

Laboratory class 0

Nomes:

1 - Josemar Figueiredo Pereira

Getting started with Python, IPython Notebook, Anaconda, and Scikit-learn

Welcome to the Machine Learning course. In this course we will introduce the basic concepts of machine learning. The goal of this course is for your effort to be spent on learning the fundamental concepts and algorithms behind machine learning in a hands-on fashion. These concepts transcend any single package. What you learn here you can use whether you write code from scratch, use any existing ML packages out there, or any that may be developed in the future.

The learning approach in this course is to start from use cases and then dig into algorithms and methods, what we call a case-studies approach. The lab classes are focused on understanding how ML can be used in various cases studies, and the practical assignments will dig into the details of algorithms and methods for each of the main ML areas. In the lab classes, you will not be implementing algorithms from scratch, but rather building intelligent applications that use ML. In the practical assignments, we will be implementing and comparing a wide range of algorithms. To make it easy to implement the use cases we will be covering, we are recommending a particular set of software tools, but you can successfully complete the course with other tools out there.

Why Python

In this course, we are going to use the Python programming language to build several intelligent applications that use machine learning. Python is a simple scripting language that makes it easy to interact with data. Furthermore, Python has a wide range of packages that make it easy to get started and build applications, from the simplest ones to the most complex. Python is widely used in industry, and is becoming the de facto language for data science in industry.

We will also use the IPython Notebook in our videos. The IPython Notebook is a simple interactive environment for programming with Python, which makes it really easy to share your results. Think about it as a combination of a Python terminal and a wiki page. Thus, you can combine code, plots and text to explain what you did.

A prerequisite of attending this course is that you have learnt at least one programming language in the past. It is not our objective to teach you python. At this class we expect our students to be able pick up a language as they go. If you have not experienced python before it may be worth your while spending some time understanding the language. There are resources available for you to do this here (<https://docs.python.org/2/tutorial/> (<https://docs.python.org/2/tutorial/>)) that are based on the standard console. An introduction to the Jupyter Notebook (formerly known as the IPython Notebook) is available here (<http://ipython.org/ipython-doc/2/notebook/index.html> (<http://ipython.org/ipython-doc/2/notebook/index.html>)).

Lab classes vs Programming assignments/Research Project

In lab classes, the focus is on exploring each case study, without having to implement your own algorithms from scratch, and benefiting from the performance advantages that Scikit-learn provides. In programming assignments, you will be implementing many of these algorithms from scratch, having had the foundation of seeing them perform in practice on real applications.

In other words, in lab classes, we focus on exploring the use cases we'll tackle throughout the course. A huge goal is to familiarize ourselves with the core ML concepts that we will use during programming assignments and research project. In those, there will be much more implementation of ML algorithms, so the specific ML package becomes less important. However, in lab classes, we want to move quickly through all the use cases, and Scikit-learn will help us do just that.

Learning outcomes

This reading will walk you through the steps you will need to follow to install and get started with Python, IPython Notebook, Anaconda, and Scikit-learn.

Installing Python, IPython Notebook, and Pandas

Starting IPython Notebook

Writing variables, functions and loops in Python

Doing basic data manipulations in Python with Pandas' Frames

Getting started using these resources

All resources needed during lab classes are already installed in the laboratory machines. However, if you want to use them in your own computer, following we present how to do it.

Download and install Python, IPython Notebook through Anaconda: <https://www.continuum.io/downloads> (<https://www.continuum.io/downloads>)

Download and install Scikit-learn distribution: <http://scikit-learn.org/stable/> (<http://scikit-learn.org/stable/>)

Assignment question 1

Who invented Python and why? What was the language designed to do? What is the origin of the name "python"? Is the language a compiled language? Is it an object oriented language?

Question 1 answer

Write your answer to the question in this box

- Python é uma linguagem de programação, a qual foi criada por Guido van Rossum, sendo publicada pela primeira vez em 1991. Todavia, em 1989 a linguagem já havia sido concebida, no Instituto de Pesquisa Nacional para Matemática e Ciência da Computação (CWI), na Holanda.
- A motivação para criação da linguagem tinha como o objetivo prover um sucessor para linguagem ABC. Entre as novas características a linguagem seria capaz de tratar exceções, além de prover uma interface com o sistema operacional Amoeba por meio de scripts. Adicionalmente, o foco era que a linguagem tivesse características que permitissem o aumento da produtividade dos desenvolvedores.
- O nome Python foi inspirado no grupo britânico Monty Python criador do programa Monty Python's Flying Circus, embora exista um réptil de mesmo nome, o qual foi adotado como símbolo do produto. O código Python é interpretado e interativo, contudo, existe a possibilidade de compilar o código, o que resulta em aumento significativo de desempenho, ao custo da perda da interatividade.
- Python é uma linguagem de alto nível, interpretada, baseada em scripts de tipagem dinâmica e forte, além de oferecer suporte a diversos paradigmas de programação, entre eles: imperatividade, orientação a objetos e funcional.
- Fonte: <http://pyscience-brasil.wikidot.com/python:python-oq-e-pq> (<http://pyscience-brasil.wikidot.com/python:python-oq-e-pq>)
- Fonte: <https://pt.wikipedia.org/wiki/Python> (<https://pt.wikipedia.org/wiki/Python>)

Assignment question 2

Read on the internet about the following python libraries: Scikit-learn, numpy, matplotlib, scipy and pandas. What functionality does each provide python? What is the pylab library and how it is related to each other libraries?

