UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ECONOMIA

[DRAFT HO-350: Tópicos Especiais em Economia]

INTRODUÇÃO À CIÊNCIA DE DADOS PARA CIÊNCIAS SOCIAIS E HUMANAS

1S-2022

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 empíricas 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 Ciência de dados, visando apresentar a base para o estudo de técnicas mais complexas no campo do Aprendizado de Máquina.

Pré-requisitos

Nenhum. Aulas introdutórias sobre Álgebra Linear serão disponibilizadas para aqueles que precisem dessa base. As aulas introdutórias de Estatística e R serão ministradas nos primeiros encontros.

Programa

Parte 1: 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: Estatística descritiva

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, TLC
- Medidas de dispersão
- Medidas de associação: correlação e covariância
- Distribuições de dados: uniforme, Bernoulli, Binomial, Poisson, Normal, Exponencial, t-Student, Qui-quadrado, F
- Teste de hipótese (introdução)
- Teste z, t, F (ANOVA), 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 e ACM
- Agrupamento de dados mistos

Parte 5: Algoritmos supervisionados

Esmiuçaremos a teoria por trás dos algoritmos supervisionados mais tradicionais DE agrupamento e classificação na Ciência de Dados e os aplicaremos em bases de dados diversas.

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

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

Referências

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

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

F. Provost e T. Fawcett (2016). Data Science para negócios. Ed. Alta Books.

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

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

R. Hoffmann (2016) Análise estatística de relações lineares e não lineares. Edição do autor. Portal de livros abertos da USP.

https://doi.org/10.11606/9788592105716

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