UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ECONOMIA

HO-238: INTRODUÇÃO À CIÊNCIA DE DADOS PARA CIÊNCIAS SOCIAIS E HUMANAS (ICD)

1S-2023

Ivette Luna
iluna@unicamp.br
https://www.linkedin.com/in/ivetteluna/
http://lattes.cnpq.br/2854855744345507

Descrição

É inegável que a disponibilidade de dados em todas as áreas de conhecimento é cada vez maior quanto ao seu volume, a sua velocidade de divulgação e a sua diversidade, no seu tipo e nas fontes. Nesse cenário, pesquisas multidisciplinares e abordagens *data driven* vêm a contribuir nas análises em diversas áreas de conhecimento. Urge por tanto, conhecimento sobre analítica de dados, visando a extração de informação e o desenvolvimento de modelos adequados como instrumentos de suporte para os tomadores de decisão.

Objetivos

A disciplina apresenta os fundamentos da Analítica e Ciência de dados, visando apresentar a base para o estudo de técnicas mais complexas no campo do Aprendizado de Máquina.

Pré-requisitos

É desejável o conhecimento da estatística básica: medidas descritivas, distribuições e teste de hipótese. Os conceitos serão também apresentados ao longo do desenvolvimento das aulas. Quanto a Álgebra Linear, os conceitos mínimos necessários serão comentados durante as aulas. As aulas introdutórias de R e de estatística descritiva serão ministradas nos primeiros encontros, porém, já em um formato hands-on, pelo que é importante estudar antes (ver nota 2 ao final do programa).

Programa

Parte 1: As transformações tecnológicas, a Ciência de Dados e o trabalho dos cientistas

Introdução. Definindo o que é essa ciência, o que os cientistas de dados fazem e quais ferramentas e algoritmos eles usam no dia a dia. Apresentação da disciplina e o contexto geral da área.

Parte 2: Introdução ao R

Aulas introdutórias de R, uma das linguagens mais usadas em Ciência de Dados visando o conhecimento básico para o acompanhamento da disciplina. Contudo, as atividades podem ser desenvolvidas também em Python, caso o aluno tenha experiência com essa linguagem e assim o prefira.

Parte 3: Análise exploratória de dados

Revisaremos os principais conceitos e ferramentas estatísticas para uma análise exploratória de dados e na validação de modelos das próximas seções.

- Dados: tipos, tipos de variáveis (qualitativas, quantitativas)
- Análise gráfica
- População e amostra, viés
- Medidas de posição e de dispersão
- Medidas de associação: correlação e covariância
- Distribuições de dados e o TLC
- Teste de hipótese (introdução), teste z, t, teste qui-quadrado

Parte 4: Algoritmos não supervisionados

Quantas vezes tivemos que identificar padrões ou grupos? Ou identificar variáveis (mesmo que latentes) relevantes para explicar os padrões observados? Entraremos na lógica dos algoritmos de agrupamento mais tradicionais na Ciência de Dados, identificando as condições necessárias para o uso de cada técnica.

- Cluster hierárquico – dendograma, algoritmo aglomerativo, medidas de distância (Single Linkage, Complete Linkage, Avarage Linkage, Centróide, Ward)
- Cluster não hierárquico - K-Means
- Análise de componentes principais - PCA
- Análise de correspondência – Anacor, ACM e AC para dados mistos (FAMD)

Parte 5: Algoritmos supervisionados

Esmiuçaremos a teoria por trás dos algoritmos supervisionados mais tradicionais de agrupamento e classificação na Ciência de.

- K-vizinhos mais próximos - KNN (RL não se inclui por se ver em Econometria)
- Modelos de árvore de decisão (CART)
- Avaliação de modelos de classificação: matriz de confusão, AUC, curvas ROC, Precisão, revocação e especificidade, F1 etc
- Bagging e floresta aleatória (*random forests*)
- Boosting

Parte 6: Introdução a Redes neurais artificiais

Uma técnica bastante tradicional e que recentemente se popularizou pelo *deep learning*, atualmente possível pelo poder computacional disponível.

- História e conceitos: componentes e funções de ativação
- Processo de aprendizado: otimização e treinamento supervisionado
- Rede perceptron multicamadas (MLP)
- Support Vector Machine (SVM)

Referências

W. de O. Bussab e P. A. Morettin (2017). Estatística básica, 9° ed., Ed. Saraiva

Ch. Wheelan (2013). Estatística: o que é, para que serve, como funciona, 5ª ed., Ed. Zahar

L. P. Fávero e P. Belfiore (2017). Manual de análise de dados. Ed. Elsevier.

P. Bruce e A. Bruce (2019). Estatística prática para cientistas de dados, Ed. O'Reilly

K. Faceli et al (2021). Inteligência Artificial: uma abordagem de aprendizado de máquina, 2ª ed.; ed. LTC

Referências open source

R4DS, Hadley Wickham (2017): https://r4ds.had.co.nz/

R for grad students, Y. Wendy Huynh (2019): https://bookdown.org/yih_huynh/Guide-to-R-Book/

Análise estatística de relações lineares e não lineares. Edição do autor. Portal de livros abertos da USP. R. Hoffmann (2016): https://doi.org/10.11606/9788592105716

Data Science for Business, F. Provost e T. Fawcett (2013): https://www.researchgate.net/publication/256438799_Data_Science_for_Business

Deep learning book (para redes neurais): https://www.deeplearningbook.com.br/

Nota 1: ao longo do semestre se indicará bibliografia complementar referente às aplicações de cada seção.

Links legais (scripts, dados, fóruns e canais)

The Algorithms (exemplos de scripts em diversas linguagens, sobre modelos de ML):

https://the-algorithms.com/category/machinelearning

Quick-R (pequenos macetes para resolver quase tudo no R):

https://www.statmethods.net/

Stack Overflow (fórum): https://stackoverflow.com/

R-Bloggers (fórum): https://www.r-bloggers.com/

StatQuest with Josh Starmer (canal incrível)

https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw

ASN. Rocks (canal da jedi de DS):

https://www.youtube.com/channel/UCKkLm58oeFM77_MwfO06Mwg

Repositórios para cientistas de dados: https://www.cienciaedados.com/15-repositorios-no-github-para-cientistas-de-dados/

Kaggle: https://www.kaggle.com/

UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php

Base dos dados: https://basedosdados.org/

Leituras introdutórias/motivadoras

Data Scientist: The Sexiest Job of the 21st Century

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Cientistas de Dados: quem são, o que fazem e por que você quer ser um?
https://www.sas.com/pt_br/insights/analytics/cientistas-de-dados.html

Life of Data | Data Science is OSEMN
https://medium.com/@randylaosat/life-of-data-data-science-is-osemn-f453e1febc10

Nota 2:

Para se preparar com a parte estatística, usar o livro do Bussab e Morettin (ou algum outro de estatística básica):

- Ver o capítulo 3 sobre medidas descritivas.
- Ver os capítulos 10 e 12 sobre distribuição amostral e teste de hipótese.
- Ver as vídeo aulas do meu canal:
  https://www.youtube.com/channel/UCeH0Kby1SIomvTPnOGHyoXA/playlists

Minicurso de R (vídeo e slides)

- https://ilunah.wixsite.com/ivetteluna/minicurso-r