



Machine Learning* **com Python**

Parte 1 – Introdução

Tópicos

- Dados x Informação
- Aprendendo modelos
 - Tipos de aprendizado
- Problemas de Aprendizagem
- Overfitting e Underfitting
- Regularização

Dados

- Os dados são informações sobre o problema em que você está trabalhando

Dados x Informação

	Dados	Informação
Significado	Os dados são crus, fatos não organizados que precisam ser processados. Os dados podem ser algo simples e aparentemente aleatórios e inúteis até que estejam organizados.	Quando os dados são processados, organizados, estruturados ou apresentados em um determinado contexto, de modo a torná-lo útil, é chamado de informação.
Exemplo	O resultado do teste de cada aluno é uma peça de dados.	A pontuação média de uma aula ou de toda a escola é informação que pode ser derivada dos dados fornecidos
Etimologia	"Dados" vem de uma palavra latina singular, datum, que originalmente significava "algo dado". O seu primeiro uso data de 1600. Ao longo do tempo, "dados" tornou-se o plural do datum.	"Informação" é uma palavra antiga que remonta ao 1300 e tem origens de francês antigo e inglês médio. Sempre se referiu ao "ato de informar", geralmente em relação à educação, instrução ou outra comunicação de conhecimento.

Aprendendo um modelo

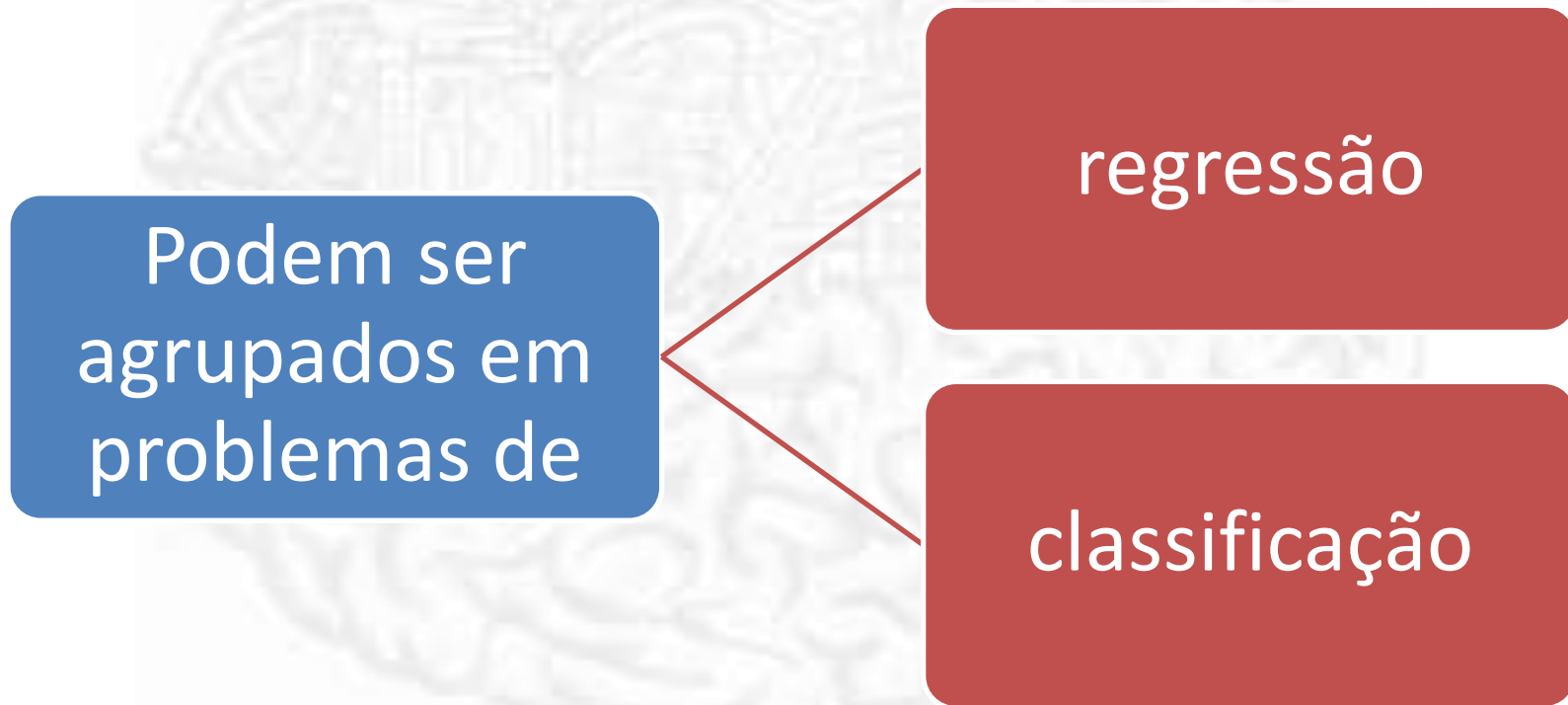
- Aprendizado supervisionado
- Aprendizado não supervisionado

