Ciência de dados FGV – memórias

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fnc-www5.fgv.br%2Fcursosgratuitos%2Fcg%2FOCWICDEAD%2Fbase\_rede%2Fpag%2Fassets%2Fintroducao\_ciencia\_de\_dados\_reading\_list.pdf&clen=1138872&chunk=true

Introdução:

O processo de descoberta de conhecimento a partir de dados compreende um conjunto de etapas que envolvem desde a organização e limpeza da informação até a validação e o uso da informação obtida.

Apresentar as técnicas matemáticas e computacionais de forma simples e intuitiva, permitindo ao estudante compreender como elas funcionam, quando e em que tipo de problemas devem ser empregadas e quais são as suas limitações. Apresentar exemplos de aplicações reais do processo de extração e geração de conhecimento via Ciência de Dados. Discutir os aspectos éticos envolvidos em Ciência de Dados.

Um dos mais importantes é Big Data, e como esse fenômeno começou a mudar a maneira como os problemas são resolvidos hoje.

Em resumo, o diagrama descrito no texto nos mostra que a Ciência de Dados é composta de três áreas importantes: matemática/estatística, computação e conhecimento do domínio. (Principles of data science (OZDEMIR, 2016) Capítulo 1, Seções What is data science? e The data science Venn diagram)

The hardest thing in data science (Disponível em: <https://buckwoody.wordpress.com/2015/12/30/the-hardest-thing-indata-science>) Fazer boas perguntas é tanto uma ciência quanto uma arte. Este post descreve esta como a tarefa mais difícil na Ciência de Dados. É por isso que entender os objetivos da empresa – ou do cliente – e as limitações dos seus dados parece ser pré-requisito fundamental para fazer perguntas interessantes.

No contexto de Ciência de Dados, os processos de extração de conhecimento e tomada de decisão são tipicamente executados seguindo um conjunto de etapas, as quais podem ser descritas como: preparação dos dados, exploração dos dados, escolha de técnicas e modelos, ajuste de modelos e, finalmente, avaliação e uso dos modelos.

Preparação dos dados

Nesta unidade, discutiremos questões relacionadas à qualidade e integridade de dados. Em particular, apresentaremos os conceitos de transformação e engenharia de atributos, além de abordarmos problemas que tipicamente aparecem em bases de dados, por exemplo, dados faltantes, inconsistentes e redundantes.

Data science, The MIT Press (KELLEHER; TIERNEY, 2018) Capítulo 3 Discute o problema de limpeza e tratamento de dados desde as etapas de aquisição dos dados, analisando possíveis fontes de erros. A questão do armazenamento dos dados também é abordada.

Vamos discutir duas ferramentas básicas bastante empregadas nessa tarefa: estatística descritiva e visualização.

Ao final desta unidade, o estudante terá conhecimento da diferença entre métodos descritivos e preditivos e quais classes de problemas esses métodos buscam resolver.

Moving into data science as a career: mathematical models (Disponível em: <https://towardsdatascience.com/moving-into-data-science-as-a-careermathematical-models-e13f30690b00>) Discute o problema de modelagem matemática em ciência de dados a partir da perspectiva das habilidades matemáticas e computacionais dos profissionais que atuam na área.

Avaliando métodos e modelos Nesta unidade, discutiremos a questão da avaliação de métodos descritivos e preditivos. Em particular, estudaremos os diferentes tipos de métricas que podem ser utilizados e as suas particularidades com relação ao tipo de dado e método. Discutiremos ainda, no caso de modelos preditivos, como organizar os dados a fim de não incorrermos em avaliações errôneas ou equivocadas.

A general introduction to data analytics (MOREIRA et al., 2018) Seção 5.2 Fornece uma visão geral do problema de avaliar a qualidade dos métodos e dos modelos empregados no processo de extração e de geração de conhecimento.

MÓDULO IV – MÉTODOS MATEMÁTICOS E COMPUTACIONAIS

como tratar dados faltantes, como detectar informações discrepantes – os chamados outliers – e como realizar transformações de escala, muito utilizadas em problemas reais.

A general introduction to data analytics (MOREIRA et al., 2018) Seções 4.1 a 4.4 Descreve os principais problemas presentes em bases de dados reais, fornecendo uma discussão simplificada de como tratar tais problemas.

Análise de componentes principais Nesta unidade, estudaremos uma das mais importantes técnicas de análise e transformação de dados, a chamada Análise de Componentes Principais – do inglês Principal Component Analysis (PCA). A técnica PCA será apresentada de modo que o estudante possa compreender como ela funciona, qual o tipo de análise que ela proporciona e quais são as suas limitações.

A tutorial on principal component analysis (SHLENS, 2014) Este artigo fornece uma visão intuitiva do método PCA, apresentando em seguida um detalhamento matemático de como o método funciona. Estudantes menos familiarizados com noções de álgebra linear podem concentrar-se nas seções I, II e III.A.

