UCLA Samueli
Computer Science

CS145: Introduction to Data Mining (Spring 2024)

Discussion 1: **Python Tutorial**

Instructor: Dr. Ziniu Hu
Teaching Assistant: Yanqiao Zhu

The Scalable Analytics Institute (ScAI)
Department of Computer Science
University of California, Los Angeles (UCLA)

Installing Python

- We STRONGLY recommend the anaconda environment
- https://www.anaconda.com/distribution

Jupyter notebooks

- You can install with pip or use anaconda:
  - http://jupyter.readthedocs.io/en/latest/install.html
  - It comes with Anaconda
- Used in the homework assignments for clarity but horrible for fast development cycles
- To start it: open Anaconda Prompt or your terminal and type jupyter notebook, or find the shortcut in your start menu

Jupyter notebooks

- Google Colab Notebooks
  - Modify your notebook online
  - Download in .ipynb format
  - Excellent for writing code incrementally and testing as you go
  - https://colab.research.google.com/

Packages

- You will need
  - numpy
  - seaborn and matplotlib
  - scikit-learn
- If you have Anaconda, you have all of these already
- If you need additional packages
  - conda config --env --add channels conda-forge
  - conda install <package_name>
- Or you can use pip:
  - pip install <package_name>

Packages

numpy

- Used for numerical computing/matrix operations
- Your data is going to be in a matrix, so manipulate it with numpy
- Python numpy Tutorial: https://cs231n.github.io/python-numpy-tutorial/

scikit-learn

- Used for basic ML algorithms, tools and techniques
- No integration of neural nets
- User guide: https://scikit-learn.org/stable/user_guide.html

The supervised learning recipe

- Get training data
- Pick a model class
- Pick a loss function
- Pick a learning objective to optimize

Debugging tips

- Print it
- Google it
- Try using dummy data
- Ask (ChatGPT) for help!
- Take a walk
- Take a nap