Aprendizado supervisionado

- Na aprendizagem supervisionada, você possui variáveis de entrada (x) e uma variável de saída (Y)
- Você usa um algoritmo para aprender a função de mapeamento da entrada para a saída.

$$Y = f(X)$$

Problemas de Aprendizagem



Função de Aproximação

- Modelagem preditiva
 - desenvolver um modelo usando dados existentes para fazer uma previsão de novos dados onde não temos a resposta.
 - Tipos:
 - Classificação
 - Regressão

Aprendizado não supervisionado

- Você só possui dados de entrada (X) e nenhuma variável de saída correspondente
- Modelar a estrutura ou a distribuição subjacente nos dados, a fim de aprender mais sobre os dados.

Aprendizado semi-supervisionado

- Você possui:
 - uma grande quantidade de dados de entrada (X) e
 - apenas alguns dos dados são rotulados (Y)



APRENDIZADO SUPERVISIONADO

Erro de previsão

- O objetivo de qualquer algoritmo supervisionado é estimar melhor a função de mapeamento (f) para a variável de saída (Y) dados os dados de entrada (X)
- O erro de previsão para qualquer algoritmo de aprendizado de máquina pode ser dividido em três partes:
 - Erro de viés (Bias Error)
 - Erro de variância (Variance Error)
 - Erro irreduzível (Irreducible Error)

Erro de viés (bias)

- Viés (Bias) são os pressupostos simplificadores feitos por um modelo para tornar a função alvo mais fácil de aprender.
 - Low Bias: sugere menos pressupostos sobre a forma da função alvo.
 - High-Bias: Sugere mais suposições sobre a forma da função alvo.

Erro de variância

- A diferença é a quantidade que a estimativa da função alvo irá mudar se diferentes dados de treinamento fossem usados.

Compromisso entre viés e variância

- Um algoritmo supervisionado deve ser alcançar baixo viés com baixa variação.
- Deve ter um bom desempenho de previsão.
- Compromisso:
 - Aumentar o viés diminuirá a variância.
 - Aumentar a variância diminuirá o viés.



DEFINIÇÕES E TERMINOLOGIA

Definições e terminologia

Exemplos

Características (features)

Etiquetas (labels)

Principais classes de problemas de aprendizagem

Classificação

Regressão

Ranking

Clustering

Redução de dimensão ou aprendizado múltiplo



CLASSIFICAÇÃO E REGRESSÃO

Classificação x Regressão

- A classificação é a tarefa de prever um rótulo de classe discreto.
- A regressão é a tarefa de prever uma quantidade contínua.

