

PÓS-GRADUAÇÃO

Linguagens de
programação para
ciência de dados
(*Python com Spark*)



PÓS-GRADUAÇÃO

Machine Learning em Python

Bloco 1

Marcelo Tavares de Lima





► Objetivos

- Apresentar os principais conceitos de aprendizado de máquina (*Machine Learning*).
- Descrever sobre os modelos supervisionados e não supervisionados.
- Apresentar exemplos práticos com o uso de linguagem de programação Python.



► Introdução

- Aprendizado de máquina (*Machine Learning*, em inglês) é uma área muito importante da Inteligência Artificial.
- Teve início por volta de 1950, onde alguns métodos estatísticos foram descobertos.
- *Machine Learning* é usado para resolver problemas em diversas áreas, como Engenharia, Medicina, Biologia e Computação.



► **Aprendizado de Máquina (*Machine Learning*)**

- Área da Inteligência Artificial (IA) que tem como objetivo fazer a extração de conhecimento a partir de um conjunto de dados.
- Requer conhecimento em matemática e computação.
- É o conceito de ensinar máquinas (computadores) a executar tarefas com ou sem supervisão, por meio de algoritmos de programação, segundo Müller (2016) e Marsland (2014).



► **Aprendizado de Máquina (*Machine Learning*)**

- Muitos aplicativos e sites que utilizamos contêm algoritmos que usam conceitos de *Machine Learning*: recomendações de quais filmes assistir, qual comida pedir ou quais produtos comprar, rádio on-line personalizado e sugestão de novos amigos em redes sociais.
- Empresas como Facebook, Amazon ou Netflix, utilizam aprendizado de máquinas em suas atividades, segundo Kwok e Bei (2013), Bell *et al.* (2010) e Castelli *et al.* (2017).



► **Aprendizado de Máquina (*Machine Learning*)**

- Estágios do aprendizado de máquina:
 - Seleção do conjunto de dados.
 - Análise de dados.
 - Desenvolvimento de algoritmo.
 - Análise dos resultados gerados pelo algoritmo.



► **Aprendizado de Máquina (*Machine Learning*)**

- Os tipos mais bem-sucedidos de algoritmos de aprendizado de máquina são aqueles que automatizam processos de tomada de decisão, generalizando a partir de modelos conhecidos.
- São conhecidos como aprendizado supervisionado, ou seja, o usuário fornece ao algoritmo entradas (rotuladas) e saídas desejadas, e o algoritmo encontra uma maneira de produzir a saída desejada sem ajuda humana.



► **Aprendizado de Máquina (*Machine Learning*)**

- No aprendizado de máquinas não supervisionado, apenas os dados de entrada são conhecidos e nenhum dado de saída conhecido é fornecido ao algoritmo.
- Nesse modelo, o algoritmo não tem informações/ atributos que possa aprender para os resultados de saída.

► Aprendizado de Máquina (*Machine Learning*)

Figura 1 – Visão geral dos modelos mais utilizados em *Machine Learning*



Fonte: Pereira (2019).

► Aprendizado de Máquina (*Machine Learning*)

Figura 2 – Visão geral dos modelos mais utilizados em *Machine Learning*

Machine Learning



Fonte: Pereira (2019).



► Entendendo o conjunto de dados

- É a etapa mais importante no processo de aprendizado de máquina.
- É necessário entender o que está acontecendo nos dados antes de começar a construção de um modelo (supervisionado ou não supervisionado).
- Cada algoritmo tem suas particularidades e você deve ser capaz de entender isso e escolher quais configurações do problema funcionam melhor para um dos modelos.



► Bibliotecas em Python para *Machine Learning*

- Fornecem aos cientistas de dados uma variedade de funções desenvolvidas, que facilitarão no desenvolvimento dos algoritmos.
- *Scikit-learn*: possui ferramentas para mineração de dados.
- *NumPy*: possui um conjunto de funções matemáticas de alto nível para tratamento de dados.



