

Curso de Especialização em Aprendizagem de Máquina em Inteligência Artificial

Disciplina: Aprendizagem de Máquina

AULA 01

Prof. Gustavo Gattass Ayub



O que é aprendizagem de máquina?

Nota: Daqui por diante sempre que usarmos a abreviação **AM** estamos nos referindo a Aprendizagem de Máquina.

■ Algumas definições...

Machine Learning is defined as an automated process that extracts patterns from data.

John Kelleher and Brian Mac Namee and Aoife D'Arcy

Machine Learning is the science (and art) of programming computers so they can learn from data.

Aurélien Guéron

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .

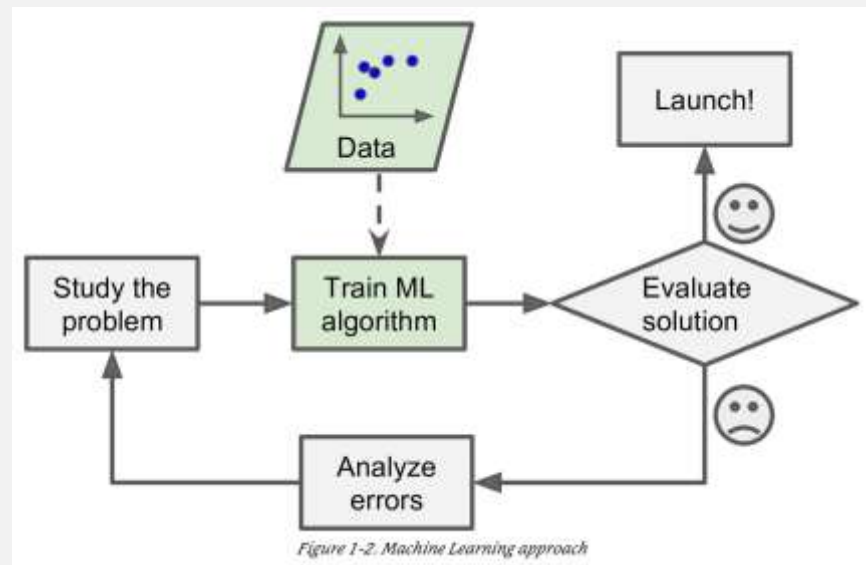
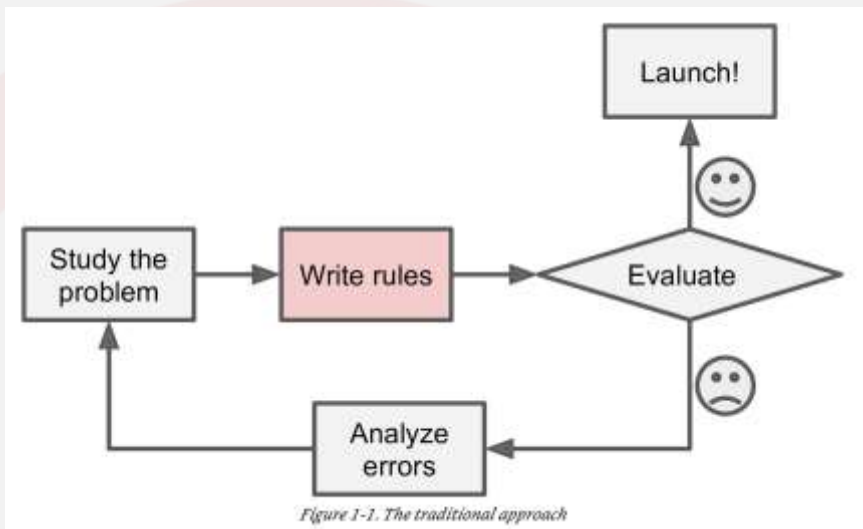
Tom Mitchell, 1997

■ Algumas definições...(Cont.)

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Wikipedia, adaptado de Arthur Samuel & Bishop

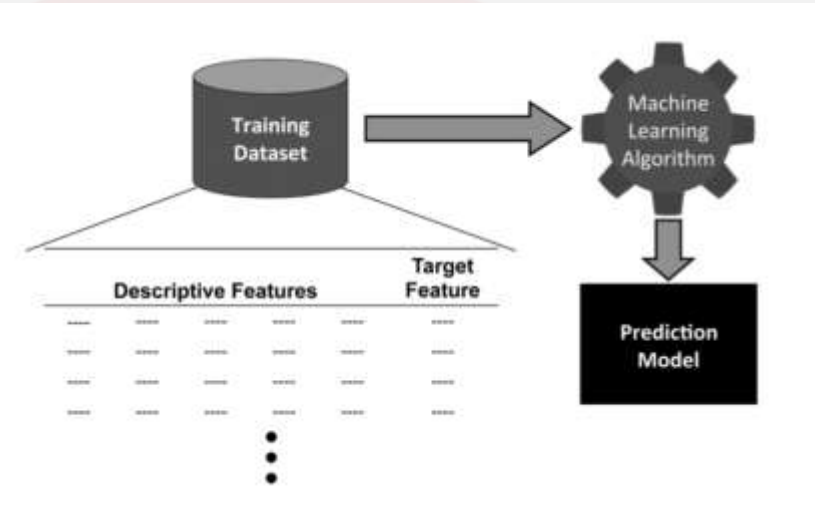
Diferentes abordagens



Fonte: GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

■ Quando usamos AM?

