Lab 0

August 29, 2017

1 Laboratory class 0

Nomes:
1 - Marco Antonio Santo
2 - Lucas Teodoro

2 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 pratical 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.

2.1 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. Futhermore, 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 undertanding the language. There are resources available for you to do this here (https://docs.python.org/2/tutorial/) that are based on the standard console. An introduction to the Jupyter Notebook (formely known as the IPython Notebook) is available here (http://ipython.org/ipython-doc/2/notebook/index.html).

2.2 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 assingments, 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.

2.3 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

2.4 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

Download and install Scikit-learn distribution: http://scikit-learn.org/stable/

2.5 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 is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991. With the porpose of handling and interfacing with the operating system Amoeba.
- The Python name comes from Monty Python, a humoristic group from british.

- Python is a multi-paradigm programming language: object-oriented programming and structured programming are fully supported.

2.6 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: Simple and efficient tools for data mining and data analysis.
- Numpy: NumPy is the fundamental package for scientific computing with Python. 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.
- Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
- Scipy: SciPy is a Python-based ecosystem of open-source software for mathematics, science, and engineering.
- Pandas: 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.
- PyLab: To make PyLab an easy to use, well packaged, well integrated, and well documented, numeric computation environment so compelling that instead of having people go to Python and discovering that it is suitable for numeric computation, they will find PyLab first and then fall in love with Python. PyLab was designed to help the begginers in python to use most of your funcionalities.

2.7 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

- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more. Notebooks can be shared with others using email, Dropbox, GitHub and the Jupyter Notebook Viewer. Leverage big data tools and code can produce rich output such as images, videos, LaTeX, and JavaScript. Interactive widgets can be used to manipulate and visualize data in realtime.
- The Notebook has support for over 40 programming languages, including those popular in Data Science such as Python, R, Julia and Scala.

2.8 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 sabe the files? See this guide:

2.9 Assignment question 4

Descrive the following functions of SFrames:

head()
tail()
show()
appy()

Question 4 answer Write your answer to the question in this box

- head(): The first n rows of the SFrame, been n the number of rows to fetch.
- tail(): The last n rows of the SFrame, been n the number of rows to fetch.
- show(): Visualize the SFrame with GraphLab Create canvas. This function starts Canvas if it is not already running. If the SFrame has already been plotted, this function will update the plot.
- apply(): Apply to each row selected commands.

2.9.1 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 intellingent applications that use machine learning!