UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ECONOMIA

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

1S-2024

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 atuação. Urge por tanto, conhecimento sobre analítica de dados, visando a extração de informação e conhecimento que auxiliem na compreensão de problemas na Economia e nas Ciências Sociais, assim como o desenvolvimento de modelos adequados como instrumentos de suporte para os tomadores de decisão.

Objetivos

A disciplina tem por objetivo apresentar os fundamentos da Analítica e Ciência de dados, assim como a apresentação e aplicação 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 programação e de estatística descritiva serão ministradas nos primeiros encontros, porém, já em um formato hands-on, pelo que é importante estudar o elementar de estatística descritiva*.

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)
- Medidas de posição e dispersão
- Revisão de distribuições de dados e o TLC
- Medidas de associação: correlação e covariância
- Revisão de teste de hipótese e p-valor, algumas estatísticas e transformação de dados

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 e K-Medoids
- Revisão de autovalores e autovetores para Análise fatorial
- Análise de componentes principais - PCA
- Análise de correspondência – Anacor, ACM e AF 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
- Métricas de desempenho de modelos de classificação e de regressão
- Entropia e Gini como métricas de incerteza
- Modelos de árvore de decisão (CART)
- Validação cruzada, grid-search e tuning
- Comitê de máquinas: bagging e boosting
- Floresta aleatória (*random forests*), XGBoost

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.

- Conceitos, componentes e funções de ativação
- Revisão do gradiente e otimização não linear de primeira ordem
- Rede perceptron multicamadas (MLP)
- Processo de aprendizado: treinamento supervisionado (gradiente descendente)

Sugestão de livros para estudo dos temas

(Não é pra ler todas as opções, basta escolher uma por tema, a que seja do seu agrado)

Contexto

Inteligência Artificial: Avanços e tendências. organizadores Fabio G. Cozman, Guilherme Ary Plonski, Hugo Neri. São Paulo: Instituto de Estudos Avançados, 2021. (Livro da USP).

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ECONOMIA

Revisão de distribuições de dados e o TLC, p-valor

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

Ver o capítulo 3 sobre medidas descritivas.

Ver os capítulos 10 e 12 sobre distribuição amostral e teste de hipótese.

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

Os primeiros capítulos do Statquest também são muito bons para uma revisão leve dos conceitos básicos de estatística para ML (livro no Classroom).

Técnicas não supervisionadas - conceitos, Clustering e medidas de similaridade, Clustering hierárquico, K-Means

Practical Machine Learning in R (2020). Edição Inglês por Fred Nwanganga, Mike Chapple.

Practical Guide to Cluster Analysis in R: Multivariate Analysis I (2017). Alboukadel KASSAMBARA. STHDA.

Análise fatorial – PCA, Anacor, ACM e FAMD

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

Correspondence Analysis in practice (2017). Greenacre, Michael. Chapman & Hall/CRC, 3rd. Edition.

Practical Guide to Principal Component Methods in R: Multivariate Analysis II (2017). Alboukadel KASSAMBARA. STHDA.

K-vizinhos mais próximos - KNN

Cap. 4 do livro da Katti Facelli

Cap. 12 do livro de Joel Grus (Data Science do Zero)

Sobre medidas de erro, ver:

Cap. 9 do livro da Katti Facelli

Cap. 8 do livro do StatQuest

Cap. 6 do livro de Fred Nwanganga (Practical Machine Learning i R)

Sobre validação cruzada e bootstrap ver o cap. 9 do livro de Fred Nwanganga (Practical Machine Learning i R)

Modelos de árvore de decisão (CART), Bagging e florestas aleatórias (*random forests*), Boosting

Cap. 6 do livro da Katti Facelli

Cap. 10 do livro do StatQuest

Cap. 8 do livro do Fred Nwanganga (Practical Machine Learning i R)

Cap. 17 do livro de Joel Grus (Data Science do Zero), que aborda também as florestas aleatórias

Cap. 8 do livro do Gareth James (An Introduction to Statistical

Learning with Applications in R)

Redes neurais

Sobre o gradiente descendente, ver

Cap. 5 do livro do StatQuest

Cap. 8 do livro de Joel Grus (Data Science do Zero)

Cap. 12 do livro do StatQuest

Cap. 7 do livro da Katti Facelli

Cap. 18 do livro de Joel Grus (Data Science do Zero)

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/

Data Analytics with R, Adam Smith e Rafael Greminger (2022 updates): https://www.adamnsmith.com/MSIN0010/index.html

Ver outros em https://www.kaggle.com/general/274029

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

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/

Extra - Missings e outliers

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

Gábor Békés, Gábor Kézdi (2021). Data Analysis for Business, Economics, and Policy.

Análise Multivariada de Dados (2009). Edição Português por Joseph F. Hair Jr., William C. Black, Barry J. Babin, Rolph E. Anderson.

Nota*:

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

- Ver as vídeo aulas do meu canal: https://www.youtube.com/channel/UCeH0Kby1SIomvTPnOGHyoXA/playlists