- Problemas complexos que não podem ser equacionados ou facilmente equacionados (a partir de formulas ou regras)
- Estamos diante de problemas (ou desafios) com os seguintes complicadores:
 1. Capacidade de modelar o problema a partir de regras
 2. Volume de variáveis ou dados
 3. Volatilidade das variáveis exigindo maior adaptabilidade



No processo de aprendizagem buscamos construir um modelo preditivo que permita prever (ou estimar) as características desejadas (“target feature” ou estimador ou função objetivo) a partir de características descritivas (ou “descriptive features” ou parâmetros de entrada).

Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

■ Adaptabilidade

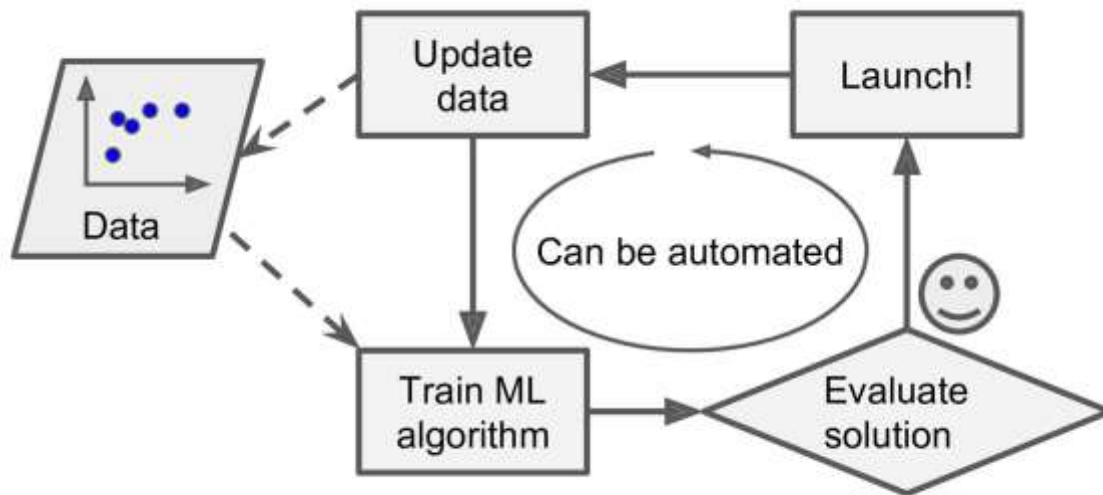
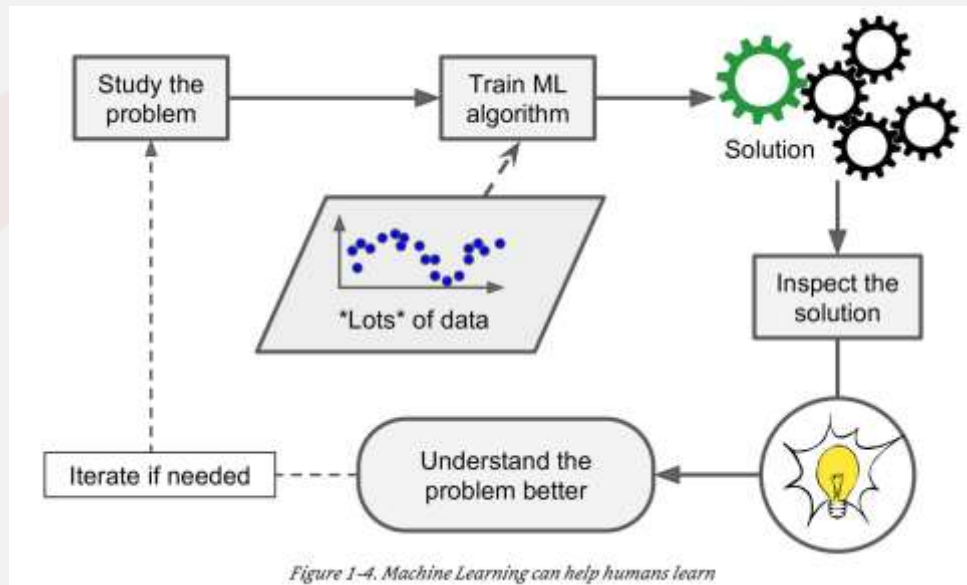


Figure 1-3. Automatically adapting to change

Fonte: GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Mineiração de Dados

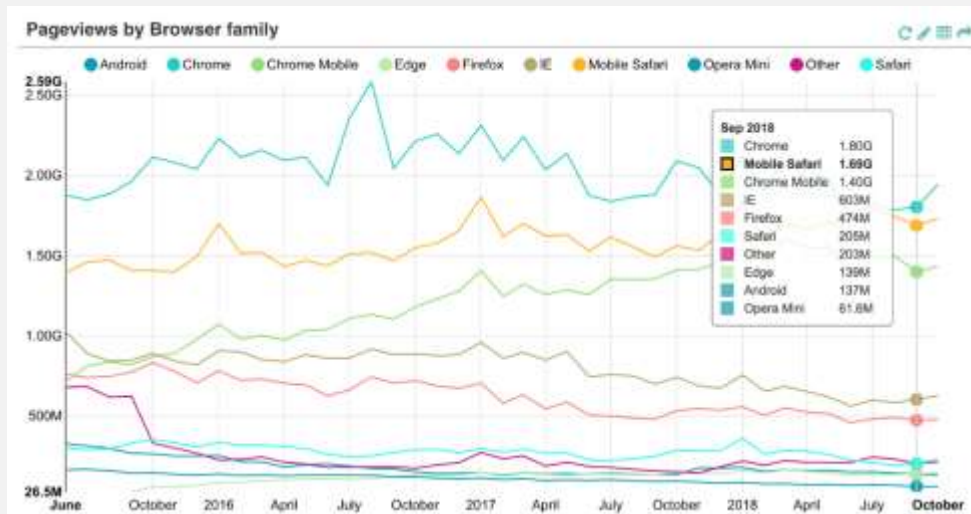


Fonte: GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn and TensorFlow

Analytics

***Analytics** is the discovery, interpretation, and communication of meaningful patterns in data; and the process of applying those patterns towards effective decision making. In other words, analytics can be understood as the connective tissue between data and effective decision making, within an organization.*

