

Nesta palestra serão apresentados alguns problemas típicos de estimação em altas dimensões, como o caso de modelos de regressão linear com mais variáveis que observações ou de modelos gráficos onde o número de vértices é maior que o tamanho da amostra. Introduziremos estimadores do tipo LASSO com regularização baseada na norma l1, e explicaremos quais são as vantagens deste tipo de proposta. Além da exposição teórica, apresentaremos alguns exemplos de aplicação.

Aplicações de machine learning em saúde

Alexandre Dias Porto Chiavegatto Filho
FSP–USP

Resumo

O rápido aumento na quantidade de dados tem aberto novas oportunidades para a saúde brasileira. Entre as várias novidades proporcionadas pelo big data, destaca-se o uso de modelos preditivos de machine learning para melhorar a qualidade e a avaliação dos serviços de saúde. A palestra tem como objetivo apresentar aplicações práticas desses modelos na área da saúde, além de seus benefícios e limitações.

FlexCode: modelando incertezas em problemas de predição

Rafael Izbicki
UFSCar

Resumo

Grande parte das ferramentas de aprendizado de máquina tem como objetivo criar boas predições. Contudo, raramente é possível fazê-las com 100% de acurácia. Assim, em muitas aplicações, apenas fornecer predições não explora toda a informação presente nos dados. Nesta apresentação, mostraremos uma ferramenta que é capaz de modelar incertezas em problemas de predição. Também mostraremos seu desempenho na predição da geolocalização de tweets, assim como para diversos problemas de cosmologia.

Multinomial regression via binomial regressions

Guaraci Requena – IME-USP,
Carlos Alberto de Bragança Pereira – UFMT/IME-USP e
Adriano Polpo – UFSCar

Resumo

The most used multinomial regression model is the baseline-category logit. However, this is not the only way to build it, neither in relation to the baseline category nor in relation to the logit link function. As we may factorize the multinomial distribution for D categories in terms of D-1 binomial ones - through recursive and exhaustive binary partitioning of the set of categories - we may define the multinomial regression in terms of binomial ones, bringing all the flexibility concerning the link functions. Likewise, to define D-1 binary classifiers - from those binomial regressions - leads us to build a multi-class classifier. Unfortunately, the class of factorizations could be very extensive according to the number of categories (approximately 35×10^6 for 10 categories, for example), so we could have a very large class of distinct multinomial models and/or multi-class classifiers. Facing this problem, we suggest two step-by-step approaches through minimizations of involved binary classification risks, based on the one-versus-one and one-versus-rest approaches. In order to study their performances, we apply them in a psychiatric problem, precisely in Obsessive-Compulsive Disorder, in which the aim is to classify the patient, who has features observed from a global severity scale (Y-BOCS), in a dimensional severity scale (DY-BOCS), seeking a more precise phenotype.

Manifold learning for non-linear dimensionality reduction in hyperspectral image unsupervised classification

Eduardo Kazuo Nakao e
Alexandre Luis Magalhães Levada
UFSCar

Resumo

Hyperspectral images characteristics of high dimensionality and strong inter-pixel correlation indicate that its possible that the vectors of those images matrixes are embedded in an non-linear manifold instead of an euclidean one. To help elucidating this supposition, one can try to reduce those images dimensionality using linear and manifold learning methods prior to a clustering algorithm and then compute an evaluation metric on the different results. If the validation index scores higher on non-linear reduction clustering results, this is an indication that hyperspectral images in fact contains non-linear relations. One approach of experimenting on this scenario for example is executing Principal Component Analysis, Isometric Feature Mapping and Locally Linear Embedding (each one in separate fashion) and then deploy K-Means and Gaussian Mixture Model clustering methods at each separate result for several different images. Then the performance can be measured by Kappa Coefficient if original classes labels are provided (external evaluation). It's worth noticing that the selection of target reduced dimensionality must be done in basis of some know criteria, for example the division of the first largest eigenvalues of the reduction methods transformation matrices by the sum of all of those eigenvalues. Another point to notice is that the External clustering evaluation criteria only works well when the clustering generated class labels are semantically the same as the original classes labels. This can be achieved by solving the minimal pairing cost on a bipartite graph

(this optimal allocation problem can be solved by the Munkres algorithm for example). There is a known dataset of AVIRIS sensor hyperspectral images that can be used for experimentation.

Aplicação do FBST em modelos Bayesianos de alta dimensão

João Carlos Poloniato Ferreira,
Rafael Bassi Stern e
Rafael Izbicki
UFSCar

Resumo

Neste trabalho estudamos o problema de controlar o nível significância do Full Bayesian Significant Test (FBST) em modelos para densidade de probabilidade. Para isto, mostramos um método que define uma posteriori da densidade de probabilidade com infinitos parâmetros. Para conduzirmos o FBST nessa situação introduzimos a definição do e-valor modificado que é uma maneira de calcular a medida de evidência do FBST controlando o nível de significância do teste já que o cálculo usual não apresenta bons resultados quando são testados muitos parâmetros. Apresentamos então os resultados de um estudo de simulação com diferentes distribuições de densidade analisando o comportamento da função poder do FBST comparada com a função poder do teste de Kolmogorov-Smirnov (KS).

Um experimento sobre previsões probabilísticas

Marcio Alves Diniz, Rafael Izbicki
Danilo Lourenço Lopes e Luis Ernesto Salasar
UFSCar

Resumo

Durante a última Copa do Mundo lançamos a plataforma "Fifa Experts", onde as pessoas podiam informar as probabilidades que atribuíam a cada possível resultado dos jogos da Copa. Depois de cada jogo, as previsões recebiam uma pontuação e os participantes eram classificados. Dois modelos matemáticos também foram incluídos como participantes. Nesta apresentação discutimos brevemente a experiência e apresentamos resultados preliminares da análise dos dados coletados.

Firm location: an approach using spatial point process

Adriano Barasal Morales e Márcio Poletti Laurini
FEARP-USP

Resumo

We propose an application of spatial statistics to model the location patterns of new services firms in the city of São Paulo. In this paper, we assume that the spatial location of these firms was generated through a two-dimensional point process and thus we applied two distinct models: one based on non-stochastic intensity based on the Poisson process, and a stochastic intensity model based on the Log Gaussian Cox process (LGCP). The results show the usefulness of these models the construction of spatial location models, combining different data sources and introducing new perspectives on the empirical study of location economics. Keywords: Firm location, spatial statistics, Poisson point process, LGCP, INLA.

