

Machine Learning and Data Mining

April, 2021

Outline

- Introduction and motivation (Machine Learning and Data Mining)
- Introduction to both fundamental programming concepts and the Python programming language (the programming language R)
- The Principal Components Analysis (PCA)
- The Limit Order Book (LOB)
- The Long short-term memory (LSTM) neural network

What is data mining?

Data mining

- the process of analyzing the data sets using machine learning and statistics **to find insights and to detect pattern in the observed data sets**
- explosive data growth, we are able to store much more data than before \implies automation of the massive data sets analysis
- “We are drowning in information but starved for knowledge.” - John Naisbitt
- The main job of data mining is to extract and pick out the potentially useful, understandable and hidden information, previously unknown information/knowledge from the data set.
- Alternative names: data analysis, business intelligent, knowledge extraction, information harvesting, etc.
- the process of analyzing data sets from different prospective and summarizing the relevant knowledge

The steps in data mining

The steps in data mining

- Data cleaning (noise, outliers, missing values, duplicate data)
- Data integration
- Data selection
- Data preprocessing/ Data transformation
- Data mining: select the mining approach, choose the mining algorithm
- Pattern evaluation
- Knowledge presentation and visualization
- Use of discovered information

Basic Data Mining Tasks

- Classification: Assign each data element into the one of the predefined classes (e.g. Spam e-mail detection, etc.)
- Regression: map item into the real valued prediction variable (e.g. predict a value of savings in future)
- Clustering: similar to classification but the groups are not predefined (e.g. group people into communities from a given social network)

Installation

- Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, etc.)
- <https://www.anaconda.com/>
- The Jupyter Notebook is an open-source web application that lets you easily write and iterate Python code for data analysis.
- <https://jupyter.org/>
- PyCharm
- <https://www.jetbrains.com/pycharm/>

Installation

- Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, etc.)
- <https://www.anaconda.com/>
- The Jupyter Notebook is an open-source web application that lets you easily write and iterate Python code for data analysis.
- <https://jupyter.org/>
- PyCharm
- <https://www.jetbrains.com/pycharm/>

Installation

- <https://www.python.org/downloads/>
- `pip install numpy`

If you use `pip`, you can install it with:

```
pip install notebook
```

Congratulations, you have installed Jupyter Notebook! To run the notebook, run the following command at the Terminal (Mac/Linux) or Command Prompt (Windows):

```
jupyter notebook
```