Classificação

- A modelagem preditiva de classificação é a tarefa de aproximar uma função de mapeamento (f) das variáveis de entrada (X) para variáveis de saída discretas (y).
- As variáveis de saída geralmente são chamadas de rótulos ou categorias. A função de mapeamento prevê a classe ou categoria para uma determinada observação.
- Por exemplo, um e-mail de texto pode ser classificado como pertencente a uma das duas classes: "spam" e "não spam".

Regressão

- A modelagem preditiva de regressão é a tarefa de aproximar uma função de mapeamento (f) das variáveis de entrada (X) para uma variável de saída contínua (y).
- Uma variável de saída contínua é um valor real, como um número inteiro ou valor de ponto flutuante. Estas são muitas vezes quantidades, como quantidades e tamanhos.
- Por exemplo, uma casa pode estar prevista para vender por um valor específico em dólares, talvez na faixa de US \$ 100.000 a US \$ 200.000.

Converter Regressão em Classificação

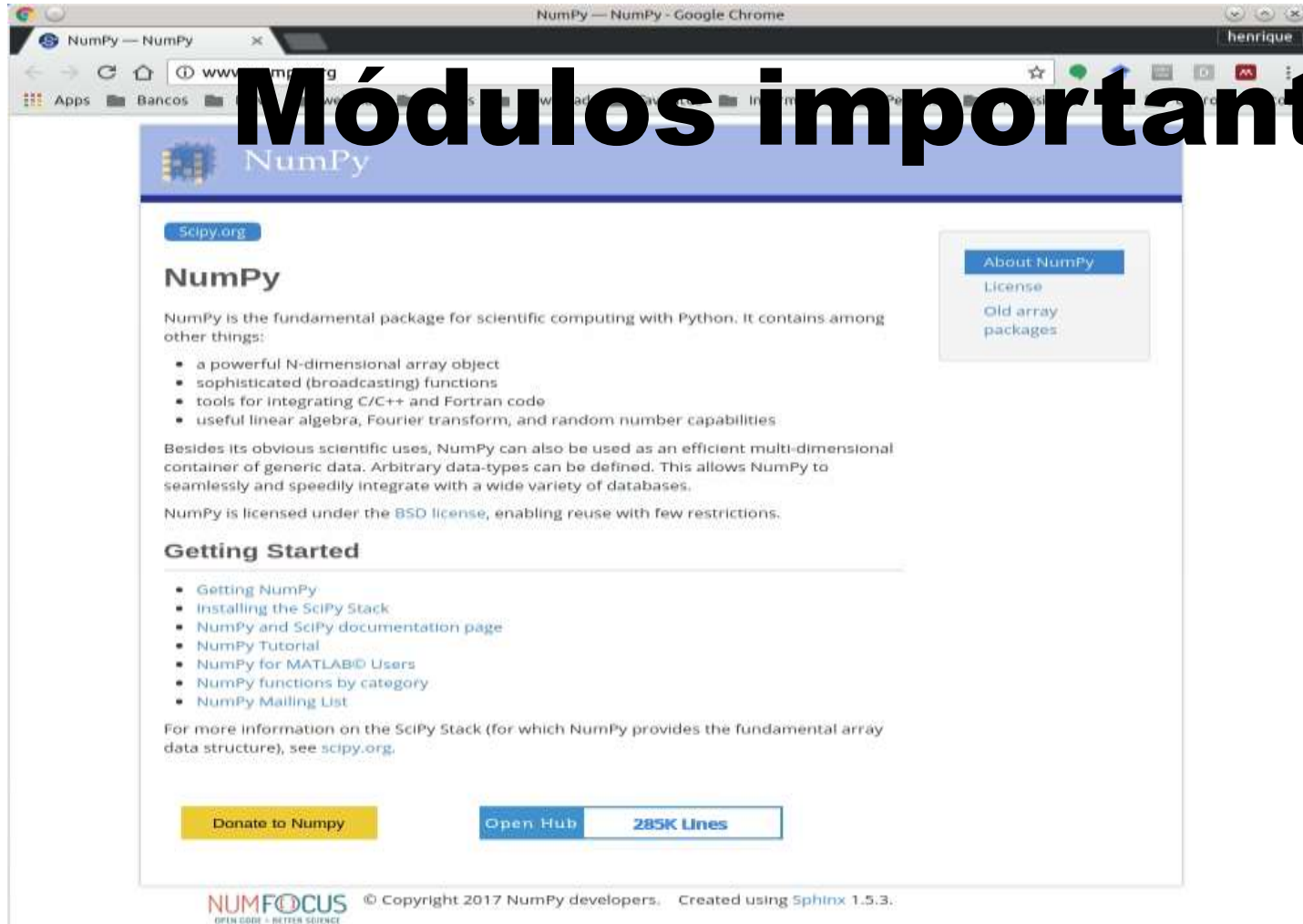
- Em alguns casos, é possível converter um problema de regressão para um problema de classificação.
 - Por exemplo, a quantidade de soja a ser enviada em um navio a prever pode ser convertida em containers discretos.
 - Isso geralmente é chamado de discretização