Mineração de regras de associação espaço-temporais temáticas aplicada a imagens de explosões solares

Carlos Roberto Silveira Junior, Marcela Xavier Ribeiro e Marilde Terezinha Prado Santos
UFSCar

Resumo

Introdução. A análise de clima espacial é uma tarefa complexa que envolve dados espaço-temporais provenientes de imagens de satélite somado a dados de boletins diários. Tais dados são caracterizados como séries temporais de imagens georeferenciadas e séries temporais de dados semânticos (dados alfanuméricos que descrevem as imagens), respectivamente. A mineração de regras de associação pode auxiliar na análise desses dados, como um mecanismo para a revelação de padrões novos e úteis para o especialista de domínio. No entanto, os métodos existentes de mineração de regras de associação espaço-temporais ainda são limitados e, em consequência disso, não atendem adequadamente às expectativas para extração de padrões que relacionam informações espaço-temporais em imagens e dados semânticos. Objetivo. Assim sendo, este trabalho tem por objetivo apoiar a análise do clima espacial a partir do desenvolvimento de um método de mineração de regras de associação espaço-temporais que permita relacionar dados solares semânticos e visuais. O foco são séries de imagens solares oriundas de satélites. Proposta. O método desenvolvido é composto por: um novo processo de ETL - direcionado ao domínio solar; um novo algoritmo de mineração de regras de associação espaço-temporais, e; um novo classificador que utiliza as regras espaço-temporais para determinar o comportamento futuro de novos dados solares. O algoritmo de mineração proposto avança o atual estado da arte da área de mineração de regras de associação por dividir a aplicação das restrições espaço-temporais em duas etapas diferentes do

processamento: a aplicação das restrições espaciais é feita durante a extração de itemsets frequentes e a aplicação das restrições temporais durante a geração das regras de associação espaço-temporais temáticas. Desta forma, é possível a obtenção de regras que representam a evolução de um determinado conjunto de eventos e como eles se relacionam entre si. Por fim, essas regras são utilizadas pelo classificador associativo que foi proposto neste trabalho para predizer o comportamento solar com base em suas características visuais atuais. Resultados. O método proposto gerou regras que foram usadas para a classificação, apresentando uma precisão de até 87,3% na classificação de imagens solares, sendo que esse valor de precisão varia com o extrator de características utilizado para representar as imagens. A maior precisão (87,3%) foi obtida utilizado SURF como extrator de características e a menor precisão (82,7%) foi utilizado o Histograma como extrator de características. Os resultados obtidos foram analisados pelo especialista de domínio que avaliou como eficaz e válido o método proposto.

Sustentabilidade e curadoria digital para coleções de patrimônio cultural

Ana Carolina Simionato, Maria Ligia Triques,
Débora Marroco Ninin e Marcos Teruo Ouchi
PPGCI-UFSCar

Resumo

No atual cenário tecnológico e diante ao crescente volume de dados, a área de Ciência da Informação busca métodos mais efetivos para a organização, representação do conteúdo digital de coleções de patrimônio cultural. Assim, objetiva-se a discutir sobre os processos de criação e gerenciamento de dados e metadados, a partir do estudo da Curadoria Digital e do emprego dos modelos de dados, destacando-se a importância da otimização e sustentabilidade do reuso dos dados em centros de informação. Nessa perspectiva, as questões emergentes do atual cenário têm conduzido esses estudos para o viés da manutenção do contexto digital, denominado como Curadoria Digital. A partir dos pressupostos da Curadoria Digital, evidencia a preocupação em assegurar a sobrevivência e o acesso contínuo do material digital, conduzindo a novas práticas teórico-aplicadas para o gerenciamento dos dados. No entanto, a Curadoria Digital envolve diversas ações de gerenciamento de dados, que consistem em identificar, digitalizar, higienizar, descrever, armazenar e preservar, compartilhar e avaliar os dados. A descrição é tomada como princípio e integrado de todas as ações, sem descartar as particularidades de cada tipo de acervo. Para tanto, os processos de representação anteriores a esse cenário, buscavam atender a necessidades específicas às tipologias de acervos, e hoje, as mesmas detentoras das coleções de patrimônio cultural devem agir e repensar o gerenciamento de metadados por meio das atuais tecnologias que fornecem modos comuns e interoperáveis de acesso, uso e reuso de recursos. Como também, devem reconfigurar os procedimentos comuns que delineiam à um retrabalho no uso

dos metadados em arquivos, bibliotecas e museus, a partir de uma sustentabilidade de dados, resultante do planejamento dos sistemas de gerenciamento nos modelos de dados, definição dos metadados e de padrões de metadados. Considera-se que os modelos conceituais acarretam em uma desconstrução do registro, em que os dados de uma estruturação monolítica passam a ser retratados por relações, desse modo, os instrumentos de representação e os catálogos são parte de um novo paradigma de ligação de acervos no ambiente digital. A proporção abstrativa dos modelos de dados configura-se em um espaço de informações que interage com a Web, formando uma rede de dados que integra diversos recursos informacionais. Além disso, seu potencial está na possibilidade de disseminar a terceiros seus conteúdos, promovendo amplo reuso e acesso aos dados, e principalmente, o gerenciamento desse conteúdo e assim, caracterizando a Curadoria Digital.

Generalising dynamic Bayesian networks to accommodate causal and symmetric signals in high-dimensional time series

Osvaldo Anacleto Júnior
ICMC-USP

Resumo

We present the dynamic chain graph model, which extends dynamic Bayesian networks by considering high-dimensional time series exhibiting not only a causal drive mechanism between their components but also symmetric relationships among them. This model can accommodate non-linear and non-normal time series and simplifies computation by decomposing a high-dimensional problem into separate, simpler sub-problems of lower dimensions. The advantages of the new model will be illustrated by forecasting traffic network flows and also by modelling gene expression data from transcriptional networks. A hierarchical extension of the model will be also introduced.

Curva de crescimento fetal personalizada

Agatha Sacramento Rodrigues, Mariza Marie Fujita e
Rossana Pulcineli Vieira Francisco
Departamento de Obstetrícia da FM-USP

Resumo

O acompanhamento do crescimento fetal alerta o obstetra para a necessidade de cuidados assistenciais adequados. Atualmente no Brasil, as curvas de crescimento das medidas biométricas de Hadlock (1991) são usadas como referência. No entanto, estas são curvas baseadas na população americana, que apresenta características diferentes da população brasileira. Ainda, as curvas de Hadlock são funções apenas da idade gestacional e pode ser interessante considerar curvas de crescimento fetal para medidas biométricas que, além da idade gestacional, levem em conta características maternas e da gestação, ou seja, curvas personalizadas.

