Introduction to Scikit-Learn: Machine Learning with Python

Preliminaries

Tony Yao-Jen Kuo

Goals of this Tutorial

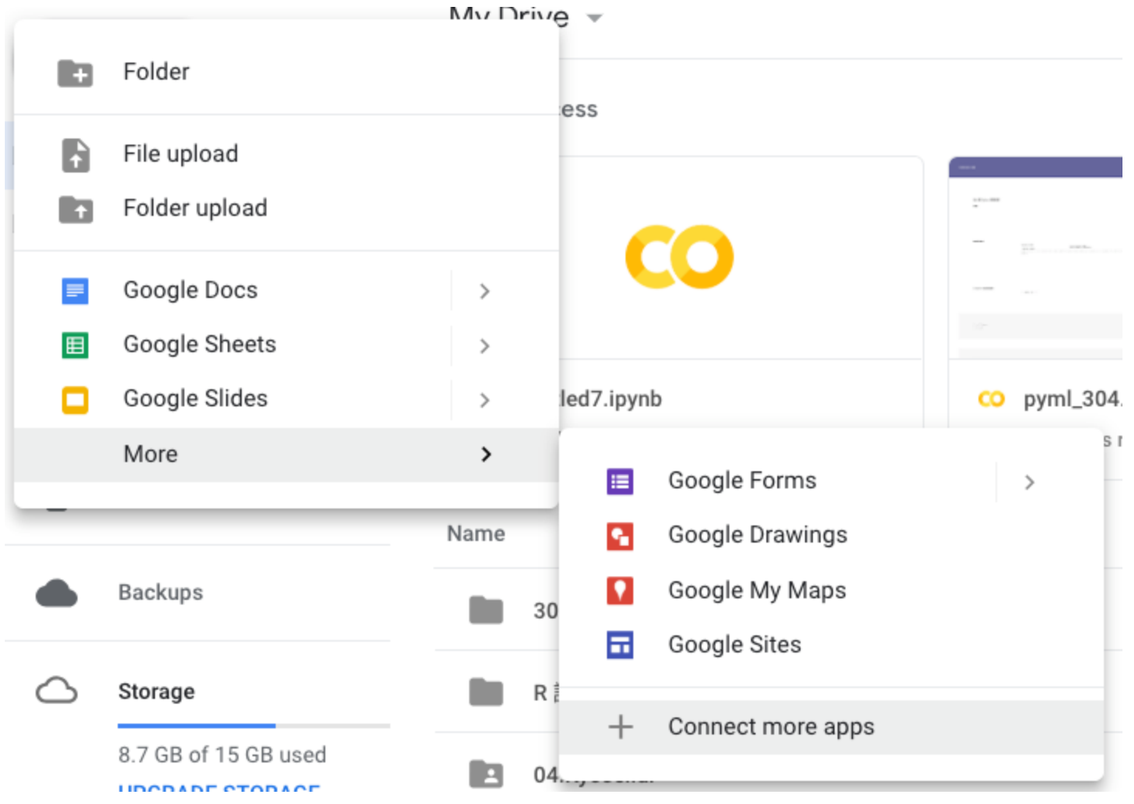
- **Introduce the basics of Machine Learning**, and some skills useful in practice.
- **Introduce the syntax of scikit-learn**, so that you can make use of the rich toolset available.

Setup & introduction

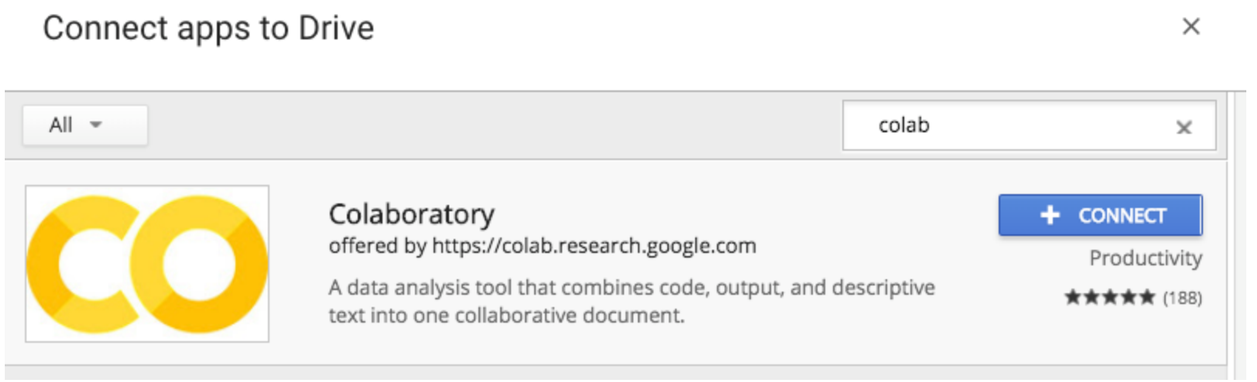
- Making sure you are familiar with Google Colab (https://colab.research.google.com)