PYTHON

Módulos importantes

- Numpy
- Scipy
- Matplotlib
- Pandas
- Scikit-learn
- Tensorflow
- Jupyter

Módulos importantes



NumPy — NumPy - Google Chrome

NumPy

Scipy.org

NumPy

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

NumPy is licensed under the [BSD license](#), enabling reuse with few restrictions.

Getting Started

- [Getting NumPy](#)
- [Installing the SciPy Stack](#)
- [NumPy and SciPy documentation page](#)
- [NumPy Tutorial](#)
- [NumPy for MATLAB® Users](#)
- [NumPy functions by category](#)
- [NumPy Mailing List](#)

For more information on the SciPy Stack (for which NumPy provides the fundamental array data structure), see [scipy.org](#).

Donate to NumPy

Open Hub 285K Lines

NUMFOCUS
OPEN CODE - BETTER SCIENCE

© Copyright 2017 NumPy developers. Created using [Sphinx](#) 1.5.3.

About NumPy
License
Old array packages

Módulos importantes

The screenshot shows the SciPy.org website in a Google Chrome browser. The page features a blue header with the SciPy logo and the text "SciPy.org" and "Sponsored by ENTHOUGHT". Below the header, there are four circular icons for "Install", "Getting Started", "Documentation", and "Report Bugs", each with a SciPy logo. Below these is an orange RSS icon labeled "Blogs". A paragraph describes SciPy as a Python-based ecosystem of open-source software for mathematics, science, and engineering. Below this, there are three columns of core packages, each with an icon and a description: NumPy (Base N-dimensional array package), Matplotlib (Comprehensive 2D Plotting), and SymPy (Symbolic mathematics). To the right of these columns is a sidebar with a list of links: About SciPy, Install, Getting Started, Documentation, Bug Reports, Topical Software, Citing, Cookbook, SciPy Conferences, Blogs, and NumFOCUS. At the bottom right, there is a section titled "CORE PACKAGES:" with links to Numpy, SciPy library, Matplotlib, IPython, and SymPy.

NumPy — NumPy

SciPy.org — SciPy.org

https://www.scipy.org

Apps Bancos Drive webmail Cursos download Favoritos Informacoes Pesquisa Outros Favoritos

SciPy.org

Install Getting Started Documentation Report Bugs

Blogs

SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

NumPy
Base N-dimensional array package

Matplotlib
Comprehensive 2D Plotting

Sympy
Symbolic mathematics

SciPy library
Fundamental library for scientific computing

IPython
Enhanced Interactive Console

pandas
Data structures & analysis

About SciPy
Install
Getting Started
Documentation
Bug Reports
Topical Software
Citing
Cookbook
SciPy Conferences
Blogs
NumFOCUS