Foram observados 1445 exames ultrassonográficos em 434 gestações únicas no período gestacional de 12 a 42 semanas cujo parto ocorreu entre 2014 e 2017 no hospital universitário da Universidade de São Paulo (HU/USP), com pelo menos duas avaliações da mesma gestação ao longo do pré natal.

No presente trabalho, construímos curvas de crescimento fetal por meio de modelos lineares mistos ao levar em conta a dependência de exames de uma mesma gestação. Consideramos as seguintes covariáveis: peso materno, altura materna, sexo fetal e número de partos anteriores. No processo de modelagem, 70% dos dados são separados como amostra treinamento e 30% para testar o modelo. Modelos lineares sem e com penalizações de Lasso e de Ridge foram considerados. A escolha dos parâmetros de penalização foi realizada por meio de validação cruzada pelo método de 10-fold. O modelo escolhido é aquele com menor raiz do erro quadrático médio (REQM) na amostra teste. Um aplicativo shiny foi elaborado para visualização das curvas para que

obstetras do Brasil possam calcular a curva esperada de uma gestação dada suas características.

Adaptative significance levels in normal mean hypothesis testing

Alejandra Estefanía Patiño Hoyos e Victor Fossaluza
IME-USP

Resumo

The Full Bayesian Significance Test (FBST) for precise hypotheses was presented by Pereira and Stern [Entropy 1(4) (1999) 99-110] as a Bayesian alternative instaed of the traditional significance test using p-value. The FBST is based on the evidence in favor of the null hypothesis (H0). An important practical issue for the implementation of the FBST is the determination of how large the evidence must be in order to decide for its rejection. In the Classical significance tests, it is known that p-value decreases as sample size increases, so by setting a single significance level, it usually leads H0 rejection. In the FBST procedure, the evidence in favor of H0 exhibits the same behavior as the p-value when the sample size increases. This suggests that the cut-off point to define the rejection of H0 in the FBST should be a sample size function. In this work, we focus on the case of two-sided normal mean hypothesis testing and present a method to find a cut-off value for the evidence in the FBST, by minimizing the linear combination of the type I error probability and the expected type II error probability for a given sample size.

Performance analysis of deep neural networks in piRNAs classification

Alisson Hayasi da Costa,
Renato Augusto Correa dos Santos
e Ricardo Cerri
UFSCar

Resumo

Modern machine learning techniques, such as Deep Learning, have been successful in many complex Bioinformatics tasks. The capacity of Deep Neural Networks to handle large volumes of data has made them essential tools for multiple areas of knowledge. However, developing the best model for a given task is a hard work. Deep Neural Networks have a very large number of hyperparameters, making them as powerful as complex to be adjusted. Therefore, in order to better understand the behavior of Deep Neural Networks when applied to biological data, we present in this paper a performance analysis of a Deep Feedforward Network in piRNAs classification. Different configurations of activation functions, initialization of weights, number of layers and learning rate are experienced. The effects of different hyperparameters are discussed and certain organizations are proposed for similar domains of data.

Recursos sonoros e linked open data: MusicBrainz

Amanda Azevedo dos Santos e Ana Carolina Simionato
UFSCar

Resumo

Diante da ascensão do uso e desenvolvimento das tecnologias da informação e comunicação (TIC), produção e compartilhamento de informações e conteúdos digitais tornou-se parte da rotina, em reflexo a esse novo comportamento, é difícil definir e contabilizar o crescimento do volume informacional. Crescimento que propicia dificuldades para localização e acesso de diversos recursos informacionais como imagens e músicas. A música como arquivo no formato MP3 apresenta dados denominado Identify a MP3 (ID3), "Para representação de músicas em formato MP3 existe o padrão ID3, que é um conjunto de metadados incorporado ao próprio arquivo de áudio.". (FERREIRA, 2015, p.13). Os metadados também podem ser assimilados como descritores ou atributos, que podem enriquecer, identificar e auxiliar na interação entre dados e softwares. (POLLOK, 2011). Frente ao desenvolvimento de estudos sobre internet, o pesquisador Tim Berners-Lee criou um ambiente que usava uma rede de comunicação, popularmente conhecida como internet que viabiliza o compartilhamento de arquivos, textos, áudios, imagens e vídeos ambiente que foi nomeado como Wide Word Web ou WWW. (BERNERS-LEE, 1989). Assim, por meio do estudo e desenvolvimento dessas tecnologias, surge o Linked Data, no Linked Open Data é possível localizar datasets relacionados ao contexto musical, e analisar o modo como esse recurso é oferecido aos usuários, assim como, quais ferramentas, metadados utilizados para aperfeiçoar a recuperação deste conteúdo. Um dataset sobre música é o MusicBrainz MusicBrainz é um projeto que fornece dados como identificadores únicos e específicos no contexto musical, como artistas de música, álbuns e as músicas, com o uso de URIs relacionados com música, o MusicBrainz também desenvolve produtos.

Seu banco de dados denominado MusicBrainz Database é estruturado pelo PostgreSQL e contém os metadados das músicas e quanto aos metadados utilizados eles segundo Metabrainz (2017, não paginado, tradução nossa): Artista: Nome, nome de classificação, apelido, tipo, datas de início e término, comentário de desambiguação; Grupos de liberação: Título, crédito do artista, tipo, comentário de desambiguação; Lançamentos: Título, crédito do artista, tipo, status, idioma, data, país, rótulo, número de catálogo, código de barras, meio (s), ID (s) de disco, comentário de desambiguação; Suporte: Formato, lista de faixas (título, crédito do artista, duração) Gravação: Título, crédito do artista, duração, relacionamentos, comentário de desambiguação; Trabalho: Título, relações, comentário de desambiguação; Etiquetas: Nome, nome de classificação, apelido, país, tipo, código, datas de início e término, comentário de desambiguação; Relações e URLs: Os relacionamentos são uma maneira de vincular as entidades acima e permitir que o MusicBrainz capture a maioria dos dados contidos nas notas de linha de um CD. CD Stubs: Título, artista, código de barras, ID do disco, comentário de desambiguação.

