Projeto ABCIA

Módulo 3

Aula 01 - INTRODUÇÃO A ML COM PYTHON

Prof. Msc. Acauan C. Ribeiro

UFRR | DCC DEPTO DE CIÊNCIA DA COMPUTAÇÃO | Softex | ABCIA ABC DA INTELIGÊNCIA ARTIFICIAL

Objetivo da Aula

- Construir **modelo** de Machine Learning a partir de **base de dados pública e gratuita** por meio da linguagem de programação Python, em um ambiente de desenvolvimento de fácil acesso. (Ex.: Google Colab, jupyter.org, Kaggle, etc...)

Setup - Várias Opções

The Jupyter Notebook

(Formerly known as the IPython Notebook)

The IPython Notebook is now known as the Jupyter Notebook. It is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media. For more details on the Jupyter Notebook, please see the Jupyter website.

https://jupyter.org/

Anaconda Repository

Our repository features over 8,000 open-source data science and machine learning packages, Anaconda-built and compiled for all major operating systems and architectures.

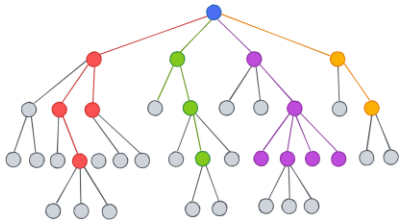
https://www.anaconda.com/

Google colab

https://colab.research.google.com/

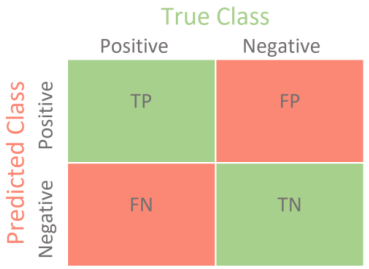
Conceitos que você vai precisar

- Decision Trees

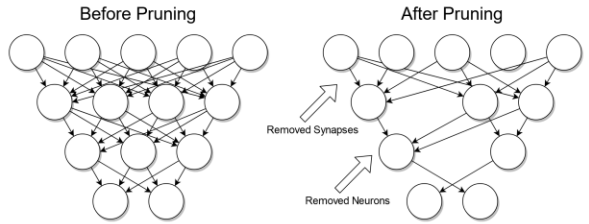


- Cross Validation

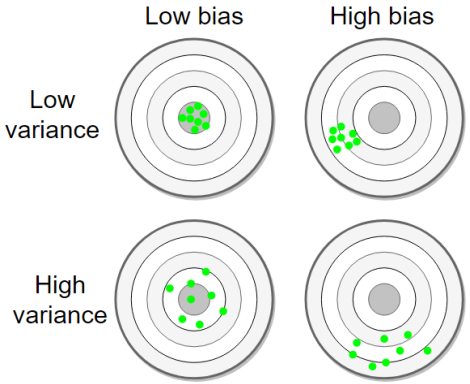
- Confusion Matrices



- Cost Complexity Pruning



- Bias and Variance and Overfitting



Material de Referência / Estudo

Pense em Python

https://penseallen.github.io/PensePython2e/

Python Data Science Handbook

https://jakevdp.github.io/PythonDataScienceHandbook/

Artigo para Ler

https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/35179.pdf

Machine Learning Crash Course - Google

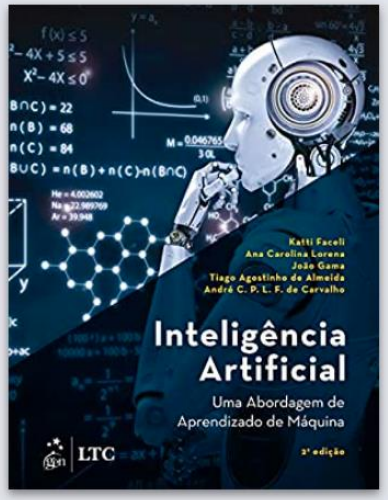
https://developers.google.com/machine-learning/crash-course?hl=pt-br

Curso de Machine Learning - Kaggle

https://www.kaggle.com/learn/intro-to-machine-learning

Livros Textos

Mãos A Obra: Aprendizado ...
Géron, Aurélien



Inteligência Artificial - Uma...
Carvalho, André Carlo...

Onde conseguir Datasets?

Repositórios populares de open data

- UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/)
- Conjunto de dados no Kaggle (https://www.kaggle.com/datasets)
- Conjunto de Dados na AWS (https://registry.opendata.aws/)

Metaportais de Dados (Listam repositórios open data)

- Data Portals (http://dataportals.org/)
- OpenDataMonitor (http://opendatamonitor.eu/)
- Quandl (http://quandl.com/)

Outras páginas que listam datasets

- Lista de conjuntos de dados de aprendizado de máquina do Wikipedia (https://homl.info/9)
- Quora.com (https://homl.info/10)
- Conjunto de dados do Reddit (https://www.reddit.com/r/datasets/)

Exemplo inicial

Link para o dataset: https://drive.google.com/file/d/1FwxvMv-FiQHkuAQUewXtgti-BLmaKRG9/view?usp=sharing

Passos (Pipeline):
-> Ler dados
-> Explorar Dados
-> Definir modelo
-> Treinar modelo (train dataset)
-> Testar modelo (test dataset)
-> Avaliar modelo a partir de uma métrica

Objetivo 1: Rodar nosso primeiro modelo.
Objetivo 2: Ver se conseguimos melhorar o modelo para ter resultados melhores.

Outro Exemplo - Dados de Saúde

- **Importação de Dados**

- **Dados ausentes (vazios)**
  - Identificando dados ausentes
  - Tratando dados ausentes

Link para o DataSet:
https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

- **Formatando os dados para Decision Trees**
  - Dividindo as variaveis em Dependent and Inpependent Variables
  - One-hot-Encoding

- **Construído uma primeira versão da Árvore de Classificação**

- **Melhorando a árvore com Cost Complexity Pruning**
  - Visualizando Alpha
  - Usando **Cross Validation** para encontrar o melhor valor para Alpha

- **Construção, Desenho, Interpretação e Evolução de uma Árvore de Classificação**

Roteiro para aprender ML

Algumas referências indicam começar aprender **Machine Learning** por algoritmos mais simples de compreender, seguindo até uma ordem como:

- Árvores de Decisão
- Regressão Linear
- Regressão Logística
- SVM
- KNN
- Agrupamento

Exercício

Exercício Prático

- O exercício dessa aula vai ser prático. Você vai aplicar uma solução em um desafio de **Machine Learning**.
- Você deve enviar o resultado do seu modelo (predições) em uma competição do **Kaggle**.
- **Link da competição:** https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques
- Tire um print da sua posição no Ranking (leaderboard) comprovando que você fez e enviou a atividade e ela foi aceita no Kaggle.

UFRR

DEPTO DE CIÊNCIA DA COMPUTAÇÃO

Softex

ABCIA
ABC DA INTELIGÊNCIA ARTIFICIAL

Até a próxima aula!

Continue praticando :)