Wikipedia

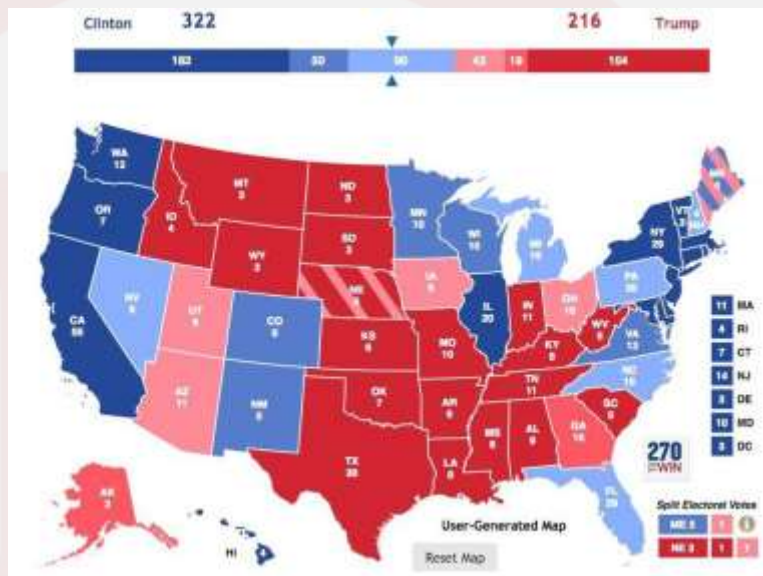


***Analytics** is the art of finding out what's in your data.
Analytics deals with what you know.
Statistics deals with that you don't.
Statistical Thinking, Cassie Kozyrkov*

Estatística

Statistics is the science of making decisions under uncertainty.

“The foundations of statistics” Leonard Savage, 1954



■ Tomada de Decisão

Decision-making (in psychology) is regarded as the cognitive process resulting in the selection of a belief or a course of action among several alternative possibilities.

Decision-making is the process of identifying and choosing alternatives based on the values, preferences and beliefs of the decision-maker. Every decision-making process produces a final choice, which may or may not prompt action. **Wikipedia.**

Common Issues (Wikipedia):

- Analysis Paralysis
- Information Overload ("a gap between the volume of information and the tools we have to assimilate", "excessive information affects problem processing and tasking, which affects decision-making")
- Post-decision Analysis

■ AM vs Mineiração de Dados

- Muitas vezes os termos “aprendizagem de máquina” e “mineiração de dados” aparecem como sinônimos mas isso não é correto.
- Aprendizagem de máquina é uma tecnologia de predição, que foca no processo de aprendizado e identificação de padrões a partir de dados de treinamento enquanto que em mineiração o foco está na descoberta de informações (na maioria das vezes aplicando aprendizagem de máquina) para a descoberta de novas informações.
- **KDD:** Knowledge Discovery and Data Mining

■ Conjunto de treinamento

- Conjunto de treinamento contém uma instância (dados históricos) do problema sendo observado com dados (ou sinais) que permitem descrevê-lo bem como apresentam os dados alvo, ou objetivo. Um bom conjunto de treinamento apresenta bons exemplos de instâncias do problema (amostra).

Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

■ Objetivos do processo de aprendizagem

- O objetivo do processo é selecionar um modelo preditivo que seja capaz de uma boa generalização, ou seja, a partir do conjunto de treinamento o modelo deve operar com boa precisão com dados mais gerais.

Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

Exemplo

ID	OCCUPATION	AGE	LOAN-SALARY	
			RATIO	OUTCOME
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

What is the relationship between the **descriptive features** (OCCUPATION, AGE, LOAN-SALARY RATIO) and the **target feature** (OUTCOME)?

```
if LOAN-SALARY RATIO > 3 then
    OUTCOME='default'
else
    OUTCOME='repay'
end if
```

Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

■ Desafios

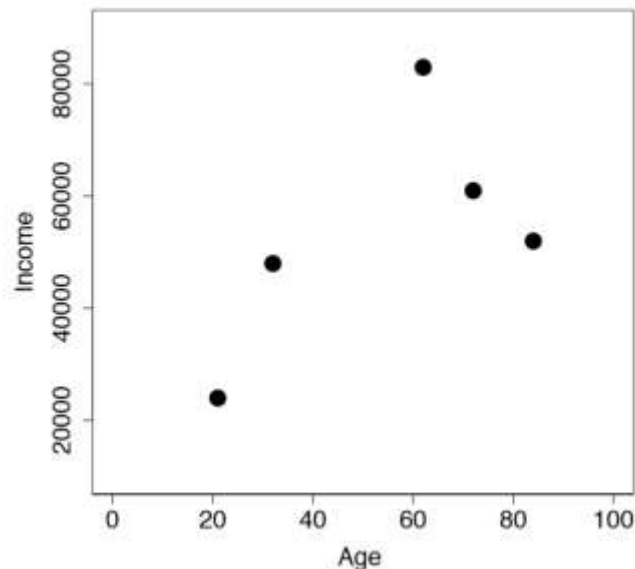
- Como dito anteriormente um dos objetivos do treinamento é a busca por modelos que permitam generalizar no entanto nesse processo podemos cair em duas situações:
- Underfitting que ocorre quando o modelo produzido a partir de um conjunto de treinamento é simplista demais para a generalização ou
- Overfitting que ocorre quando o modelo produzido é complexo demais a ponto de se tornar sensível a ruído.

Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

■ Ilustrando o conceito

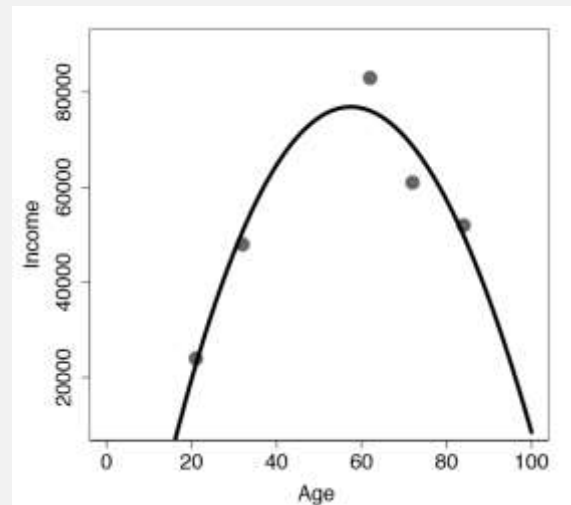
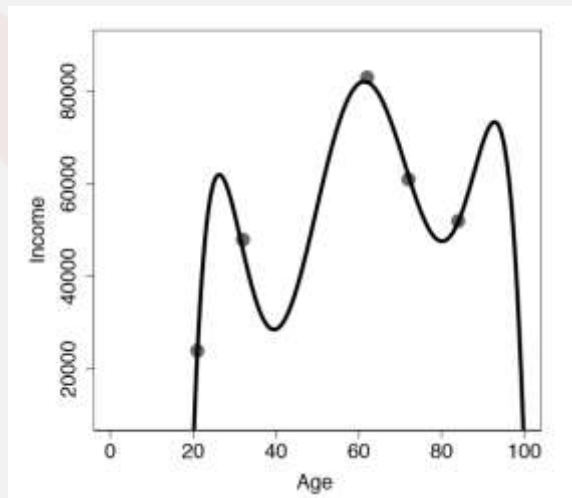
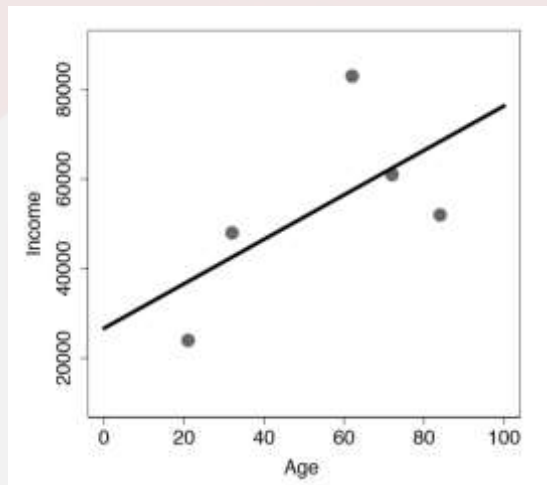
Table: The age-income dataset.

ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000



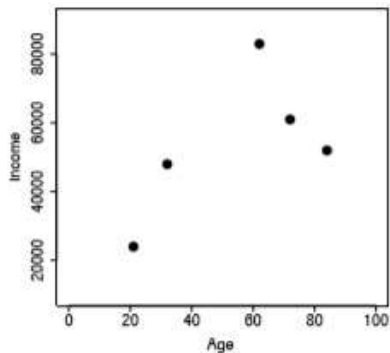
Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

■ Ilustrando o conceito (cont.)

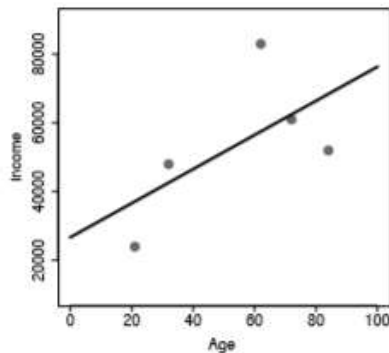


Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

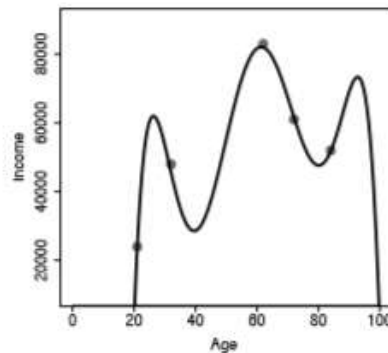
■ Ilustrando o conceito (cont.)



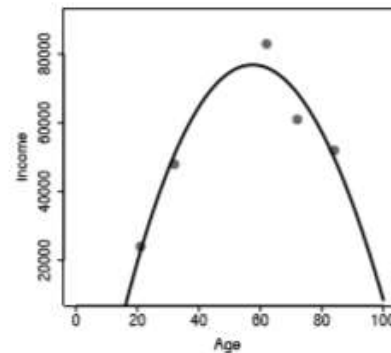
(a) Dataset



(b) Underfitting



(c) Overfitting



(d) Just right

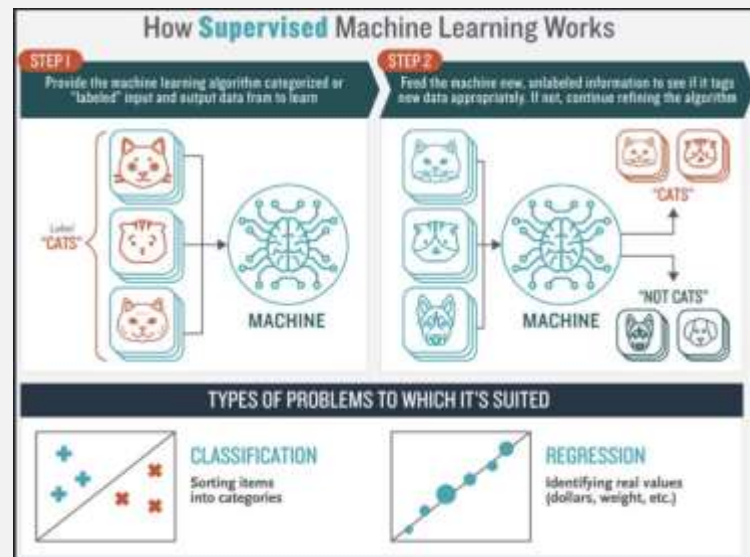
Fonte: KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