► Bibliotecas em Python para *Machine Learning*

- *SciPy*: coleção de funções para computação científica.
- *Matplotlib*: possui ferramentas para mineração de dados.
- *Pandas*: permite a elaboração de visualizações complexas e também realiza análises complexas.
- *IPython*: permite computação iterativa.




► **Aprendizado supervisionado: regressão**

- O principal objetivo da regressão é encontrar um modelo que mais se ajusta ao conjunto de dados fornecido.
- Em um problema de regressão, estamos tentando prever os resultados em uma saída contínua, o que significa que estamos tentando mapear variáveis de entrada para alguma função contínua.



► **Aprendizado supervisionado: classificação**

- O objetivo é prever um rótulo de classe, que é uma escolha dentre uma lista de possibilidades.
 - Em um problema de classificação, estamos tentando prever os resultados em uma saída discreta. Em outras palavras, estamos tentando mapear variáveis de entrada em categorias distintas.
- 

PÓS-GRADUAÇÃO

Machine Learning em Python

Bloco 2

Marcelo Tavares de Lima





► Regressão linear em *Python*

- Vamos desenvolver um algoritmo para tentar prever o preço de imóveis na cidade de Boston, nos Estados Unidos.
- Para esse exemplo, utilizaremos uma base de dados gratuita do pacote (biblioteca) *scikit-learn*, chamada Boston *house prices dataset*.
- Todas as informações desse conjunto de dados, estão disponível no site dos criadores do pacote (SCIKIT-LEARN, 2019).



► Regressão linear em *Python*

- Na Leitura Fundamental há uma série de passos que devem ser realizados inicialmente, antes de executar a programação em *Python*.

► Regressão linear em *Python*

- Código para calcular desvio padrão.

```
from sklearn.datasets import load_boston
from sklearn.linear_model import
LinearRegression
from sklearn.model_selection import
train_test_split
# carregando os dados
house_data = load_boston()
data = house_data['data']
target = house_data['target']
# separando o conjunto em 2 grupos:
treinamento e teste
data_train, data_test, target_train,
target_test = train_test_split(data, target,
test_size=0.33, random_state=42)
```

Há mais linhas não apresentadas AQUI.



► Regressão linear em *Python*

- Analisando os resultados, nosso algoritmo indica que houve 73% da variação nos preços dos imóveis na cidade de Boston. O conjunto de dados contém apenas 506 observações, talvez teríamos um resultado melhor se o *dataset* fosse maior e com mais atributos.



► Classificação em *Python*

- Como exemplo, utilizaremos um conjunto de dados gratuito do pacote *scikit-learn* que contém informações sobre tumores de câncer de mama. Utilizaremos um classificador para descobrir quais tumores são malignos e benignos.
- Todas as informações dessa base de dados estão disponível no site (SCIKIT-LEARN, 2019).



► Classificação em *Python*

- Na Leitura Fundamental há uma série de passos que devem ser realizados inicialmente, antes de executar a programação em *Python*.

► Classificação em *Python*

- Código em *Python*.

```
from sklearn.datasets import load_breast_cancer  
from sklearn.model_selection import train_test_split  
from sklearn.naive_bayes import GaussianNB  
from sklearn.metrics import accuracy_score  
# Carregando o dataset  
dataset = load_breast_cancer()  
# Organizar os dados  
labels = dataset ['target']  
features = dataset ['data']
```

Continua...

► Classificação em *Python*


- Código em *Python*.