Dados abertos: criação de novos negócios

Aulida Berenice Moretti dos Santos e
Ana Rita Tiradentes Terra Argoud
FATEC

Resumo

Com o avanço nas tecnologias de informação e comunicação criou-se um novo cenário no uso e compartilhamento de dados, com o advento da Internet foi possível considerar políticas de acesso aberto às publicações científicas. Atualmente, visando à transparência e colaboração sob diversos aspectos, cada vez mais o conceito de dados abertos vem sendo abordado, sendo tendência que haja essa abertura em dados governamentais e dados científicos. Os dados abertos, conhecido também como o movimento Open Data, surgiu com o intuito de oferecer transparência na divulgação dos dados e sua possível reutilização. A Open Knowledge International é uma importante fundação sem fins lucrativos que visa incentivar o uso dos dados abertos na sociedade, e define que: "Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras." (OPEN KNOWLEDGE INTERNATIONAL, s.d.). Os dados abertos têm influencia na construção de novas informações por meio de sua reutilização, trazendo a oportunidade de investimento tanto de instituições ou organizações que pretendem transformá-los em um conhecimento que traga benefícios para a sociedade. A Open Knowledge Brasil desenvolve e possui parceria em alguns projetos nesse sentido, como o "Índice de Dados Abertos", "Gastos Abertos", "Vai Mudar", etc. Assim Santarem Segundo (2015) afirma que as instituições no âmbito público ou privado vem investindo na organização de acesso à informação, pois consideram um bom diferencial no que tange tomada de decisão em diversas instâncias. Com a política de Dados Governamentais Abertos (DGA) criado em 2009 pelo

então presidente dos Estados Unidos Barak Obama, essa política disponibiliza dados governamentais de domínio público para a livre utilização dos cidadãos. Grandes números de pesquisadores e empresa tem buscado mais informações sobre DGA, além de compreender seus atributos e características ligados ao DGA. A abertura da base de dados do governo vai além da transparência e combate à corrupção que já se justifica, ela também é um meio de incentivo ao uso desses dados abertos para a criação de ferramentas e aplicativos, além de novas empresas ou as já existentes utilizam esses dados em seus serviços e produtos criando novos modelos de negócios. Dentro do contexto de serviços voltados para a população que fazem uso do DGA, tem se o "Cadê o ônibus?" (NANO, 2012) trata-se de um aplicativo para smartphones que utiliza as informações de linhas, rotas, horários de saída e chegada do site da SPTrans, que é uma empresa pública de transporte no município de São Paulo.

Feature extraction for multi-target learning

Brendon Gouveia Cambuí
UFSCar

Resumo

Multi-target learning is a generalization of the recently-popularized task of multi-label classification, where each data instance is associated with multiple target-variables simultaneously. The main challenges in this research field are related to the high dimensionality of data present in datasets with such characteristics, and also the high number of target-variables having dependencies among them. In such scenarios, it is crucial to extract lower-dimensional representations from the original input-space, such that these can be provided as input to other multi-target predictors. In this research, are proposed the use of Auto-Encoders and Restricted Boltzmann Machines as feature extractors in some of multi-target datasets publicly available. Results will be evaluated considering state-of-the-art multi-target prediction methods and evaluation measures in the literature.

Predicting protein functions via interaction prediction

Bruna Zamith Santos, Ricardo Cerri e Celine Vens
UFSCar

Resumo

Proteins are macro-molecules responsible for virtually every task necessary for the maintenance of cells, having a fundamental role in the behavior and regulation of organisms. Advances in the area of Molecular Biology have allowed an almost complete listing of the proteins that make up the organisms. However, there are a large number of proteins whose function is still unknown, opening space for a new research focus in Molecular Biology. Usually, protein function prediction is performed using homology-based Bioinformatic tools, comparing a sequence with a database with many sequences belonging to previously known functions. This is a limited strategy, since it ignores the sequences' biochemical properties, and also the hierarchical relationships that may exist between the different classes. In the literature, the use of Machine Learning for the protein function prediction has shown to be promising, obtaining significant advances regarding the use of homology and other methods. Making use of Machine Learning, it is possible to model the protein function prediction problem as a Hierarchical Multi-label Classification (HMC) problem, due to the fact that protein functions are hierarchically organized and that they can occur simultaneously. Among several HMC algorithms known in the literature, some of them have treated HMC tasks considering interaction data. Interaction data are characterized by two sets of objects, each described by their own set of features, which makes it possible to predict the interactions between two instances. They are often represented as a network of relationships. Such algorithms assign a function to a protein based on the functional labels of its interacting neighbours. However, none of these methods model the HMC problem as an

interaction data problem. This project proposes modeling the protein function prediction task as a HMC problem through interaction data. Thereby, a new method for HMC of protein functions, which makes use of interaction prediction, is developed and studied.

Comparação de testes de hipóteses para duas médias em variáveis do tipo proporção

Camila Lorencetti Brolo e
Gustavo Henrique de Araújo Pereira
UFSCar

Resumo

O estudo de taxas e proporções é muito comum em diversas áreas do conhecimento. Elas assumem valores no intervalo (0;1) e são denominadas variáveis do tipo proporção. Quando o interesse é comparar a média de uma variável em dois grupos diferentes, é comum a utilização de testes de hipóteses. Neste trabalho, comparamos a performance de três testes de hipóteses para comparação de duas médias em variáveis do tipo proporção com a suposição de que a distribuição da variável resposta é Beta. Um dos testes é o mais tradicional e os outros são baseados em bootstrap. Em cada um dos cenários considerados, através de simulação de Monte Carlo, obtemos para cada método estimativas do tamanho e do poder do teste para a comparação de duas médias e por fim, aplicamos a dados reais. Dentre os resultados obtidos, salientamos a boa performance do método bootstrap 1 para o erro do tipo I, quando temos valores iguais nas duas populações do parâmetro de precisão. Entretanto, para valores diferentes do parâmetro de precisão, o método bootstrap 2 é o que melhor se comporta. Para o poder do teste, salientamos também a boa performance do teste de hipótese bootstrap 1 na grande maioria dos cenários.

Redes Bayesianas: uma comparação entre métodos de estimação de estrutura

Camila Sgarioni Ozelame – UFSCar,
Anderson Ara – UFBA,
Francisco Louzada Neto – ICMC-USP,
Marcos Jardel Henriques – USP/UFSCar e
Oilson Alberto Gonzatto Junior – USP/UFSCar

Resumo

A técnica de Redes Bayesianas baseia-se na representação da distribuição conjunta de um grupo de variáveis aleatórias através de um grafo acíclico direcionado (DAG - *Directed Acyclic Graph*), sendo tais variáveis representadas por nós da rede e a dependência condicional sendo representada por arcos. Deste modo, a tarefa de modelagem das conexões entre as variáveis pode reduzir a dimensionalidade do banco de dados e permitir melhor interpretação das variáveis envolvidas.