Figure 1: `pip install notebook`

Installation

<https://www.anaconda.com/distribution/>

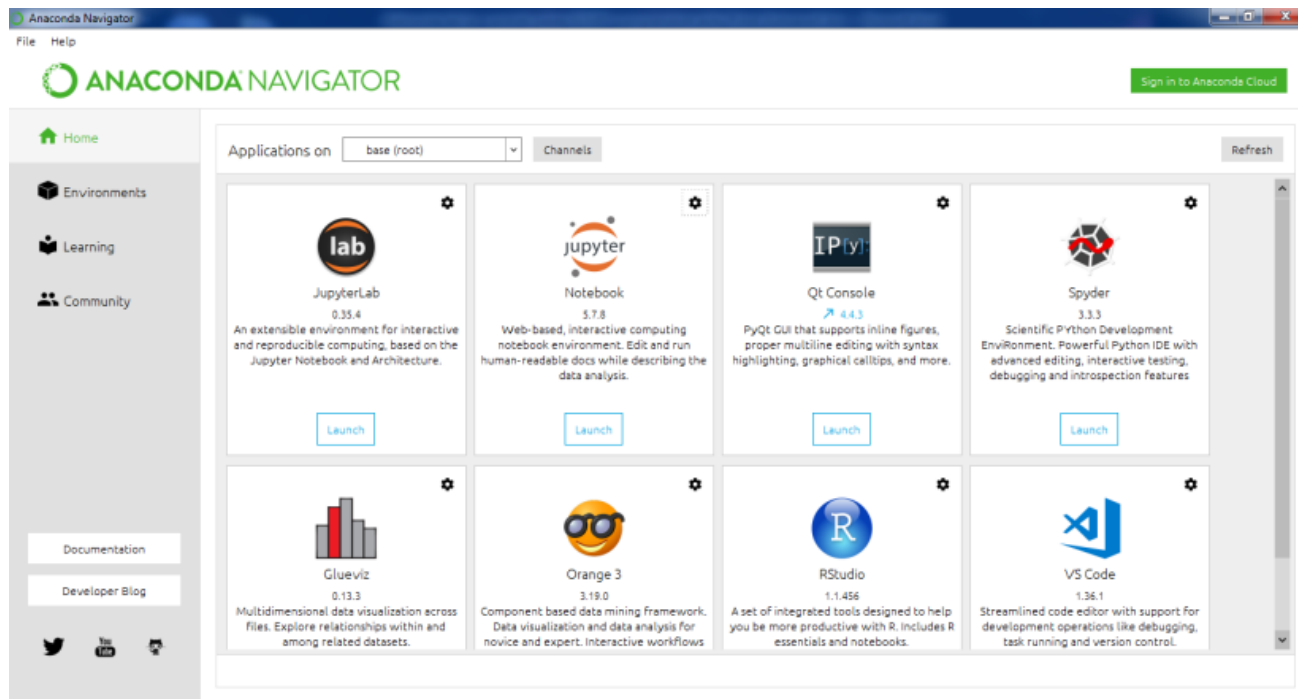


Figure 2: Programs within Anaconda Navigator

Jupyter Notebook

The screenshot shows a web browser window with the address bar at `localhost:8889/tree`. The Jupyter logo and navigation buttons (Quit, Logout) are at the top. Below the tabs (Files, Running, Clusters), there is a section for file management with buttons for 'Upload', 'New', and a refresh icon. A table lists the files and folders in the current directory, sorted by 'Last Modified'.

| | Name | Last Modified | File size |
|--------------------------|-------------|---------------|-----------|
| <input type="checkbox"/> | / | | |
| <input type="checkbox"/> | Anaconda3 | 2 months ago | |
| <input type="checkbox"/> | Contacts | 2 months ago | |
| <input type="checkbox"/> | Desktop | 3 minutes ago | |
| <input type="checkbox"/> | Documents | 2 months ago | |
| <input type="checkbox"/> | Downloads | a day ago | |
| <input type="checkbox"/> | Favorites | 2 months ago | |
| <input type="checkbox"/> | Links | 2 months ago | |
| <input type="checkbox"/> | Music | 2 months ago | |
| <input type="checkbox"/> | New folder | 4 years ago | |
| <input type="checkbox"/> | Pictures | 2 months ago | |
| <input type="checkbox"/> | pip | 6 months ago | |
| <input type="checkbox"/> | Saved Games | 2 months ago | |
| <input type="checkbox"/> | Searches | 2 months ago | |
| <input type="checkbox"/> | source | a year ago | |

Jupyter Notebook

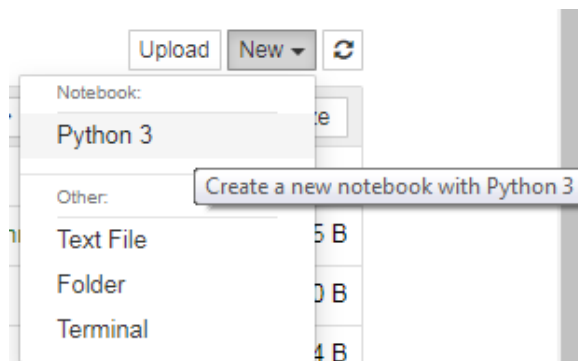


Figure 3: Creating new Jupyter notebook script

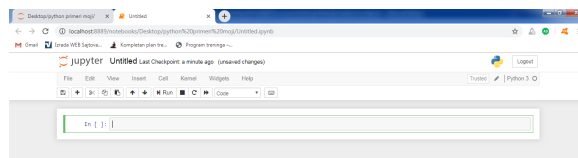


Figure 4: A blank Jupyter notebook script

Literature

- Notes from classes
- C. Bishop: Pattern Recognition and Machine Learning
- K. Murphey: Machine Learning: A Probabilistic Perspective
- S. Shalev-Schwartz, S. Ben-David: Understanding Machine Learning: From Theory to Algorithms