■ Tipos de Aprendizagem de Máquina

- Por supervisão, ou seja, como a interação humana se dá no processo de treinamento:
 - Supervisionado
 - Não Supervisionado
 - Por Reforço

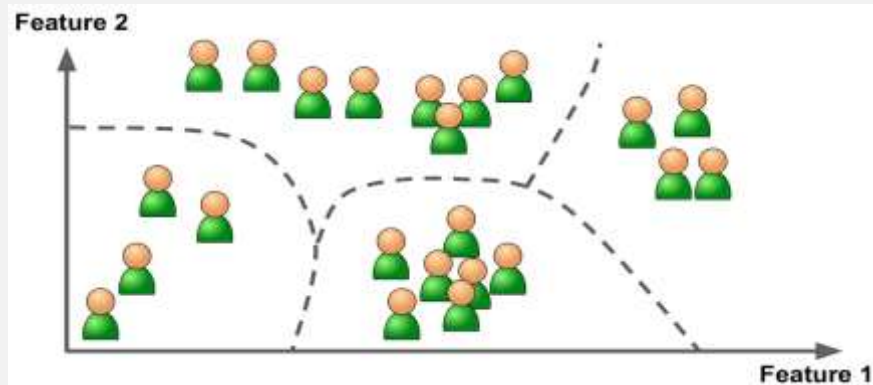
Aprendizado Supervisionado

- O conjunto de treinamento inclui labels (ou os valores desejados/objetivo).
- Eficiente em problemas de classificação e regressão.
- Exemplos de algoritmos:
 - **K-Nearest Neighbors**
 - **Regressão Linear**
 - **Regressão Logística**
 - **Support Vector Machines**
 - **Decision Trees & Random Forests**
 - **Redes Neurais**



Aprendizado Não Supervisionado

- O conjunto de treinamento não inclui labels (ou os valores desejados/objetivo).
- Eficiente em problemas de classificação e regressão.
- Exemplos de algoritmos:
 - **Clustering** (K-means, Hierarchical Cluster Analysis, Expectation Maximization)
 - **Visualization and Dimensionality Reduction** (PCA: Principal Components Analysis, Kernel PCA, Locally-Linear Embedding, t-SNE: t-distributed Stochastic Neighbor Embedding)
 - **Association Rules** (apriori, eclat)





Regressão e Classificação

■ Regressão e Classificação

- Usam métodos de aprendizagem supervisionada
- Os modelos de classificação produzem um mapeamento rotulando (ou categorizando) variáveis de entrada. Tipicamente esses modelos usam classificação binária, ou seja, mapeiam as entradas entre dois rótulos (ex. Sim/Não). Dentre algumas aplicações estão a classificação de objetos em imagens e sistemas de anti-spam.
- Os modelos de regressão por outro lado produzem um mapeamento que atribui uma quantidade (números inteiros ou reais) às variáveis de entrada.

■ Exemplos de Algoritmos

Regressão

- Regressão Linear Simples
- Regressão Multi-variada
- Regressão Polinomial

Classificação

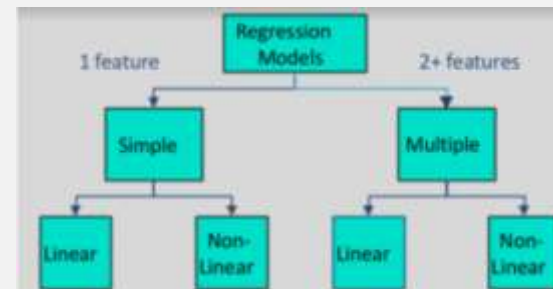
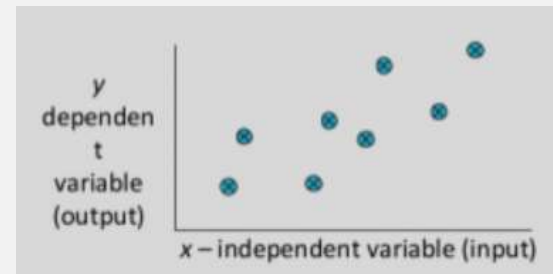
- Regressão Logística
- SVM (Support Vector-Machines)
- Decision Trees
- Random Forests
- Naïve Bayes



Regressão

Regressão

- Regression is a method of modelling a target value based on independent predictors.
- This method is mostly used for forecasting and finding out cause and effect relationship between variables.
- Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



■ Diagramas de Dispersão

Diagramas de dispersão que sugerem uma regressão linear entre as variáveis



Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)



Existência de correlação negativa (em média, quanto maior for a colheita menor será o preço)