Continuação...

```
# Separando os dados em 2 conjuntos: treinamento e teste  
treinamento, teste, treinamento_labels,  
teste_labels = train_test_split(features,  
labels, test_size=0.33, random_state=42)  
gnb = GaussianNB()  
model = gnb.fit(treinamento,  
treinamento_labels)  
preds = gnb.predict(teste)  
# Imprimindo o score do classificador  
print('Score do classificador: '  
+ str(round(accuracy_score(teste_labels,  
preds)*100,2)) + '%')
```



► **Aprendizado não supervisionado**

- São os tipos de aprendizado de máquinas onde não há saída conhecida.
 - O algoritmo de aprendizado mostra apenas os dados de entrada (não rotulados) e, por meio dele, o algoritmo deverá ser capaz de para extrair conhecimento somente desses dados, segundo Müller (2016).
- 



► **Aprendizado não supervisionado**

- O principal desafio no aprendizado não supervisionado é avaliar se o algoritmo aprende algo que de fato ajudará na resolução do problema.
- Algoritmos de aprendizado não supervisionado são aplicados a dados que não contêm informações sobre rótulos, por isso, não sabemos ou prevemos qual será a saída.
- Portanto, é muito difícil avaliar se o modelo teve um desempenho satisfatório ou não, dependerá muito do contexto do problema em questão, de acordo com Müller (2016).



► Aprendizado não supervisionado

- Tipos:
 - Agrupamento (*Clustering*).
 - *K-Means Clustering*.

PÓS-GRADUAÇÃO

Teoria em prática

Bloco 3

Marcelo Tavares de Lima





► Análise de dados com *Python*

- Você foi contratado por uma empresa produtora de soja no Brasil para resolver o seguinte problema:
 - Em um determinado período do ano, a empresa percebeu, por meio de imagens de satélite, que a plantação está sendo prejudicada por pragas.
 - Diante desse cenário, você deverá desenvolver um algoritmo capaz de identificar quais são as regiões mais afetadas pelas pragas e fazer uma classificação dos tipos de pragas encontradas, a fim de que a empresa possa tomar as providências para resolver esse problema e evitar prejuízos futuros.

Dica do professor

Bloco 4

Marcelo Tavares de Lima





► Dica de blog

INTELIGÊNCIA ARTIFICIAL SOB CONTROLE (IASC).

Disponível em:

<https://iascblog.wordpress.com/2017/03/17/aprendizado-de-maquina-supervisionado-com-python/>. Acesso em: 17 jan. 2020.

IASC - INTELIGÊNCIA ARTIFICIAL SOB CONTROLE

Blog com textos relacionados à Inteligência Artificial.

ALGORITMOS E IMPLEMENTAÇÕES, APRENDIZADO DE MÁQUINA, DATA SCIENCE, INTELIGÊNCIA ARTIFICIAL, POSTAGENS COM DOWNLOADS, SOFTWARE LIVRE, SUBÁREAS DA IA

Aprendizado de Máquina Supervisionado com Python



► Referências

BELL, R. M.; KOREN, Y; VOLINSKY, Chris. **All together now:** A perspective on the netflix prize. *Chance*, v. 23, n. 1, 2010.

CASTELLI, M.; MANZONI, L.; VANNESCHI, L. *et al.* **An expert system for extracting knowledge from customers reviews:** the case of Amazon. com, Inc. *Expert Systems with Applications*, 2017.

KWOK. L, BEI Y. **Spreading social media messages on Facebook:** an analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly* 54.1, 2013.

MARSLAND, S. **Machine Learning:** an algorithmic perspective. Chapman and Hall/CRC, 2014.

► Referências

MÜLLER, A. C., AND SARAH, G., **Introduction to Machine Learning with Python: a guide for data scientists**. O'Reilly Media, 2016.

PEREIRA, D.R. **Linguagens de programação para ciência de dados (Python com Spark)**. Londrina: Editora e Distribuidora Educacional S.A., 2019.

SCIKIT-LEARN. **Boston house prices dataset**. 2019. Disponível em: <https://scikit-learn.org/stable/datasets/index.html#boston-house-prices-dataset>. Acesso em: 17 jan. 2020.

SCIKIT-LEARN. **Breast cancer wisconsin (diagnostic) dataset**. 2019. Disponível em: <https://scikit-learn.org/stable/datasets/index.html#breast-cancer-wisconsin-diagnostic-dataset>. Acesso em: 17 jan. 2020.