Open Google Cloud Drive

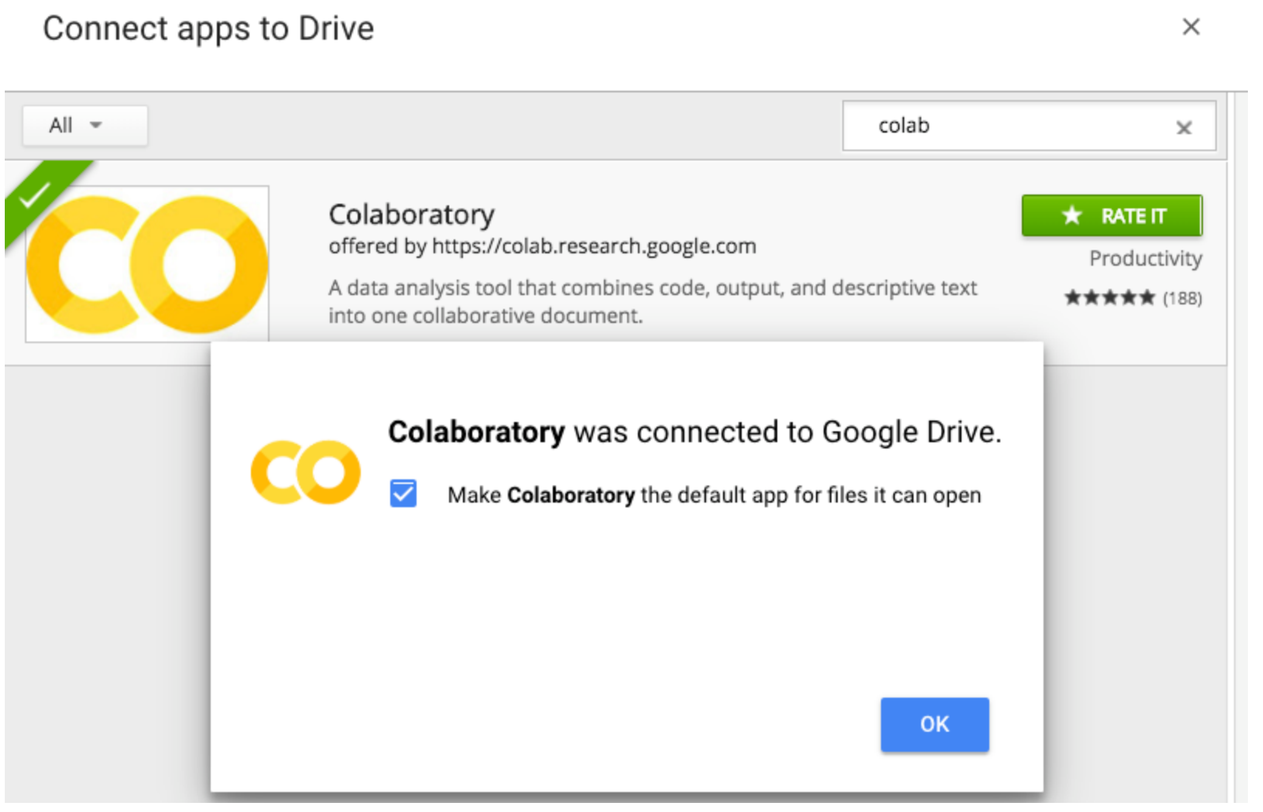
New, More, Connect More Apps



Search Colaboratory



Connect Colaboratory



Basic Principles of Machine Learning and the Scikit-learn Interface

- Machine Learning Intro
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Dimensionality Reduction and Clustering
- Validation and Model Selection

Classification

- k Nearest Neighbors
- Support Vector Machines
- Decision Trees and Random Forests

Regression

- Linear Regression
- Polynomials
- Random Forest Regressor

Unsupervised learning

- Principal Component Analysis
- K-means Clustering

Validation and Cross-validation

Preliminaries

This tutorial requires the following skills

- A Gmail account, since we are using Google Colab
- Python programming
- Understanding of `numpy` and matrix algebra
- Understanding of `scipy`
- Understanding of `pandas` and its core `Series` and `DataFrame`
- Understanding of `matplotlib.pyplot`

Checking your installation

You can run the following code to check the versions of the packages on Google Colab.

In [1]:

```
import numpy
print('numpy:', numpy.__version__)

import scipy
print('scipy:', scipy.__version__)

import matplotlib
print('matplotlib:', matplotlib.__version__)

import sklearn
print('scikit-learn:', sklearn.__version__)
```

```
numpy: 1.12.1
scipy: 1.1.0
matplotlib: 2.0.2
scikit-learn: 0.19.1
```

Useful Resources

- scikit-learn: http://scikit-learn.org (http://scikit-learn.org)
- matplotlib: http://matplotlib.org (http://matplotlib.org) (especially the gallery section)
- Jupyter: http://jupyter.org (http://jupyter.org)

References

- Python Data Science Handbook (https://www.amazon.com/Python-Data-Science-Handbook-Essential/dp/1491912057)
- Jake Vanderplas (http://www.vanderplas.com)
- Source and license info is on GitHub (https://github.com/jakevdp/sklearn_tutorial/).