Gustavo Kendi Tsuji

Modelo Totoro Canônico de Trabalho Acadêmico com abnTEX2

São Paulo

2016

Gustavo Kendi Tsuji

Modelo Totoro Canônico de Trabalho Acadêmico com abnTEX2

Modelo canônico de trabalho monográfico acadêmico em conformidade com as normas ABNT apresentado à comunidade de usuários LATEX.

Universidade de São Paulo - USP

Faculdade de Economia, Administração e Ciências Contábeis - FEAUSP

Bacharelado em Administração

Orientador: Alessandra Montini

São Paulo

2016

Gustavo Kendi Tsuji
Modelo Totoro Canônico de
Trabalho Acadêmico com abnTEX2/ Gustavo Kendi Tsuji. – São Paulo, 2016-
45 p. : il. (algumas color.) ; 30 cm.

Orientador: Alessandra Montini

Trabalho de Conclusão de Curso – Universidade de São Paulo - USP
Faculdade de Economia, Administração e Ciências Contábeis - FEAUSP
Bacharelado em Administração, 2016.

1. Big Data 2. Regressão Logística 3. Random Forest 4. Spark I. Orientador. II. Universidade xxx. III. Faculdade de xxx. IV. Título

Gustavo Kendi Tsuji

Modelo Totoro Canônico de Trabalho Acadêmico com abnTEX2

Modelo canônico de trabalho monográfico acadêmico em conformidade com as normas ABNT apresentado à comunidade de usuários LaTeX.

Trabalho aprovado. São Paulo, 28 de junho de 2016:

Alessandra Montini
Orientadora

São Paulo
2016

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Gostaria de agradecer a professora Alessandra Montini pelo apoio e orientação durante os estudos sobre um tema que tenho muito interesse em continuar aprendendo. Também estou muito grato pelos conselhos dos meus amigos Alan Dieguez e Paulo Haddad que deram dicas sobre suas experiências.

Também gostaria de agradecer a minha família que sem dúvidas me deram o suporte para que pudesse estudar e em especial, a minha noiva Eliana que sempre esteve ao meu lado.

"One's mind, once stretched by a new idea,
never regains its original dimensions.
(HOMES, Oliver Wendell)

Resumo

Segundo a ABNT (2003, 3.1-3.2), o resumo deve ressaltar o objetivo, o método, os resultados e as conclusões do documento. A ordem e a extensão destes itens dependem do tipo de resumo (informativo ou indicativo) e do tratamento que cada item recebe no documento original. O resumo deve ser precedido da referência do documento, com exceção do resumo inserido no próprio documento. (. . .) As palavras-chave devem figurar logo abaixo do resumo, antecedidas da expressão Palavras-chave:, separadas entre si por ponto e finalizadas também por ponto.

Palavras-chave: latex. abntex. editoração de texto.

Abstract

This is the english abstract.

Keywords: latex. abntex. text editoration.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
abnTeX	ABsurdas Normas para TeX

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

Sumário

Introdução

A tecnologia evoluiu a ponto de tornar possível o armazenamento de volume de dados um grande desafio. Na internet, é possível encontrar mais de 60 trilhões de páginas indexadas pelo Google (SMITH, 2016). Só o Facebook possui warehouse com mais de 300 petabytes, tendo um tráfego de mais de 600 terabytes diários (VAGATA; WILFONG, 2014).

Por trás desses dados brutos armazenados de forma estruturada ou não, existe o que Baeza-Yates e Ribeiro-Neto (1999) chama de informação. Em geral, os dados são objetos brutos que trazem pouco ou nenhum significado. A informação, então, refere-se a uma interpretação do dado dentro de um contexto com um ganho cognitivo. A utilização dessa informação para qualquer fim produz o conhecimento.

Por conta da dificuldade computacional em não só armazenar como analisar e monitorar esse volume de dados que nasceu o Big Data. A análise e extração de informações possibilitam uma melhor compreensão de vários aspectos, micro e macro na empresa. A utilização de Big Data a favor da companhia pode conferir uma grande vantagem competitiva, bem como uma diferenciação, sendo considerado um ativo estratégico muito valioso.

Este trabalho visa estudar conceitos teóricos estatísticos que analisam os dados, os algoritmos que criam as informações, bem como tecnologias que auxiliam o processo como um todo.

