Introduction to Data Science

(Lecture 8)

Dr. Mohammad Pourhomayoun

Assistant Professor

Computer Science Department

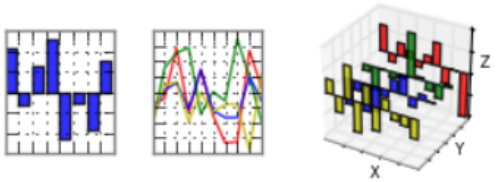
California State University, Los Angeles

Data Science with Python

IP[y]: IPython
Interactive Computing

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



scikits learn
machine learning in Python

NumPy

SciPy.org Sponsored By ENTHOUGHT

matplotlib

Scikit-Learn:
A Library for Data Science and Machine Learning

Scikit-Learn (sklearn)

- Scikit-learn (aka sklearn) is the Python Machine Learning Library.
- It includes optimal implementation of various **classification, regression** and **clustering** algorithms.
- It also includes hundreds of commands and functions for data preprocessing and processing along with a number of **default datasets** to work with.
- It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- Scikit-learn has an exceptional documentation.

Important Hint about sklearn

- Sklearn only accept **NUMERICAL features**. Thus, we have to convert non-numerical (categorical) features into numerical values.
- **Note**: In converting features (and sometimes labels), we have to be cautious to avoid defining a confusing "ordering" between categorical values (we will talk about it later in this course).
- Depending on the classification algorithm, We usually use **LabelEncoding** to convert labels, and **OneHotCoding** to convert features.

5 Steps To Make Prediction In sklearn

- **Step1:** Importing the sklearn class (the machine learning algorithm) that you would like to use for prediction FROM sklearn library.
- **Step2:** Set up the Feature Matrix and Label Vector.
- **Step3:** Defining (instantiating) an "object" (instance) of the sklearn class as an initial predictive object.
- **Step4:** Training Stage: Train the above predictive model using the training dataset.
- **Step5:** Testing (Prediction) Stage: Making prediction on new observations (Testing Data) using the trained model.
- **Step6:** Evaluating the machine leanring model and results

Data Science Practical Tutorial

- Let's open file ***CS4661-PythonDataScienceTutorial-Lab3.ipynb*** in Jupyter notebook to continue the tutorial.