Introduction to scikit-learn: Machine Learning in Python

By Thomas J. Fan

[Link to slides](#)

Scikit-learn is a Python machine learning library used by data science practitioners from many disciplines. We will start this training by learning about scikit-learn's API for supervised machine learning. scikit-learn's API mainly consists of three methods: fit to build models, predict to make predictions from models, and transform to modify data. This consistent and straightforward interface abstracts away the underlying algorithm, thus enabling us to focus on our particular problems. We will learn the importance of splitting your data into train and test sets for model evaluation. Next, we will learn about combining preprocessing techniques with machine learning models using scikit-learn's Pipeline. The Pipeline allows us to connect transformers with a classifier or regressor to build a data flow where the output of one layer is the input of another. Finally, we will look at the Pandas output API recently introduced in version 1.2. After this training, you will have the foundations to apply scikit-learn to your machine learning problems.

Obtaining the Material

With git

The most convenient way to download the material is with git:

```
git clone https://github.com/thomasjpfan/ml-workshop-intro-v2
```

Please note that I may add and improve the material until shortly before the session. You can update your copy by running:

```
cd ml-workshop-intro-v2
git pull origin main
```

Download zip

If you are not familiar with git, you can download this repository as a zip file at:
github.com/thomasjpfan/ml-workshop-intro-v2/archive/main.zip. Please note that I may add and improve the material until shortly before the session. To update your copy please re-download the material a day

before the session.

Running the notebooks

Local Installation

Local installation requires conda to be installed on your machine. The simplest way to install conda is to install miniconda by using an installer for your operating system provided at docs.conda.io/en/latest/miniconda.html. After conda is installed, navigate to this repository on your local machine:

```
cd ml-workshop-intro-v2
```

Then download and install the dependencies:

```
conda env create -f environment.yml
```

This will create a virtual environment named ml-workshop-intro-v2. To activate this environment:

```
conda activate ml-workshop-v2
```

Finally, to start jupyterlab run:

```
jupyter lab
```

This should open a browser window with the jupterlab interface.

Run with Google's Colab

If you have any issues with installing conda or running jupyter on your local computer, then you can run the notebooks on Google's Colab:

1. [Loading data with Scikit-learn](Loading%20data%20with%20Scikit-learn)
2. [Supervised learning with scikit-learn](Supervised%20learning%20with%20scikit-learn)
3. Preprocessing
4. Pipelines
5. [Pandas output](Pandas%20output)

License

This repo is under the MIT License.