

Introduction to Python

Basics of Data Science

Day 5 | Duration: 2 Hours

Topics

- Numpy
- Train-Test learning

NumPy

NumPy

- NumPy is the fundamental library for data processing and data science
 - A multi-dimensional array type not in vanilla Python
 - Functions for matrices and linear algebra

ndarray

- The core objects used by NumPy are ndarrays
 - True multi-dimensional arrays – located in a single, contiguous block of memory.
 - Designed to be used with a single type.

Train-Test

Train-Test

- Train-Test is a machine learning technique where we create a model of part of our data and then test that model against another part of our data.
- This provides a check against our model fitting the idiosyncrasies of our particular data rather than the actual problem.

Train-Test Process

1. Divide your data set into a training and testing set.
 1. This is typically an 80-20 split
2. Perform whatever ML training you're doing on just the training data
 1. This could be as simple as fitting it to a regression or as complex as running an evolutionary algorithm against it for days.

Train-Test process

3. Once your model fits your training data sufficiently (r-squared), correlate the same model to the test data.
4. If the model also correlates well to the test data, there's a good chance it's a predictive model.