Aprendizado de Máquina Open Source

Msc.João Paulo V.M Martins
UnB - FGA

Introdução

- Aprendizado de Máquina (Machine Learning) é uma disciplina científica focada no estudo e **construção de algoritmos que aprendem a partir de dados.**
- **Esses algoritmos constroem modelos a partir de dados de entrada e os usa para fazer predições ou tomar decisões.**

Introdução

“Campo de estudo que dá aos computadores a habilidade de aprender sem terem sido explicitamente programados.”

- Arthur Samuel

Machine Learning vs IA

- Machine Learning é uma sub área da IA.
- IA é focada em criar uma **Inteligência Artificial Geral.**
- Machine Learning é focada em otimização extrema de solução de problemas **pontuais**.

Aplicações de Machine Learning

- Reconhecimento de caracteres;
- Detecção de Faces;
- Filtro de Spam;
- Reconhecimento de fala;
- Diagnóstico médico
- Segmentação de Clientes;
- Detecção de fraudes.

Quem Usa?

- Google;
- Facebook;
- Yahoo;
- Tesla Motors;
- Microsoft;
- Baidu;
- E qualquer grande empresa que se preze.

A “onda” do Open Source

- Em 2015 o **Google** abriu sua biblioteca para o mundo. A ***TensorFlow.***
- **O Facebook, Twitter, Nvidia, Google Deep Mind,** entre outros, desenvolvem seus algoritmos e sistemas de aprendizado de máquina também em um biblioteca open source, a ***Torch7*****.**

Outra bibliotecas Open

- Utilizadas na indústria:
  - Scikit-Learn (Python);
  - Shogun (C++);
  - Mllib – Apache Spark (Python, Java, Scala);
  - H2O (Java);
  - Weka (Java).

Porque eu deveria me importar com ML?

“Machine Learning é a próxima Internet”.

Tony Tether, DARPA


ALPHAGO
00:01:00
LEE SEDOL
00:01:00
AlphaGo
Lee Sedol

Requisitos para o Campo

- Álgebra Linear;
- Cálculos;
- Cálculo Vetorial;
- Probabilidade clássica;
- Probabilidade Bayesiana.

Tipos de Aprendizados

- Supervisionado:
  - Você conta com uma entrada e uma saída desejada, **o objetivo é aprender uma regra que mapeia entradas em saídas.**
- Não Supervisionado:
  - Você tem apenas entradas, sem saídas desejadas. **O objetivo é descobrir padrões escondidos nos dados**.

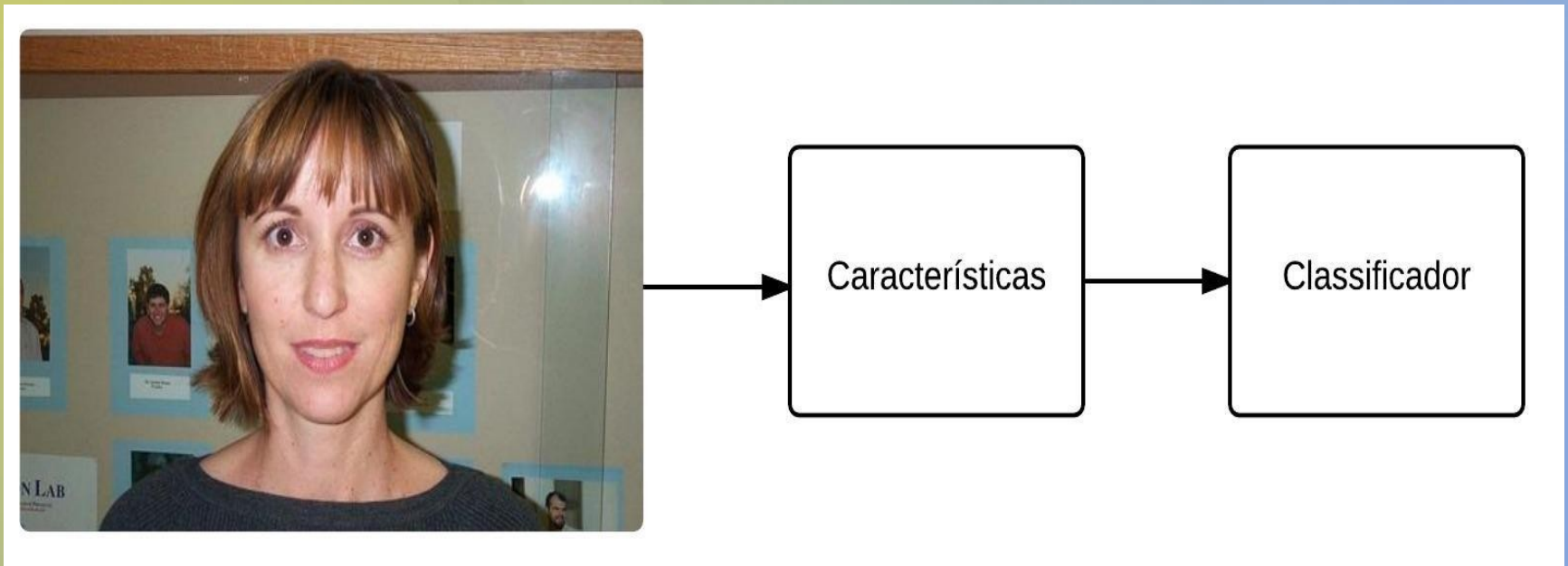
Problemas de Aprendizado

- Classificação:
  - É o aprendizado de **categorização** de objetos em categorias fixas.
- Regressão:
  - Na regressão tentamos aprender um valor real. Ex: Queremos prever **quanto** vai chover amanhã.

Problemas de Aprendizado

- Redução de Dimensionalidade:
  - Simplifica entradas mapeando-as em espaços de menores dimensões.
- Clusterização:
  - Um conjunto de entradas é dividido em grupos. **Diferente da classificação, os grupos não são dados de ante mão.**

Estrutura de Algoritmos em ML



O que vamos utilizar?

- Linguagem Python + Jupyter Notebook;
- SciPy Stack;
- Biblioteca Scikit-Learn.

Jupyter Notebook

- É uma aplicação web que permite a criação e o compartilhamento de documentos.
- Muito utilizado para data science e computação científica.
- + de 40 linguagens disponíveis.
- Evolução do IPython Notebook.

SciPy Stack

- Ecossistema baseado em Python para computação científica.
- No “pacote” vem:
  - NumPy;
  - IPython;
  - Sympy
  - Matplotlib;
  - Pandas;

Scikit-Learn

- É uma biblioteca open-source de machine learning;
- Baseada em Python;
- Diversos datasets pré moldados;
- Interessante para aprender;
- Foi criada por David Cournapeau;
- Se tornou popular em 2012.

Contato

- **E-mail:** jpmarinho.martins@gmail.com
- **Github:** https://github.com/jpvmm