Neste trabalho, comparamos a estimação da estrutura das redes utilizando os algoritmos clássicos K2 e PC. O primeiro considera, inicialmente, a independência entre os nós e a cada passo adiciona a relação mais provável entre as variáveis, otimizando um escore específico de qualidade de ajuste, o qual avalia a rede como uma função dos dados. O segundo, considera testes estatísticos baseado em métrica de informação mútua condicional e de critério de d-separação para orientar os arcos. Neste contexto, a metodologia implementada considera variáveis aleatórias categóricas e uma variável resposta, sendo então redes de classificação.

Os métodos são comparados por meio de dados artificiais com a estrutura conhecida, bem como de dados reais relativos a falhas na plantação de cana de açúcar, neste último a problemática da empresa gira em torno da dificuldade de diminuir as falhas nos talhões de cana plantados em suas terras. Para quantificar a qualidade do das estruturas propostas são utilizadas as medidas de AIC e BIC, além disso de quatro medidas de performance

de desempenho: sensibilidade, especificidade, acurácia e o coeficiente de correlação de Matthews.”

Momentos de vida: mineração de textos e Big Data para oferecimento de serviços personalizados

Dhiogo José Corrêa de Sá, Thiago de Paulo Faleiros, Priscilla de Abreu Lopes
Ricardo B. Scheicher e Eduardo F. Velludo Prado
Itera - Inovação e Desenvolvimento Tecnológico

Resumo

Em 2016 a Itera foi selecionada para participar do programa InovaBra do Bradesco. O desafio proposto era o oferecimento de serviços personalizados, maximizando a conversão de vendas e engajamento com os clientes. Foi disponibilizado um conjunto de dados com 1Mi de transações bancárias de clientes, com dados como identificação do cliente, data e valor da transação e texto de segunda linha (texto curto que descreve a transação). Com o auxílio de especialistas, foram definidas 32 categorias de gastos, sendo 26 gerais (”veículo”, ”mercado”) e 6 para produtos financeiros (”seguro”, ”cartão”), utilizadas para rotulação manual de 32 mil transações (1.000 por categoria). A geração da solução foi realizada em 5 iterações do processo: pré-processamento do texto de segunda linha, geração de modelo de classificação utilizando a técnica semissupervisionada Transductive Classification based on Bipartite Heterogeneus Network (TCBHN), classificação automática de 450Mi de transações utilizando tecnologia de processamento distribuído e validação de amostra classificada pelos especialistas, sendo que erros encontrados eram utilizados para retreinamento do modelo. Após validação final do modelo, foi realizada a segmentação de transações, definindo 10 quantis para cada categoria de transação, baseados no valor gasto. Foi realizada a recategorização de transações, incluindo a informação de quantis, e.g. ”mercadoQ5”, ”seguroQ7”. Para cada cliente, foram montadas sacolas de compras mensais, documentos compostos pelas categorias das transações efetuadas no mês de re-

ferência concatenadas. Os documentos gerados foram agrupados e tópicos foram extraídos dos grupos utilizando a técnica Latent Dirichlet Allocation (LDA), executada utilizando processamento distribuído. A análise dos tópicos e de clientes específicos proporcionou a identificação de Momentos de Vida (reforma de casa, viagem) que foi utilizada pelo banco para oferecimento de serviços personalizados a seus clientes.

Obtenção de curvas de confiabilidade em testes de vida acelerados

Diogo Barboza Moreira e Teresa Cristina Martins Dias
UFSCar

Resumo

Na área de confiabilidade em Estatística, um dos interesses está em estimar os parâmetros envolvidos no modelo que descreve o comportamento de falha e a função de confiabilidade dos produtos. Consideramos um experimento que envolve a aplicação de testes de vida acelerados, com a finalidade de modelar o comportamento de falha, utilizando unidades amostrais. Tais unidades são submetidas à condições de funcionamento não usuais, através do aumento dos níveis de variáveis que influenciam no tempo até a ocorrência do evento. São exemplos de variáveis de estresse: temperatura, voltagem e corrosão. Assumimos os modelos exponencial ou Weibull para os tempos e a relação da variável de estresse com o tempo de vida dada pelos modelos de lei de potência ou Arrhenius, sob o esquema de censura à direita, do tipo II. Neste trabalho apresentamos estimativas para a função de confiabilidade, sob diversos níveis de estresse, nos cenários citados. Os estimadores foram obtidos via método da máxima verossimilhança e implementados em um programa, criado no software R, que retorna as estimativas e as curvas de confiabilidade para tempos simulados e conjuntos de dados reais.

Proposta de um modelo de aprendizado competitivo para classificação hierárquica multirrótulo

Elaine Cecília Gatto – Faculdade Anhanguera de Bauru/UFSCar e
Ricardo Cerri – UFSCar

Resumo

A Classificação Hierárquica Multirrótulo (CHM) é um problema desafiador da área de Aprendizado de Máquina (AM), sendo considerada uma tarefa complexa dentro da Classificação de Dados, possuindo aplicações em áreas como Bioinformática, classificação de textos e imagens. Um problema de CHM pode ser formalizado como possuindo um espaço de exemplos X; um conjunto de classes C; uma ordem parcial que representa o relacionamento superclasse $\leq h$; sendo que $\forall c_1, c_2 \in C : c_1 \leq hc_2 \iff c_1$ uma superclasse de c_2; uma hierarquia de Classes $(C, \leq h)$, um conjunto de tuplas (x_i, C_i) sendo $x_i \in X$, $C_i \subseteq C \mid c \in C_i \longrightarrow c' \leq hc : c' \in C_i$; um critério de qualidade q que recompensa modelos com alto desempenho preditivo e baixa complexidade, e por fim, uma função $f : X \longrightarrow 2C$; sendo $2C$ o conjunto de potência de $C, \mid c \in f(x) \longrightarrow \forall c' \leq hc : c' \in f(x)$ e f otimiza q. As classes em um problema de CHM podem ser organizadas como uma Árvore ou como um Grafo Acíclico Direcionado. Dada essa taxonomia, os algoritmos de AM devem rotular objetos como pertencentes a múltiplos caminhos simultaneamente. Abordagens Competitivas, que aplicam aprendizado não supervisionado, têm sido recentemente aplicadas para resolução de problemas envolvendo CHM, porém, ainda há poucos trabalhos que relatam soluções aplicando Abordagens Competitivas Híbridas, mesclando aprendizado supervisionado e não-supervisionado. Portanto, este projeto de pesquisa tem como objetivo investigar como o aprendizado competitivo, usando redes neurais artificiais, pode colaborar para tarefas de CHM. Abordagens não supervisionadas podem ser interessantes dado

que, quanto mais profunda uma classe na hierarquia, menos exemplos positivos ela possui, dificultando o aprendizado supervisionado.