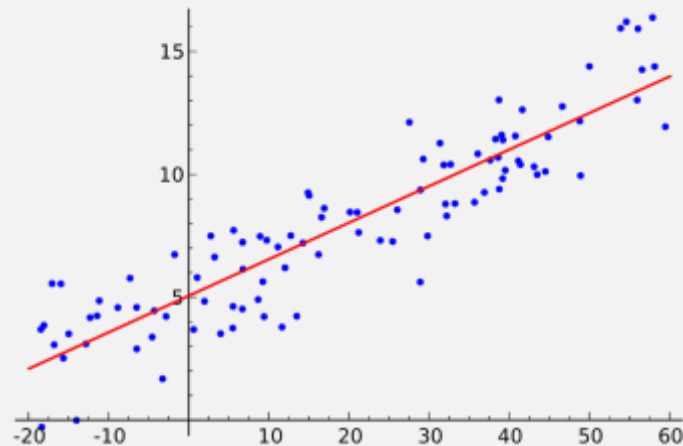
■ Diagramas de Dispersão

Diagramas de dispersão que sugerem uma regressão não linear entre as variáveis



Linear Regression

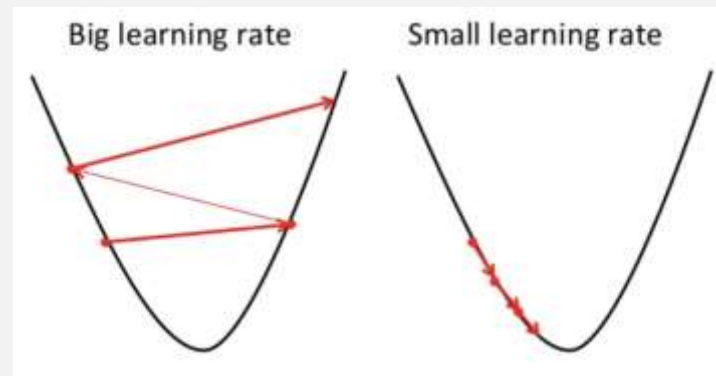
- Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.
- $Y = a + bX$
- The cost function helps us to figure out the best possible values for **a** and **b** which would provide the best fit line for the data points.



$$\text{minimize } \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

■ Gradiente Descendente

- Gradient descent is a method of updating a and b to reduce the cost function(MSE)
 - método dos quadrados mínimos
- The idea is that we start with some values for a and b and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.



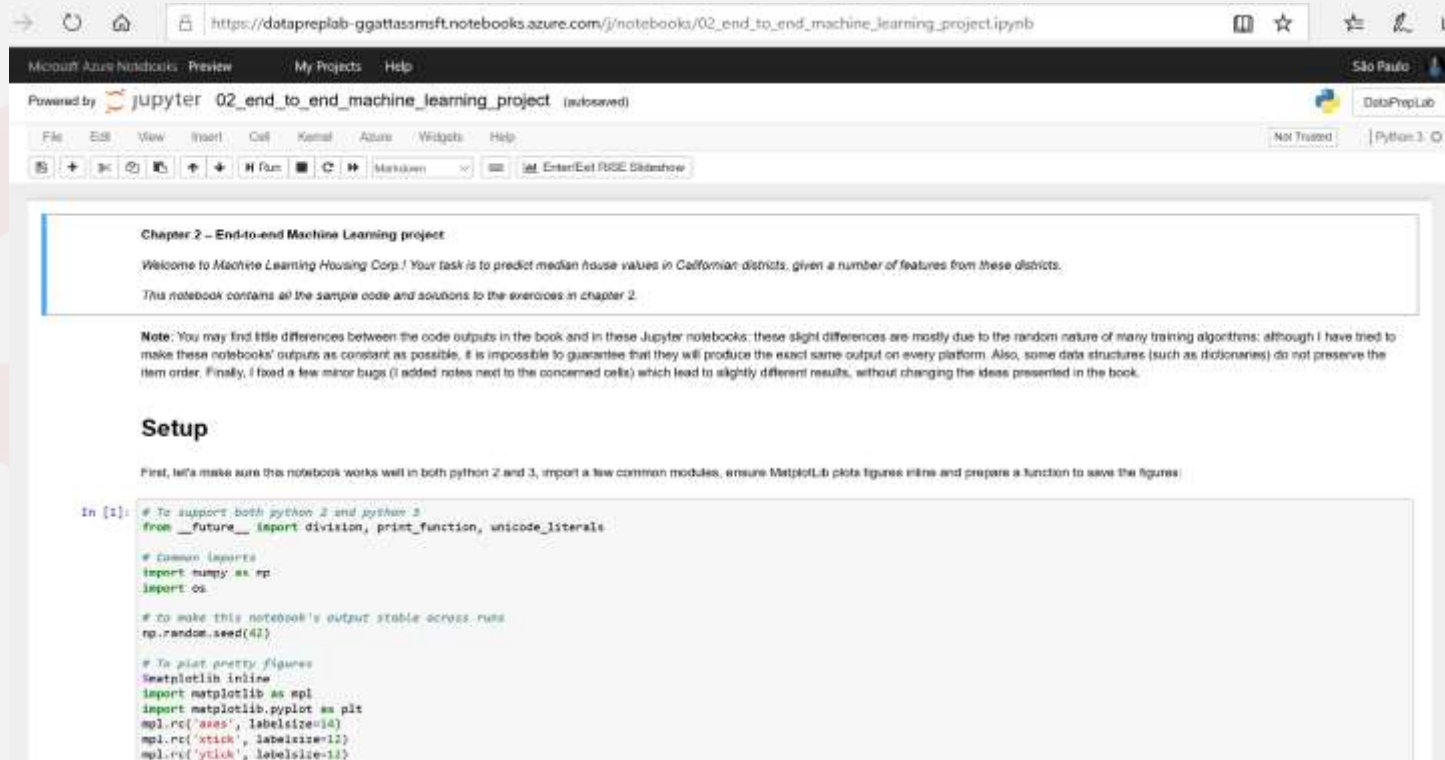
Estudo Obrigatório

Alunos devem apreender Python


Ferramentas de Data Science

- Jupiter Lab
- Anaconda
- Azure Notebooks
- AWS SageMaker (Notebooks), EMR Notebooks
- Google AI Platform Notebooks (Google Collab)