Parte I

Preparação da pesquisa

Este trabalho se baseia em um experimento quantitativo sobre problemas relacionados a segmentação e classificação de dados de uma base de dados grande, utilizando algoritmos de K médias, regressão logística e random forest. Serão feitas interpretações, análises e comparações, levantando aspectos positivos e negativos de cada metodologia.


Generated with online service GanttPro.com
2016
May
June
July
22 Week
23 Week
24 Week
25 Week
26 Week
27 Week
28 Week
29 Week
30 Week
Task name
Start date
Duration hour
Progr...
Assigned
Total Estimate 06/06/16 20:32 982.00
Super-event 06/06/16 20:32 982.00 4%
Develop workplan 06/06/16 23:32 202.00 32%
Develop agenda 10/06/16 02:32 127.00 56%
Invite speakers 10/06/16 02:32 72.00 40% Organizator
Prepare material 11/06/16 09:32 96.00 47% Organizator
Develop event form: 06/06/16 23:32 96.00 0% Empty
Design contributions 13/06/16 20:32 576.00 0%
Graphics consultati 13/06/16 20:32 144.00 0% Designer
Develop logo 19/06/16 20:32 144.00 0% Designer
Design brochure 19/06/16 20:32 120.00 0% Designer
Design invitations 19/06/16 20:32 96.00 0% Designer
Theme research 13/06/16 20:32 576.00 0%
Event environmei 13/06/16 20:32 264.00 0% Organizator
Media productior 30/06/16 20:32 168.00 0% Marketing
Develop budget 06/06/16 20:32 288.00 0% Organizator
Contact media 07/07/16 20:32 144.00 0% PR
Hire and train personn 03/07/16 18:32 144.00 0% HR
Refine and retest the 09/07/16 18:32 120.00 0% Organizator
Event reminder 14/07/16 18:32 72.00 0% PR
Event 17/07/16 18:32
Super-event
Develop workp
Develop workplan
Develop a
Develop agenda
Invite
Invite speakers
Prepar
Prepare materials
Develo
Develop event format
Design contributions
Graphics c
Graphics consultation
Develop lo
Develop logo
Design t
Design brochure
Design
Design invitations
Theme research
Event environment
Event environment
Media produ
Media production
Develop budget
Contact m
Contact media
Hire and tr
Hire and train personnel
Refine a
Refine and retest the event
Even
Event reminder
Event

Parte II

Referenciais teóricos

1 Big Data

O Big Data não representa apenas uma simples combinação de tecnologias.

Big Data possui três características intrínsicas: volume, velocidade e variedade.

1.1 K Médias

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

1.2 Regressão logística

A regressão logística é um modelo matemático de predição de eventos para um conjunto de variáveis independentes de entradas, baseando-se nas probabilidades de ocorrência desses eventos. ***As ocorrências desses eventos, por sua vez, são variáveis binárias dependentes.*** Dessa forma, a regressão logística viabiliza a classificação das observações por meio da probabilidade estimada na categoria estudada.

A variável dependente na regressão logística segue uma distribuição Bernoulli com uma probabilidade p desconhecida

Assim, a regressão logística tem como propósito estimar a probabilidade p desconhecida para uma certo conjunto de variáveis independentes.

Para estimar essa probabilidades, a regressão logística se utiliza da função logit.

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{1.1}$$

A função logit se baseia no conceito de chances. As chances é dada pela equação:

$$OR = \frac{p}{1-p}(1) \tag{1.2}$$

onde p é a probabilidade de sucesso para a ocorrência do evento x, 1-p é probabilidade de fracasso e x é um evento que segue a distribuição Bernoulli

Portanto, OR representa a razão entre as probabilidades de sucesso e fracasso de um determinado evento.