Técnicas de agrupamento

Técnicas de agrupamento são métodos descritivos que buscam encontrar conjuntos de objetos semelhantes, sendo extremamente úteis para revelar padrões que caracterizam conjuntos específicos de objetos. Nesta unidade, estudaremos duas das principais técnicas de agrupamento, o método Kmeans e agrupamento hierárquico, analisando as suas diferenças, vantagens e limitações. A general introduction to data analytics (MOREIRA et al., 2018) Capítulo 5, Seções 51 e 5.3 Traz uma discussão mais detalhada sobre a questão da escolha da métrica para diferentes tipos de dados e o impacto nos resultados de agrupamentos, apresentando exemplos.

Modelos de classificação

Modelos preditivos de classificação buscam classificar objetos a partir de informações contidas nos seus atributos. Nesta unidade, estudaremos um conjunto de técnicas de classificação largamente empregadas em problemas de Ciência de Dados. Aprenderemos como tais modelos funcionam, quais são as suas limitações e como avaliar a sua eficácia. Existe um grande número de técnicas de classificação, e várias delas são largamente empregadas em problemas reais. A fim de apresentar as técnicas com um bom nível de detalhe, dividimos esta unidade em duas partes. A primeira parte inicia discutindo uma técnica bastante simples, vizinhos mais próximos, abordando em seguida os métodos de regressão logística e Naive Bayes, muito empregados em problemas reais. A segunda parte desta unidade aborda as técnicas SVM e Árvores de Decisão, que são consideradas mais sofisticadas. O conceito de comitês, bootstrap e boosting também são abordados na segunda parte desta unidade.

Support Vector Machine (SVM) Tutorial (Disponível em: <https://blog.statsbot.co/support-vector-machines-tutorial-c1618e635e93>) Discute a técnica SVM com um linguajar não matemático, apresentado diversos exemplos e ilustrações do conceito de Kernel e como ele funciona.

Os estudos devem também ter tornado clara a importância da técnica PCA como ferramenta de análise e transformação de dados.

MÓDULO VI – QUESTÕES ÉTICAS EM CIÊNCIA DE DADOS

Neste último módulo, vamos ver alguns conceitos muito importantes. Nesta era do Big Data, temos o poder de acessar muitos dados, no entanto, isso também traz uma grande responsabilidade de saber como agir eticamente com o único objetivo de tornar o mundo melhor. No entanto, existem alguns pontos que podem parecer óbvios, mas é bom conhecê-los. As próximas três unidades lidam mais de perto com cada uma delas.

Privacidade e segurança

A Ciência de Dados pode ajudar a tornar o mundo muito melhor, mas também pode prejudicar a sociedade ou os indivíduos. Nas leituras abaixo, descreveremos alguns aspectos éticos e de privacidade que todo cientista de dados deve conhecer. Por exemplo, devemos ter cuidado na forma como comunicamos os nossos resultados, manter a segurança de dados confidenciais e cuidar da privacidade em dados agregados, entre outros.

BIBLIOGRAFIA AGGARWAL, C. C. Outlier Analysis. Springer, 2017. CADY, Field. The data science handbook. Hoboken, NJ: John Wiley & Sons, 2017. p. COHEN, Gerald A. Deeper into bullshit. Contours of agency: essays on themes from Harry Frankfurt, 321-339, 2002. Disponível em: uploads/8/1/2/6/8126749/cohen\_and\_frankfurt\_2002.pdf>. <http://phil480.weebly.com/ FRANKFURT, Harry. On bullshit, 1986. Disponível em: <http://www2.csudh.edu/ccauthen/ 576f12/frankfurt\_\_harry\_-\_on\_bullshit.pdf>. KELLEHER, J. D.; TIERNEY, B. Data science, The MIT Press, 2018. MOREIRA, J. M.; CARVALHO, A.; HORVÁTH, T. A general introduction to data analytics. Hoboken: Wiley, 2018. NG, A.; SOO, K. Numsense! Data science for the layman: no math added. Annalyn Ng & Kenneth Soo, 2017. O’NEIL, Cathy; SCHUTT, Rachel. Doing data science: straight talk from the frontline. Sebastopol, California: O’Reilly Media, Inc., 2013. OZDEMIR, Sinan. Principles of data science. Packt Publishing Ltd., 2016. PROVOST, Foster; FAWCETT, Tom. Data science for business: what you need to know about data mining and data-analytic thinking. O’Reilly Media, Inc., 2013. SALTZ, J. S.; STANTON, J. An introduction to data science. SAGE Publications, 2018. SELTMAN, H. J. Experimental design and analysis, 2018. Disponível em: <http:// www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>. SHLENS, J. A tutorial on principal component analysis, 2014. Disponível em: <https:// arxiv.org/abs/1404.1100>. SKIENA, S. The data science design manual, Springer, 2017.  36 SMITH, L. A tutorial on principal components analysis. Technical Report OUCS-2002-12, 2002. Disponível em: <http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca\_tutorial.pdf>. Report SOLEY-BORI, M. Dealing with missing data: key assumptions and methods for applied analysis, Technical n. 4, Boston University, www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf>. 2005. 2013. Disponível em: <https:// TAN, P.-N.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. Introduction to data mining, Pearson, Disponível ch7\_clustering.pdf>. em: <https://www-users.cs.umn.edu/~kumar001/dmbook/ WITTEN, Ian H. et al. Data mining: practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann, 2016.

Fim do resumo do resumo kkkk