Question 2 answer

Write your answer to the question in this box

- Scikit-learn: É um módulo Python que integra um vasto conjunto de algoritmos para resolução de problemas de aprendizado supervisionado e não supervisionado de máquina. Permitindo que leigos utilizem uma linguagem de alto nível para resolução destes problemas. Tendo como proposta a ênfase em usabilidade, desempenho, documentação e consistência da API, é distribuída com o mínimo de dependências, num formato de licenciamento simplificado BSD, sendo adequada para uso acadêmico e comercial. Todo seu código-fonte, binários e documentação podem ser obtidos em <http://scikit-learn.sourceforge.net> (<http://scikit-learn.sourceforge.net>). Fonte: <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf> (<http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>)
- Numpy: É um módulo fundamental para computação científica em Python. Entre as suas funcionalidades temos funções para: Operações de Álgebra Linear, Transformadas de Fourier, manipulação de números aleatórios, bem como, funcionalidades para integração com C/C++ e Fortran. Este módulo também pode ser utilizado de forma eficiente para a manipulação de dados multidimensionais de diversos tipos, o que confere a capacidade de integração a diversas fontes de dados. Tudo isso sobre uma licença do tipo BSB, o que permite o reuso com poucas restrições. Fonte: <http://www.numpy.org/> (<http://www.numpy.org/>)
- Matplotlib: É uma biblioteca para construção de gráficos em Python. Possui os mais diversos tipos de gráficos como: barras, tortas, pontos, superfícies, etc... Seu licenciamento é baseado em PSF (Python Software Foundation). Fonte: <https://matplotlib.org/> (<https://matplotlib.org/>)
- Scipy: É uma biblioteca para computação científica que permite integração numérica e otimização. Além disso, Scipy também é o nome de um ecossistema para computação científica, composto de diversas bibliotecas entre elas: Numpy, Pandas, Matplotlib, IPython, Sympy e Scipy.
- Pandas: É uma biblioteca para manipulação de estrutura de dados e análise. Sua proposta é fornecer uma interface mais amigável. Seus principais tipos de dados são: series e dataframe.
- O Pylab é um módulo do pacote Matplotlib que permite gerar gráficos de duas dimensões, com edição interativa, animações. Além disso, este módulos possui diversos tipos de gráficos, suporte a LaTeX, além de suporte a diversas extensões de imagem. Hoje, não é recomendável importar o módulo diretamente, mas sim a Matplotlib.
- Fonte: <https://www.scipy.org/> (<https://www.scipy.org/>)

Assignment question 3

What is Jupyter Notebook and why was it inveted? Give some examples of functionality it gives over standard python. What is the jupyter project? Name two languages involved in the Jupyter project other than python.

Question 3 answer

Write your answer to the question in this box

- O Jupyter Notebook é uma aplicação web de código aberto, a qual permite a criação e compartilhamento de documentos, os quais podem conter código interativo, equações, visualizações, além de texto explicativo. Suas aplicações incluem a manipulação e transformação de dados, simulação numérica, modelagem estatística, aprendizado de máquina, entre outros.
- A história deste projeto começa em 2001, quando Fernando Pérez iniciou o desenvolvimento do IPython. Quatro anos depois, já com o apoio de Robert Kern eles tentaram construir o sistema notebook. Contudo, o protótipo nunca foi completamente funcional. Apesar disso, como o IPython continuou a ser desenvolvido, em 2014 eles, finalmente, conseguiram lançar o projeto Jupyter utilizando o IPython como núcleo. Atualmente, por se tratar de uma iniciativa de código aberto, o projeto, hoje recebe o nome de Jupyter Lab.
- O Python padrão produz como resultado aplicações desktop, as quais são interpretadas e tem seu resultado, apresentado na console. Com o uso de algumas bibliotecas como o Django ou Flask, por exemplo, é até possível criar páginas web, contudo, estas páginas ainda não são interativas. Ou seja, nativamente, não permitem que um código seja escrito e interpretado na web de forma dinâmica e iterativa, o que é obtido com o Jupyter Notebook App.
- Além do núcleo em Python, foram utilizadas AJAX e HTML no Projeto Jupyter.

Download the data and sample code and familiarize yourself with the notebooks

Before doing the assignments in this course, familiarize yourself with the two following notebooks:

Download the notebook that covers getting started with Python: Getting started with iPython Notebook

Download the notebook that covers getting started with Pandas: Getting Started with Pandas

Download the simple people dataset: people-example.csv

Save all these files in the same directory. If you are not sure where to save the files? See this guide:

Assignment question 4

Describe the following functions of Pandas Frames:

head()

tail()

show()

apply()

Question 4 answer

Write your answer to the question in this box

- `head()`: Retorna as primeiras `n` linhas da estrutura, sendo que o valor padrão do argumento `n` é 5.
- `tail()`: Retorna as últimas `n` linhas da estrutura, sendo que o valor padrão do argumento `n` é 5.
- `show()`: Não foi possível encontrar referências a este método. A suspeita é que tenha sido descontinuado.
- `apply()`: Aplica uma função sobre uma coluna ou linha da estrutura de dados. Sintaxe:
`DataFrame.apply(func, axis=0, broadcast=False, raw=False, reduce=None, args=(), **kwds)`
- Fonte: <https://pandas.pydata.org/pandas-docs/stable/api.html> (<https://pandas.pydata.org/pandas-docs/stable/api.html>)

Congratulations!!! Now you are ready to get started! From here, you will be ready to do all the assignments in the course, and build awesome intelligent applications that use machine learning!

In [1]:

In []: