Python for Machine Learning

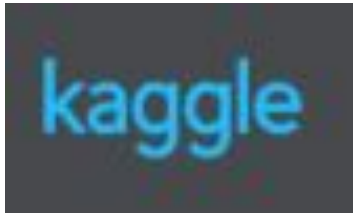
Connor Watson

Outline

I. Data Sets
   A. Kaggle
II. Anaconda
   A. numpy
   B. pandas
   C. sklearn
III. Machine Learning
   A. scikit-learn

Kaggle

- Website with data sets / projects
- Competitions
  - Earn $$$$
- View other peoples' work
- Beginner's project:
  - https://www.kaggle.com/c/titanic
  - Predict which passengers survived or not



Anaconda

- Python for Data Science
  - numpy, pandas
  - scikit-learn
  - matplotlib, seaborn
  - and more!!
- Includes R as well

Jupyter Notebooks

- Interactive environment to run code
- Execute in groups or line by line
- Show plots on screen

numpy

- Save coding time
  - Less loops - apply operation to all items
- Faster execution
  - Single type - avoid type checking
- Less memory
  - List: array of pointers (4B+) to Python objects (16B+)
  - Array: Itemsize same throughout

Arrays vs List

Python List

Memory Address	Pointer to Data Location
0050h	1204h
0054h	F1A3h
0058h	2B87h

Memory Address	Data
1204h	5.75
2B87h	1.94
F1A3h	38.6

NumPy Array

Memory Address	Data
0050h	5.75
0054h	38.6
0058h	1.94

Metadata

- Type - shared data type
- Size - memory size of each item
- Shape - array dimensions
- Data - access through indexing

Data Type	Description
bool_	Boolean (True or False) stored as a byte
int8	Byte (-128 to 127)
int16	Integer (-32768 to 32767)
int32	Integer (-2.15E-9 to 2.15E+9)
int64	Integer (-9.22E-18 to 9.22E+18)
uint8	Unsigned integer (0 to 255)
uint16	Unsigned integer (0 to 65535)
uint32	Unsigned integer (0 to 4.29E+9)
uint64	Unsigned integer (0 to 1.84E+19)
float16	Half precision signed float
float32	Single precision signed float
float64	Double precision signed float
complex64	Complex number: two 32-bit floats (real and imaginary components)
complex128	Complex number: two 64-bit floats (real and imaginary components)

numpy vs pandas

numpy

- Low-level data structure (array)
- Multiple dimensional arrays / matrices
- Wide range of math operations

pandas

- High-level data structure (dataframe)
- Better for tabular data / time-series data
- Data-alignment
- Replace/ignore missing data
- SQL like operations
- Comprised of numpy/scipy

pandas

- Structured Data - Table based information (DataTable)
  - Rows - observations
  - Columns - features / attributes

	First	Last	Age
0	Jane	Doe	23
1	Terell	Smith	24
2	Elizabeth	Chen	22
3	Nishant	Patel	25

Creating Data

- Create from scratch
- Read from a file:
  - TXT
  - CSV
  - HD5
  - Excel
  - JSON

Processing Data

- Easily query rows based on a condition
  - Find missing data
  - Use outlier detection
  - Change multiple rows at once

scikit learn

- Several objects for machine learning models
- A **model** can be **trained** to make **predictions**
  - Create and **fit** the model on **train** data
  - **Predict** on the **test** data and check for accuracy
  - **Optimize** parameters to get **better accuracy**
  - Predict on **new** data points

Models

- We use mathematical models to learn about data and make predictions
- Some need to know about the data
  - classification
  - regression
- Some can learn patterns without any prior knowledge
  - clustering

How Does it Work?

- Many sklearn models follow use the following methods:
  - init - initialize the model with some defined parameters
  - fit() - learn patterns about the train data
  - predict() - make predictions on new data using the learned patterns
  - score() - check the accuracy of the model

Documentation

- https://docs.scipy.org/doc/
- https://pandas.pydata.org/pandas-docs/stable/
- https://scikit-learn.org/stable/documentation.html

NJIT Data Science

https://join.slack.com/t/njitdatascienceclub/shared_invite/enQtNzEzMzc4Mjk1ODQ2LWI1ZWQ5NjcyZjJlYTBhM2EyYmY2ODQzZjQ3MmM0NzFhNWY3YTYzNzMyYzYwNDc5ZGNjYmIyOGY3NWVjMGQ1OTc

- Workshops (like this one)
- Mini lectures
- Industry sponsors

Expand to your Projects?