Tools and Setup

Alireza Samar

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

MLDS
MACHINE LEARNING
FOR DATA SCIENCE

Anaconda

Anaconda puts nearly all of the tools that we're going to need into a neat little package: the Python core language, an improved REPL environment called `Jupyter`, numeric computing libraries (`NumPy`, `pandas`), plotting libraries (`seaborn`, `matplotlib`), and statistics and machine learning libraries (`SciPy`, `scikit-learn`, `statsmodels`). We'll use Anaconda's installer to handle setting up the environment that we'll work in.

Install Anaconda

In order to keep the size of the download small, we actually use a minimum set of packages called **Miniconda**.

Miniconda installer packages:

- Windows
- macOS

Once this downloads, you can follow the instructions for installing on your operating system: at this link.

Note: It's easiest just to use Anaconda's defaults in the installer. You don't have to change anything unless you're sure you want something different!

MLDS

Common packages for data science in Python

Click the link below to download an environment file. This file contains a list of common packages and libraries for doing data science in Python. Remember where you save the file `environment.yml`. You'll need that path shortly. You don't need to open that file right now.

- Windows
- macOS

Once the download finishes, open the command line by doing the following:

- Windows - Hit "Start" and then type "Command Prompt" and use that terminal.
- macOS - Type `Cmd+Space` and then enter `Terminal` in the search box to open the terminal.

MLDS

You’re almost there ...

Run the following commands, which will install the package and put you in the tutorial environment.

- `conda env create -f <PATH_TO_ENVIRONMENT.YML>` - You'll need to replace `<PATH_TO_ENVIRONMENT.YML>` with the actual path where the file was downloaded. For macOS, that's often (`/Users/<USERNAME>/Downloads/environment.yml`). For Windows, it is usually `C:/Users/<USERNAME>/Downloads/environment.yml`. You'll have to replace `<USERNAME>` with your username on your machine.

That will download all a set of packages that are commonly used for data science in Python. When it finishes, you can activate the environment with the following command:

- Windows - `activate tutorial`
- macOS - `source activate tutorial`

MLDS

Run Jupyter notebook!

In this step, we'll make sure everything is working by running the Jupyter Notebook. Jupyter Notebook is a tool for doing interactive data science work in your browser. * In your command prompt **with the tutorial environment activated** (Note: you'll be able to tell because your command prompt will say `(tutorial)` at the start of it.) * Type the following command `jupyter notebook .` * A browser window will open, showing the Jupyer environment. By default, you will be in a file browser view. * In the file browser, find where you have a Jupyter notebook. You can download this fun Jupyter notebook and then open it in the file browser. * Click on one of the notebook (*.ipynb) files to get started!

MLDS

Important to know!

To stop Jupyter notebook:

- Hit `Ctrl+c` to stop the Jupyter notebook server running on your machine. (Make sure to use `Ctrl+s` in the notebook to save it first!)

To leave the tutorial environment (with all our fun packages) and go back to your normal environment:

- Windows - `deactivate tutorial`
- macOS - `source deactivate tutorial`

MLDS

Thanks!

Machine Learning for Data Science Interest Group
Advanced Informatics School
Universiti Teknologi Malaysia

@utmmlds
ais.utm.my/mlds

February 2017