Estimadores, pontual e intervalar, para dados com censuras intervalar

Estela Maris Pereira Bereta e Teresa Cristina Martins Dias
UFSCar

Resumo

Dados com censura intervalar ocorrem com frequência em estudos de diversas áreas, em situações em que o evento de interesse é observado com periodicidade. Neste caso, o tempo exato da ocorrência não é conhecido (observado), porém sabe-se que o evento ocorreu dentro de um intervalo (conhecido) de tempo. Este tipo de observação é tratada em análise de sobrevivência usando técnicas apropriadas que considera censura intervalar. Pradhan e Kundu (2014) apresentam vários métodos de estimação pontual (algoritmo EM, aproximação de Lindley e *importance sampling*) no caso de tempos censurados de forma intervalar, com distribuição exponencial e Weibull. Também, os autores apresentam um algoritmo para a construção de intervalos de confiança, na abordagem Bayesiana. Sob a abordagem Bayesiana, obtemos estimativas pontuais e, para as estimativas intervalares apresentamos uma alternativa ao método proposto por Pradhan e Kundu (2014). Para ilustrar a teoria, simulamos dados para diferentes tamanhos amostrais, considerando tempos com distribuição Weibull.

Haphazard intentional sampling techniques in network design of monitoring stations

Julio M. Stern
IME-USP

Resumo

In contemporary empirical science, sampling randomization is the the golden standard to ensure unbiased, impartial, or fair results, see Pearl (2000) and Stern (2008). Randomization works as a firewall, a technological barrier designed to prevent spurious communication of vested interests or illegitimate interference between parties in the application of interest, that may be a scientific experiment, a legal case, an auditing process, or many other practical applications. In randomized experiments, a common issue is avoid random allocations yielding groups that differ meaningfully with respect to relevant covariates. This is a critical issue, as the chance of at least one covariate showing a ”significant difference”between two treatment groups increases exponentially with the number of covariates. Haphazard Intentional Sampling is a statistical technique developed with the specific purpose of yielding sampling techniques that, on one hand, have all the benefits of standard randomization and, on the other hand, avoid exponentially large (and costly) sample sizes. This approach, proposed at Lauretto et al (2012) and Fossaluza et al (2015), combines intentional sampling using goal optimization techniques with random perturbations that induce good decoupling properties. On one hand, this method has a computational cost that is cubic with the number of covariates. On the other hand, this method yields experimental designs that avoid exponentially large sample sizes, allowing great economical benefits that, nevertheless, do not compromise the statistical integrity of the experiment or auditing process. In this article, we apply the aforementioned Haphazard Intentional Sampling as a statistical technique to study how to rationally re-engineer networks

of measurement stations for atmospheric pollution and/or gas emissions. We show how such re-engineering or re-design can substantially decrease the operation cost of monitoring networks while providing, at the same time, support for arriving at conclusions or taking decisions with the same statistical power as in conventional setups.

Random Bernstein polynomials – a nonparametric bayesian estimation of densities via ABC

Leandro Augusto Ferreira – IME-USP/Sharecare e
Victor Fossaluza – IME-USP

Resumo

In recent years, many statistical inference problems have been solved by using Markov Chain Monte Carlo (MCMC) techniques. However, it is necessary to derivate the analytical form for the likelihood function. Although the level of computing has increased steadily, there is a limitation caused by the difficulty or the misunderstanding of how computing the likelihood function. The Approximate Bayesian Computation (ABC) method dispenses the use of the likelihood function by simulating candidates of posterior distributions and using an algorithm to accept or reject the proposed candidates. This work presents an alternative nonparametric estimation method of smoothing empirical distributions with random Bernstein polynomials via ABC method. The Bernstein prior is obtained by rewriting the Bernstein polynomial in terms of k mixtures of beta densities and mixing weights. Some examples are used to illustrate the method proposed.

Semi-supervised predictive clustering tree for protein subcellular localization

Leonardo Utida Alcântara – UFSCar,
Ricardo Cerri – UFSCar e
Isaac Triguero Velázquez – University of Nottingham

Resumo

The prediction of protein subcellular localization is a really important classification task, because the location of proteins inside a cell is directly related to these protein's functions. There are many proteins that reside at the same time in two or more locations within a cell or move between multiple locations, because of this we need to attack this problem using multi-label classification (MLC) algorithms. The supervised MLC approach is well-established in the literature; however, it presents some disadvantages such as: (i) the need for a large amount of labelled instances to train the classifier; (ii) this approach ignores the fact that unlabelled instances can provide valuable information for the classification; and (iii) there are a lot of areas, including bioinformatics, in which unlabelled data is abundant but manually labelling an instance is too expensive and time-consuming. Semi-Supervised Learning (SSL) is a subfield of machine learning, in which the learner tries to exploit both labelled and unlabelled data at the same time to improve the accuracy of the learning algorithm. The main goal of this project was to investigate how to use SSL to improve the classification in MLC datasets, focusing on the protein subcellular localization prediction problem. In order to do this, we proposed a new SSL algorithm based on the predictive clustering tree framework, that is capable of constructing a decision tree using the information from both labelled and unlabelled instances to decide the best attribute to split the data at each node. We tested our new approach in many SSL scenarios for inductive and transductive learning, evaluating the accuracy of our classifier in many multi-label datasets (MLDs),

3 proteins datasets (from virus, plants and fungus) and 4 well-known multi-label datasets. The results showed that our method could improve the classification in the SSL scenarios for most of the MLDs, proving that our method is able to exploit labelled and unlabelled data.

Implementação de Inferência Bayesiana não Paramétrica para Processos Pontuais Espaciais

Lucas Eduardo de Moraes e Danilo Lourenço Lopes
UFSCar

Resumo

Padrões pontuais espaciais são dados coletados na forma de pontos, distribuídos dentro de uma região do espaço surgem em diversas situações, como na biologia, epidemiologia ou criminologia. Neste tipo de dados, as localizações são chamadas de eventos, e tanto a quantidade quanto as localizações desses eventos são aleatórias. Padrões pontuais podem apresentar comportamentos como: agrupamento, inibição, regularidade ou não apresentar nenhuma estrutura específica. Um dos objetivos principais da análise de dados deste tipo é entender e descrever tais padrões pontuais. Neste trabalho apresentamos uma base para entender o que caracteriza um processo pontual, modelos que buscam descrever da melhor forma possível padrões observados e apresentar métodos de estimação de características de interesse, como a função de intensidade, utilizando técnicas de inferência Bayesiana não paramétrica.

Alocação intencional fortuita: um estudo de caso em avaliação de software

Marcelo de Souza Lauretto – EACH-USP,
Higor A. de Souza – IME-USP e
Marcos L. Chaim – EACH-USP

Resumo

Métodos de alocação intencional são procedimentos não probabilísticos de seleção e alocação de indivíduos, com o objetivo de atingir critérios de representatividade e balanceamento. Tal abordagem é indicada para pesquisas exploratórias ou estudos-piloto nos quais a raridade de indivíduos, restrições éticas ou de custo limitam severamente os tamanhos das amostras e impedem a adoção da amostragem aleatória tradicional. Em trabalhos anteriores, apresentamos a *alocação intencional fortuita*, um método de alocação baseado no balanceamento ótimo de covariáveis de interesse, combinado com perturbações aleatórias. Neste trabalho, estendemos a abordagem de alocação fortuita e apresentamos um estudo de caso em avaliação de software. Experimentos numéricos mostram que a alocação fortuita provê grupos experimentais bem balanceados, mesmo na presença de um baixo número de participantes.

Using Markov Chain in Data Science

Márcio Luis Lanfredi Viola – UFSCar,
Jesús E. García – UNICAMP e
Verônica A. González-López – UNICAMP

Resumo

In this work we show the possibility of using the Markov chain as a tools for data analysis, particularly, in data science. Moreover, we show an example that involves statistical classification of languages according to their rhythmic features, using speech samples. The data set consists of sentences belonging to eight languages, Catalan, Dutch, English, French, Italian, Japanese, Polish and Spanish. This is an important open problem in phonology. The linguistic conjecture is that this languages are divided into three classes according to their rhythmic properties: stress-timed, syllable-timed or mora-timed. A persistent difficulty on this issue is that the speech samples correspond to several sentences produced by diverse speakers, corresponding to a mixture of distributions. The usual procedure to deal with this problem has been to choose a subset of the complete sample which seems to best represent each language. The dataset used for this analysis are recordings of various speakers, who read the phrases into a microphone. To classify the languages, firstly we will fit a model for each language, using our robust procedure and then, we use the relative entropy between the models (for each language) as a distance for the clustering of the languages.

Uma abordagem distribuída para sistemas fuzzy evolutivos multiobjetivos em problemas de Big Data

Maykon Rocha Santana
UFSCar

Resumo

Abordagens têm sido aplicadas para a geração automática de Sistemas Fuzzy a partir de conjunto de dados e com o uso de técnicas como Algoritmos Genéticos, Redes Neurais, dentre outras. Nesse cenário, questões relacionadas à dimensionalidade, à precisão e à complexidade dos sistemas obtidos tem sido campo de estudo em diversos trabalhos. Os pesquisadores do Laboratório CIG (Computational Intelligence Group) do Departamento de Computação da Universidade Federal de São Carlos (UFSCar) - São Paulo - Brasil têm desenvolvido trabalhos voltados para, a partir de conjuntos de dados, realizar a extração das Bases de Regras necessárias para a geração dos chamados Sistemas Fuzzy Baseados em Regras (SFBR). Em geral, são usados nesses trabalhos os Algoritmos Genéticos Multiobjetivos, em especial o NSGA-II, para tratar as questões referente ao balanceamento entre a interpretabilidade e a precisão dos conjuntos de regras que fazem parte dos Sistemas Fuzzy construídos. Todavia, quando se considera contextos em que os dados têm grande dimensionalidade, como em problemas de Big Data, a geração dos SFBR incorre em processamentos bastante custosos. Tendo isso em vista, é proposto o desenvolvimento de uma abordagem que viabilize escalabilidade quando do uso da computação paralela na geração de Sistemas Fuzzy com maximização da precisão (maior acurácia) e minimização da complexidade (menor número de regras e antecedentes das regras) em contextos de problemas de Big Data. O framework de computação paralela Apache Spark será usado para explorar as características de paralelismo existentes nos problemas de geração de SFBR. A eficácia da abordagem será averiguada por intermédio de experimentos realizados

em conjuntos de dados que envolvam dados de grande dimensionalidade no que diz respeito às precisões, complexidades, convergências e escalabilidades dos sistemas obtidos.

Análise estatística do desempenho do PT nas eleições proporcionais entre 2000 e 2016, e previsões para 2018

Pedro Luiz Paolino Chaim e Márcio Poletti Laurini
FEARP-USP

Resumo

Neste artigo realizamos uma análise estatística do desempenho do PT (Partido dos Trabalhadores) nas eleições proporcionais brasileiras de 2000 a 2016. Então, utilizando um modelo dinâmico de regressão Beta, aumentado com um componente de efeitos aleatórios correlacionado no espaço e no tempo, apresentamos previsões para a proporção de votos válidos recebidos pelo PT para deputado estadual e deputado federal nas eleições de 2018. Nossos resultados apontam para a relevância de se considerar a elementos geográficos na caracterização do comportamento do eleitor brasileiro. Previsões sugerem piora considerável no desempenho do PT nas eleições de 2018, quando comparado a 2014.