Exemplo: Azure Notebooks



Microsoft Azure Notebooks | Preview | My Projects | Help | São Paulo

Powered by  jupyter 02_end_to_end_machine_learning_project (autosaved) | DataPrepLab

File Edit View Insert Cell Kernel Azure Widgets Help | Not Trusted | Python 3.0

Chapter 2 – End-to-end Machine Learning project

Welcome to Machine Learning Housing Corp.! Your task is to predict median house values in Californian districts, given a number of features from these districts.

This notebook contains all the sample code and solutions to the exercises in chapter 2.

Note: You may find little differences between the code outputs in the book and in these Jupyter notebooks: these slight differences are mostly due to the random nature of many training algorithms: although I have tried to make these notebooks' outputs as constant as possible, it is impossible to guarantee that they will produce the exact same output on every platform. Also, some data structures (such as dictionaries) do not preserve the item order. Finally, I fixed a few minor bugs (I added notes next to the concerned cells) which lead to slightly different results, without changing the ideas presented in the book.

Setup

First, let's make sure this notebook works well in both python 2 and 3, import a few common modules, ensure Matplotlib plots figures inline and prepare a function to save the figures:

```
In [1]: # To support both python 2 and python 3
from __future__ import division, print_function, unicode_literals

# Common imports
import numpy as np
import os

# to make this notebook's output stable across runs
np.random.seed(42)

# To plot pretty figures
%matplotlib inline
import matplotlib as mpl
import matplotlib.pyplot as plt
mpl.rc('axes', labelsize=14)
mpl.rc('xtick', labelsize=12)
mpl.rc('ytick', labelsize=12)
```

Fonte: <https://notebooks.azure.com/>

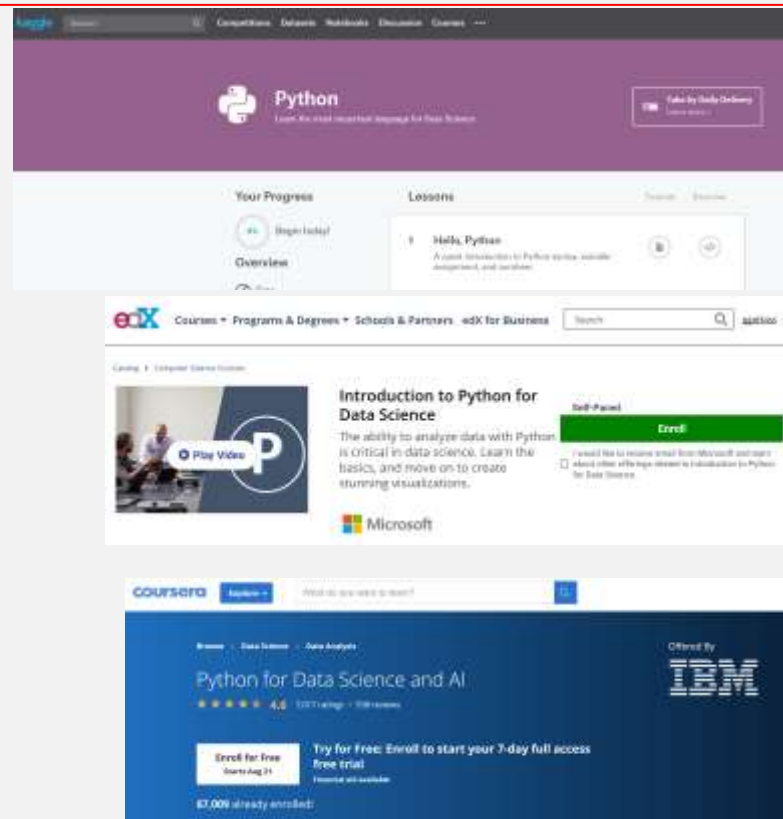
Ecosistema Python para Aprendizagem de Máquina