Utilizando inferência de estatística, podemos aplicar log em (1), ficando com:

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{1.3}$$

$$logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} \quad = \frac{e^{\alpha}}{1+e^{\alpha}} \tag{1.4}$$

O modelo de regressão logística tem a forma:

$$\begin{aligned}
\log\left(\frac{P(G=1|X=x)}{P(G=K|X=x)}\right) &= \beta_{10} + \beta_1^T x \\
\log\left(\frac{P(G=2|X=x)}{P(G=K|X=x)}\right) &= \beta_{20} + \beta_2^T x \\
\log\left(\frac{P(G=K-1|X=x)}{P(G=K|X=x)}\right) &= \beta_{(k-1)0} + \beta_{k-1}^T x
\end{aligned} \tag{1.5}$$

Onde o modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2011) é composto por K classes, e K - 1 transformações logit. * A transformação logit se faz necessária para que a restrição da soma de todas as probabilidades seja igual a 1. Utilizando a inversa da logit, temos:

$$\begin{aligned}
P(G=k|X=x) &= \frac{\exp\left(\beta_{k0} + \beta_k^T x\right)}{1 + \sum_{\ell=1}^{K-1} \exp\left(\beta_{\ell 0} + \beta_\ell^T x\right)}, k = 1, ..., K-1 \\
P(G=K|X=x) &= \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp\left(\beta_{\ell 0} + \beta_\ell^T x\right)}
\end{aligned} \tag{1.6}$$

1.3 Random Forest

Árvores são modelos presentes tanto em computação como estrutura de dados e em estatística como estrutura para tomadas de decisão. No contexto de aprendizado de máquina, a árvore de decisão refere-se a uma estrutura de modelo preditivo, um método de aprendizagem supervisionada não parametrizada utilizada para classificação e regressão (CART). Para Hastie, Tibshirani e Friedman (2011), as árvores permitem um particionamento do espaço em um conjunto de retângulos com um modelo simple em cada.

//por figura Legenda Neste exemplo (HASTIE; TIBSHIRANI; FRIEDMAN, 2011), observa-se que o particionamento num espaço bidimensional, gerando 5 regiões. Ao lado, está a árvore que gerou esse particionamento.

Trata-se de uma representação de regras que dividem as observações em grupos com características em comum. Em geral, considera-se a hipótese de que cada característica possui um domínio finito e discreto.

Árvores de classificação

Seja p os dados de entrada

Podemos representar de forma genérica o modelo da árvore de classificação por

Suponha que exista M partição que possa ser divida em regiões $R_1, R_2, ..., R_M$ e que seja possível modelar a resposta para cada região com a constante c_m,

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{1.7}$$

Utilizando como critério de minimização a soma do mínimos quadrados $\sum (y_i - f(x_i))^2$, temos que o melhor $\hat{c}_m$ é exatamente a média para y_i na região R_m:

$$\hat{c}_m = média(y_i | x_i \in R_m) \tag{1.8}$$

Para encontrar a melhor partição, é necessário recorrer a um algoritmo guloso.

Seja j uma variável de reparticionamento, s um ponto de divisão. É possível definir 1 par de semi planos

$$R_1(j,s) = X|X_j \leq s \quad \text{e} \quad R_2(j,s) = X|X_j > s) \tag{1.9}$$

Resultando a busca pela de s e j que resolva

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_1} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right] \tag{1.10}$$