Mineração de regras de associação temporais envolvendo dados quantitativos contínuos

Rafael Stoffalette João, Laís Vilioni e Silva, Marcela X. Ribeiro
UFSCar

Resumo

O crescente volume de dados gerados constantemente demanda de técnicas cada vez mais eficientes para a aquisição de informações úteis. O presente estudo objetiva a definição e construção de um novo método para refinar o processo de mineração de regras de associação ao incorporar o aspecto temporal de forma explícita, associado a dados quantitativos contínuos. Cada ponto temporal em que uma característica (atributo) assume um valor de interesse dependente da distribuição de probabilidade que a representa. Os pontos temporais são, então, componentes de um intervalo de interesse da característica. As relações temporais entre os intervalos de interesse das características são mapeadas por meio do uso da Álgebra Intervalar de Allen. Padrões podem ser identificados e regras de associação construídas no conjunto resultante de relações temporais de interesse. Até o presente momento, os experimentos resultaram regras estruturalmente semelhantes àquelas geradas pela mineração de regras de associação tradicional. Entretanto, as regras são diretamente relacionadas aos intervalos temporais associados aos dados. O que evidencia que a técnica é capaz de gerar regras de associação com uma nova semântica que expressa quais as relações entre a duração e ocorrência eventos importantes presentes na base de dados. Ao mesmo passo em que é capaz de lidar com dados quantitativos contínuos sem a necessidade da tarefa de discretizar os dados. Em termos de eficiência computacional, a nova estratégia não implicou em acréscimos consideráveis para o tempo de execução do processo de mineração de regras de associação. Esta nova estratégia tem o potencial de refinar o

processo de mineração de regras de associação em dados quantitativos contínuos, pois incorpora o aspecto da temporalidade, que resulta em regras diferentes (com novas informações) daquelas geradas pelo emprego da estratégia tradicional.

Aplicação de aprendizado ativo na tarefa de classificação de textos em fluxo de dados

Suzane Carol de Lima e Heloisa de Arruda Camargo
UFSCar

Resumo

A mineração de dados em fluxo tem como objetivo incorporar a informação do fluxo de dados em evolução ao modelo, sem ter que reaprender o modelo do zero em um processo dinâmico que deve encapsular a coleta de dados, a aprendizagem e a fase de validação em um único ciclo contínuo. Uma das técnicas de mineração de dados é a classificação e um dos desafios da classificação em fluxo de dados é obter todos os dados rotulados para realizar o treinamento do modelo. Uma das recentes abordagens utilizadas para solucionar esse problema é a utilização do aprendizado ativo. O aprendizado ativo visa amenizar o processo de rotulagem dos dados através dos mecanismos de consultas de dados não rotulados. Dessa maneira, é possível reduzir a quantidade de dados rotulados necessários para o treinamento do modelo. Este projeto objetiva desenvolver métodos de classificação de texto com o aprendizado ativo que se adapte as mudanças da distribuição dos dados. Os métodos devem manter um desempenho constante e alta precisão ao longo do tempo. Os métodos desenvolvidos incluem estratégias capazes de realizar a identificação e seleção dos dados mais informativos que se adapte às mudanças que possam ocorrer ao longo do fluxo e a utilização de um oráculo artificial que seja capaz de atribuir os rótulos dos dados selecionados sem a intervenção humana, trazendo uma contribuição adicional às estratégias de consultas de aprendizado ativo em fluxo de dados existentes.

O modelo de cópula de Frank para dados de sobrevivência bivariados: modelagem, estimação Bayesiana e pontos influentes

Taís Roberta Ribeiro - USP/UFSCar e
Adriano K. Suzuki - ICMC-USP

Resumo

Nos dias de hoje está cada vez mais comum nos depararmos com situações em que a suposição de independência entre os tempos de sobrevivência pode não ser válida. Sendo assim, essas prováveis associações entre os tempos de sobrevivência são frequentemente modeladas por meio de modelos de fragilidade. No entanto, uma outra alternativa que vem sendo cada vez mais desenvolvida ultimamente para modelar a dependência entre dados multivariados, é o uso dos modelos de cópulas. Neste trabalho propomos o modelo de cópula de Frank para modelar a dependência de dados bivariados de sobrevivência na presença de covariáveis e observações censuradas. Para fins inferenciais, realizamos uma abordagem bayesiana usando métodos Monte Carlo em Cadeias de Markov (MCMC). Algumas discussões sobre os critérios de seleção de modelos foram apresentadas. Com o objetivo de detectar observações influentes utilizamos o método bayesiano de análise de influência de deleção de casos baseado na divergência. Por fim, mostramos a aplicabilidade dos modelos propostos a conjuntos de dados simulados e reais.

Proposta de geração de regras de classificação multirrótulo simultaneamente eficazes e interpretáveis via otimização multiobjetivo com algoritmos genéticos

Thiago Zafalon Miranda e Ricardo Cerri
UFSCar

Resumo

O recente aumento do interesse por modelos de classificação interpretáveis é, em parte, consequência de regulamentos como o General Data Protection Regulation, que asseguram, entre outros, o direito à explicação a sujeitos afetados negativamente por sistemas de tomada de decisão automática. Neste trabalho, pretende-se apresentar ao que, especificamente, diferentes autores se referem quando utilizam o termo interpretabilidade, que comumente é utilizado para descrever diferentes propriedades de um modelo. Visando a geração de modelos de classificação que têm simultaneamente grande poder preditivo e elevada interpretabilidade, algoritmos genéticos multiobjetivo serão utilizados para gerar conjuntos de regras de classificação multirrótulo. Neste tipo de classificação, diferentemente da tradicional (monorrótulo), as classes a cujos objetos podem pertencer não são mutuamente exclusivas, ou seja, alguns objetos podem pertencer a múltiplas classes simultaneamente. As regras de classificação serão evoluídas de acordo com a abordagem de Pittsburg, ou seja, cada indivíduo da população representará um conjunto de regras (um classificador multirrótulo). Essa abordagem possibilita, a cada iteração do algoritmo, a identificação de uma aproximação do Conjunto Ótimo de Pareto. Resultados preliminares indicam que estratégias simples, como ordenação lexicográfica dos objetivos, não produzem consistentemente resultados satisfatórios; ora as soluções geradas são pouco diversas, ora são pouco interpretáveis.

Agnostic tests can control the type I and type II errors simultaneously

Victor Azevedo Coscrato – USP/UFSCar,
Rafael Izbicki – UFSCar e
Rafael Bassi Stern – UFSCar

Resumo

Despite its common practice, statistical hypothesis testing presents challenges in interpretation. For instance, in the standard frequentist framework there is no control of the type II error. As a result, the non-rejection of the null hypothesis (H0) cannot reasonably be interpreted as its acceptance. We propose that this dilemma can be overcome by using agnostic hypothesis tests, since they can control the type I and II errors simultaneously. In order to make this idea operational, we show how to obtain agnostic hypothesis in typical models. For instance, we show how to build (unbiased) uniformly most powerful agnostic tests and how to obtain agnostic tests from standard p-values. Also, we present conditions such that the above tests can be made logically coherent. Finally, we present examples of consistent agnostic hypothesis tests.