Python

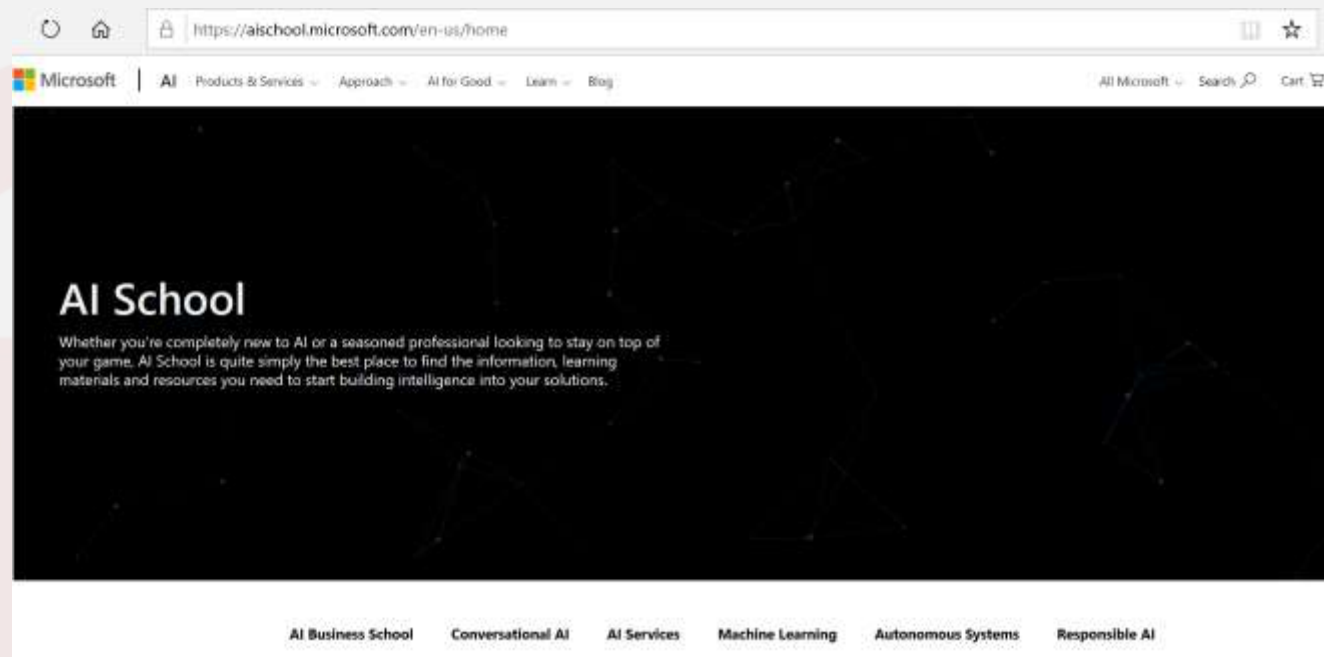
- General purpose interpreted programming language, high-level, versatile and object-oriented. One of the most widely used programming languages
- Easy to understand (learning curve), powerful to execute.
- Huge library support:
 - SciPy (NumPy, Matplotlib, Pandas): Data arrays, 2D charting, data structure/analysis
 - Scikit Learn, Tensorflow: machine learning
 - Theano: numerical computation (optimized for different processors architectures)
 - Keras: high-level neural networks (fast experimentation)
 - NLTK: natural language toolkit

Como aprender Python?

- Tutoriais
 - <https://docs.python.org/3/tutorial/index.html>
 - <https://python.swaroopch.com/>
- Exemplos
- Cursos
 - Kaggle
 - edX
 - Coursera
- YouTube
 - <https://www.youtube.com/playlist?list=PLlrXD0HtieHhS8VzuMcfQD4uJ9yne1mE6>



Microsoft AI School (Exemplos)



<https://notebooks.azure.com/home/projects/Python>

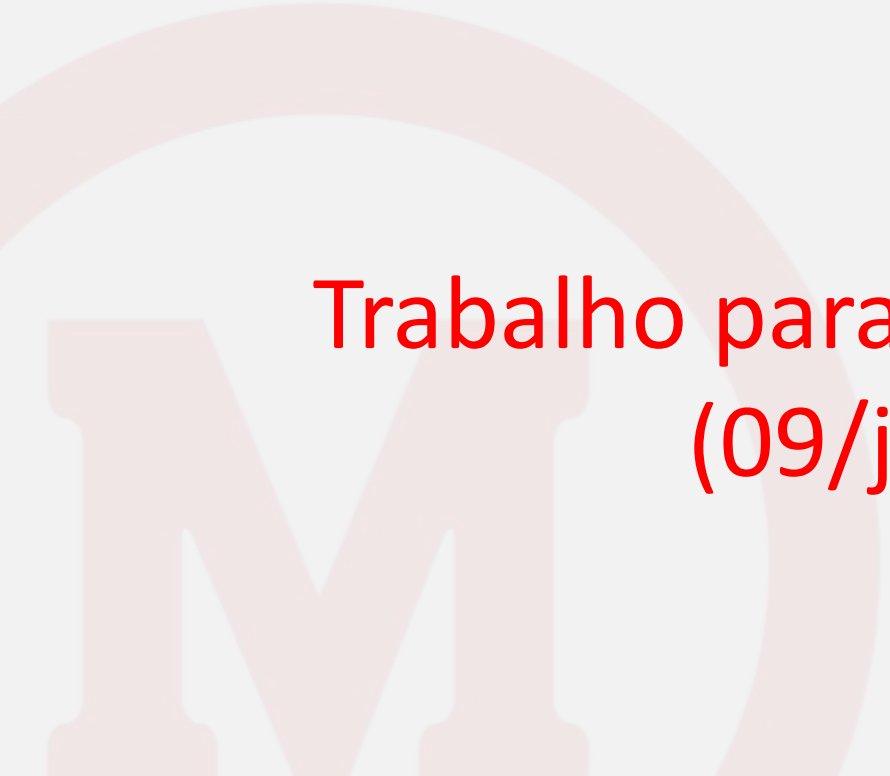
■ Setup Azure Notebooks

Criar uma conta

<https://notebooks.azure.com/>

Importar Projeto

<https://github.com/MicrosoftDocs/ms-learn-ml-crash-course-python>



Trabalho para a próxima aula (09/jun/20)

■ Para a próxima aula (09/jun/20)

Exercício Complementar (não vale nota)

- [Introduction to Machine Learning Algorithms: Linear Regression](#)
- [Linear Regression for Machine Learning](#)
- [Difference Between Classification and Regression in Machine Learning](#)

Exercício Complementar - Estudo da linguagem Python (não vale nota)

- Kaggle – [Curso Introdução ao Python](#)
- Kaggle – [Curso Introdução ao Pandas](#)

Exercício de Aprofundamento (vale nota)

- Resolver o primeiro problema – Basic Example of Linear Regression (use o exemplo 02 do MS Crash Course como exemplo)
- Link: <https://www.kaggle.com/c/atx-ex-regression/data>



Até a próxima aula

MUITO OBRIGADO!