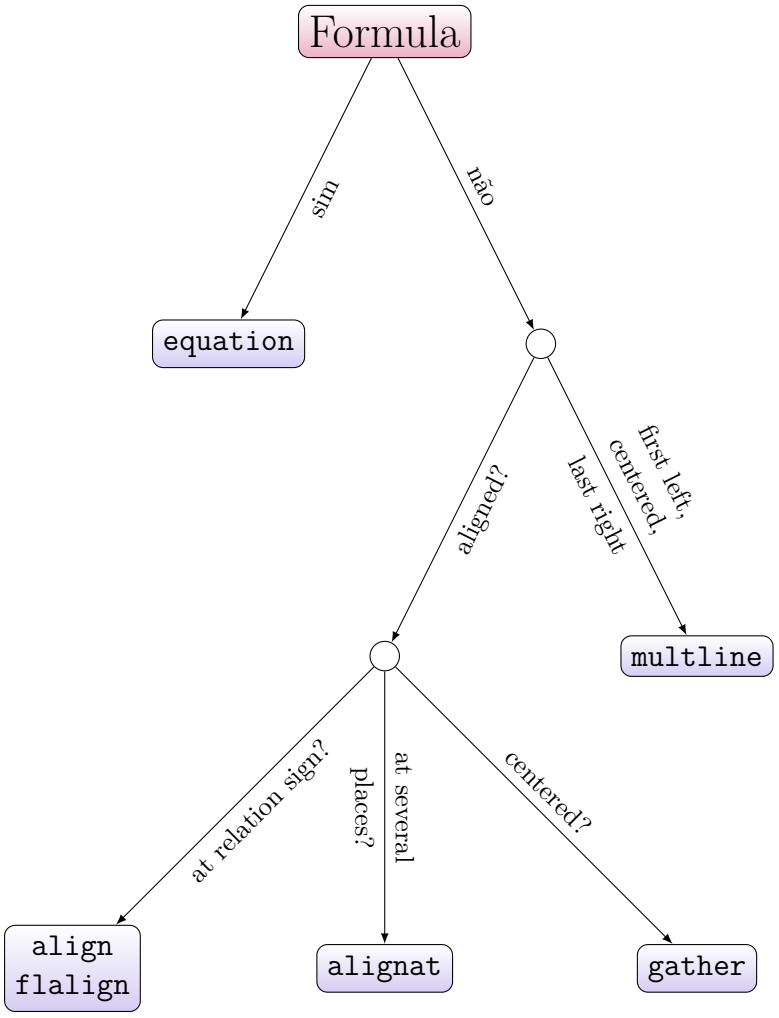

Figura 1 – Cada nó representa um atributo de um elemento da amostra. As folhas são consideradas a representação da classe a que uma observação pertence. Já o ramo é um conjunto de valores que reflete todas suas características e detalhes de um elemento

Mas como visto em 1.8, temos que para qualquer j e s, a minimização interna pode ser resolvida por:

$$\hat{c}_1 = \textit{média}(y_i | x_i \in R_1(j,s)) \quad \text{e} \quad \hat{c}_2 = \textit{média}(y_i | x_i \in R_2(j,s)) \tag{1.11}$$

Este processo é repetido até que todas as regiões sejam descobertas.

Parte III

Resultados

Referências

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 6028*: Resumo - apresentação. Rio de Janeiro, 2003. 2 p. Citado na página 11.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. 1. ed. [S.l.]: Addison Wesley, 1999. Citado na página 25.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. [S.l.]: Springer, 2011. ISBN 0387848576. Citado 2 vezes nas páginas 34 e 35.

SMITH, C. *By the numbers: a gigantic list of google stats and facts*. 2016. Disponível em: <http://expandedramblings.com/index.php/by-the-numbers-a-gigantic-list-of-google-stats-and-facts/>. Acesso em: 07 jun 2016. Citado na página 25.

VAGATA, P.; WILFONG, K. *Scaling the Facebook data warehouse to 300 PB*. 2014. Disponível em: <https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/>. Acesso em: 07 jun 2016. Citado na página 25.

Apêndices

APÊNDICE A – Spark

Sed mattis, erat sit amet gravida malesuada, elit augue egestas diam, tempus scelerisque nunc nisl vitae libero. Sed consequat feugiat massa. Nunc porta, eros in eleifend varius, erat leo rutrum dui, non convallis lectus orci ut nibh. Sed lorem massa, nonummy quis, egestas id, condimentum at, nisl. Maecenas at nibh. Aliquam et augue at nunc pellentesque ullamcorper. Duis nisl nibh, laoreet suscipit, convallis ut, rutrum id, enim. Phasellus odio. Nulla nulla elit, molestie non, scelerisque at, vestibulum eu, nulla. Ut odio nisl, facilisis id, mollis et, scelerisque nec, enim. Aenean sem leo, pellentesque sit amet, scelerisque sit amet, vehicula pellentesque, sapien.

APÊNDICE B – Pandas

Sed consequat tellus et tortor. Ut tempor laoreet quam. Nullam id wisi a libero tristique semper. Nullam nisl massa, rutrum ut, egestas semper, mollis id, leo. Nulla ac massa eu risus blandit mattis. Mauris ut nunc. In hac habitasse platea dictumst. Aliquam eget tortor. Quisque dapibus pede in erat. Nunc enim. In dui nulla, commodo at, consectetuer nec, malesuada nec, elit. Aliquam ornare tellus eu urna. Sed nec metus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.