CORE PACKAGES:
Numpy
SciPy library
Matplotlib
IPython
SymPy

Módulos importantes

NumPy — NumPy - Google Chrome

SciPy.org — SciPy.org

Matplotlib: Python plotting — Matplotlib 2.1.2 documentation - Google Chrome

Seguro | <https://matplotlib.org>

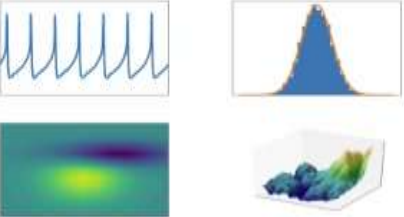
Apps Bancos Drive webmail Cursos download Favoritos Informacoes Pesquisa Profissional Outros Favoritos

matplotlib
Version 2.1.2

[home](#) | [examples](#) | [tutorials](#) | [pyplot](#) | [docs](#) » [modules](#) | [index](#)

Introduction

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the [jupyter](#) notebook, web application servers, and four graphical user interface toolkits.



Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the [sample plots](#) and [thumbnail gallery](#).

For simple plotting the [pyplot](#) module provides a MATLAB-like interface, particularly when combined with IPython. For the power users you have full control of line, color, font, annotation, image, projection, etc. via the

powered by: NumFOCUS
Depsy: 100th percentile
Travis-CI: build passing

Support matplotlib

Support NumFOCUS

Quick search

Go

Módulos importantes

Python Data Analysis Library — pandas: Python Data Analysis Library - Google Chrome

Seguro | <https://pandas.pydata.org>

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

home // about // get pandas // documentation // community // talks // donate

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](#) programming language.

pandas is a [NumFOCUS](#) sponsored project. This will help ensure the success of development of pandas as a world-class open-source project, and makes it possible to [donate](#) to the project.

A Fiscally Sponsored Project of

NUMFOCUS

OPEN CODE = BETTER SCIENCE

v0.22.0 Final (December 29, 2017)

VERSIONS

Release

0.22.0 - December 2017
[download](#) // [docs](#) // [pdf](#)

Development

0.23.0 - 2018
[github](#) // [docs](#)

Previous Releases

0.21.1 - [download](#) // [docs](#) // [pdf](#)
0.21.0 - [download](#) // [docs](#) // [pdf](#)
0.20.3 - [download](#) // [docs](#) // [pdf](#)
0.19.2 - [download](#) // [docs](#) // [pdf](#)
0.18.1 - [download](#) // [docs](#) // [pdf](#)
0.17.1 - [download](#) // [docs](#) // [pdf](#)
0.16.2 - [download](#) // [docs](#) // [pdf](#)
0.15.2 - [download](#) // [docs](#) // [pdf](#)

Fork me on GitHub

modules | index

built by: NumFOCUS

100% percentile

build passing

Support matplotlib

Support NumFOCUS

search

Módulos importantes

The image displays a stack of web browser windows, each showing the official documentation for a different Python data science library. The windows are arranged in a layered fashion, with the scikit-learn documentation being the most prominent in the foreground. The libraries shown include NumPy, SciPy, Matplotlib, pandas, and scikit-learn. The scikit-learn window features a grid of small plots and a list of key features: Simple and efficient tools for data mining and data analysis, Accessible to everybody, and reusable in various contexts, Built on NumPy, SciPy, and matplotlib, and Open source, commercially usable - BSD license. Below the main header, the documentation is organized into sections for Classification, Regression, and Clustering, each with a brief description and a list of applications and algorithms.

NumPy — NumPy - Google Chrome

SciPy.org — SciPy.org

Matplotlib: Python plotting — Matplotlib 2.1.2 documentation - Google Chrome

Python Data Analysis Library — pandas: Python Data Analysis Library - Google Chrome

scikit-learn: machine learning in Python — scikit-learn 0.19.1 documentation - Google Chrome

scikit-learn

Home Installation Documentation Examples

Google Custom Search

scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors.

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso.

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes.

Algorithms: k-Means, spectral clustering.

Módulos importantes



Instalação no Ubuntu

Veja nosso vídeo:

