“Advanced Python Programming for Everybody”

Instructor: Ernest Bonat, Ph.D.
Senior Software Engineer
Senior Data Scientist
ebonat@15itresources.com
Cell: 503.730.4556

Module 6 Source Code

https://github.com/ebonat/intel_module_6

Module 6. “Python Data Ecosystem for Data Science Projects – Part 2”

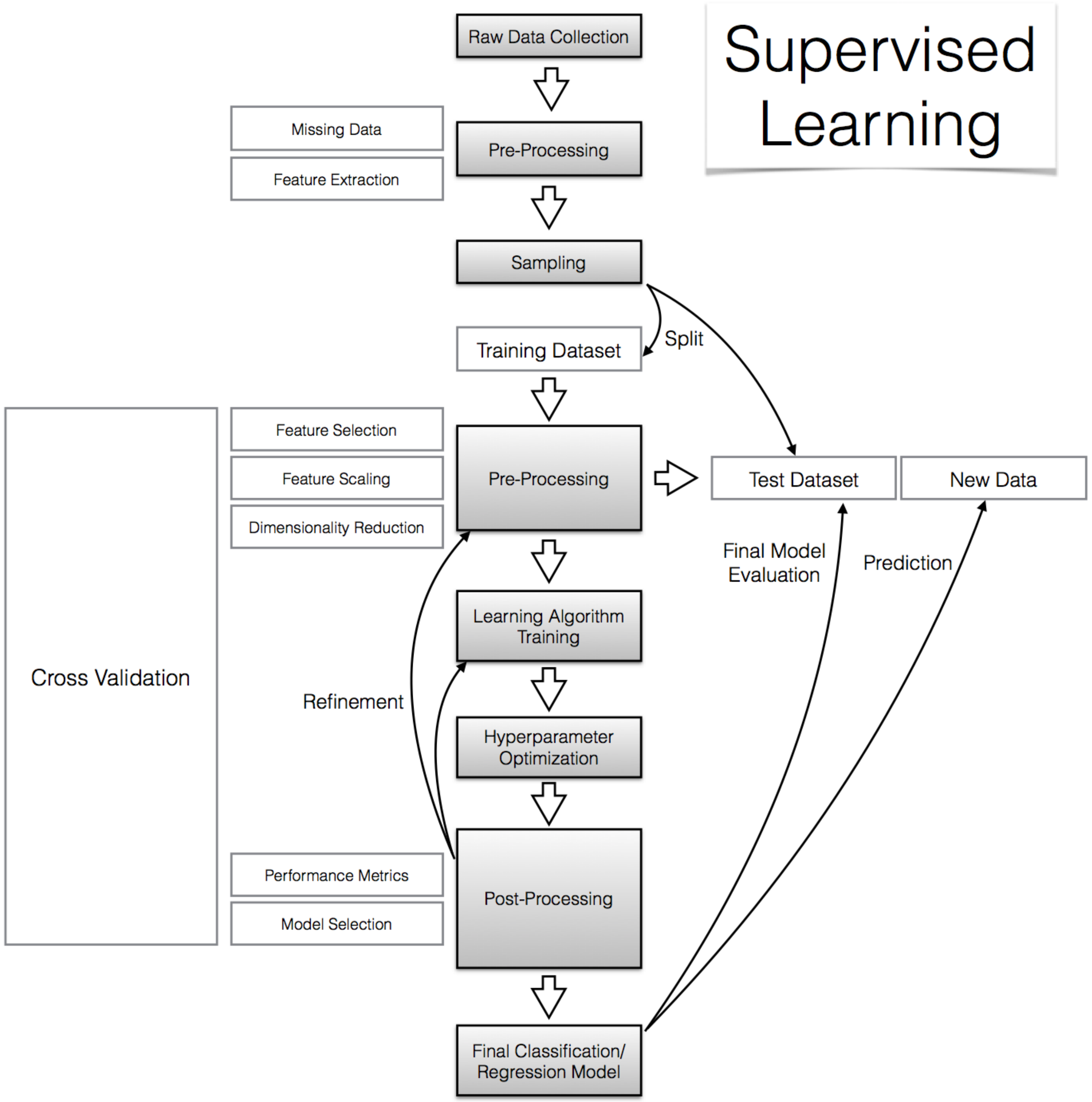
What do you really need to know to become a Data Scientist?

- **Probability and Statistics** (undergraduate level)
- **Python Programming Language** (good level!)
- **Python Data Ecosystem** (good level!):

1. **NumPy** – fundamental package for scientific computing (Numerical Python - http://www.numpy.org/)
2. **pandas** – provides easy-to-use and high-performance data structures (https://pandas.pydata.org/)
3. **SciPy** - Python-based ecosystem of open-source software for mathematics, science, and engineering (https://www.scipy.org/)

4. **scikit-learn Machine Learning** – a simple and efficient tool for data mining and data analysis (http://scikit-learn.org/)
5. **matplotlib** – a 2D plotting library which produces publication quality figures in a variety of hard copy formats and interactive environments across platforms (https://matplotlib.org/)
6. **seaborn** - statistical data visualization (https://seaborn.pydata.org/)
7. **scikit-image** – a collection of algorithms for image processing (http://scikit-image.org/)

Supervised Learning



Raw Data Collection

Missing Data

Feature Extraction

Pre-Processing

Sampling

Split

Training Dataset

Cross Validation

Feature Selection

Feature Scaling

Dimensionality Reduction

Pre-Processing

Test Dataset

New Data

Final Model Evaluation

Prediction

Learning Algorithm Training

Refinement

Hyperparameter Optimization

Performance Metrics

Model Selection

Post-Processing

Final Classification/ Regression Model

Sebastian Raschka 2014

This work is licensed under a Creative Commons Attribution 4.0 International License.

Machine Learning (ML) - at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.

Types of ML algorithms:

1. **Supervised Learning** (most popular today!)

Supervised learning can be explained as follows: use labeled training data to learn the mapping function from the input variables (X) to the output variable (Y).

Two types:

1. **Classification**: To predict the outcome of a given sample where the output variable is in the form of categories. Examples include labels such as male and female, sick and healthy

2. **Regression**: To predict the outcome of a given sample where the output variable is in the form of real values. Examples include real-valued labels denoting the amount of rainfall, the height of a person.

Popular Algorithms: **Linear Regression, Logistic Regression, Decision Trees, Random Forest, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, XGBoost, Artificial Neuro Networks (ANN = Deep Learning)**

2. **Unsupervised Learning**

Unsupervised learning problems possess only the input variables (X) but no corresponding output variables. It uses unlabeled training data to model the underlying structure of the data.

1. **Association**: To discover the probability of the co-occurrence of items in a collection. It is extensively used in market-basket analysis. Example: If a customer purchases bread, he is 80% likely to also purchase eggs.

2. **Clustering**: To group samples such that objects within the same cluster are more similar to each other than to the objects from another cluster.

Popular Algorithms: **K-Means Clustering, Principal Component Analysis (PCA), etc.**

3. **Reinforcement Learning**

Reinforcement learning is a type of machine learning algorithm that allows the agent to decide the best next action based on its current state, by learning behaviors that will maximize the reward.

Popular Algorithms: **Markov decision processes, Q-Learning, RL, Monte Carlo Simulation, etc.**

Best Supervised Learning algorithms to start to:

1. **Random Forest** (RF) - https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd,
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

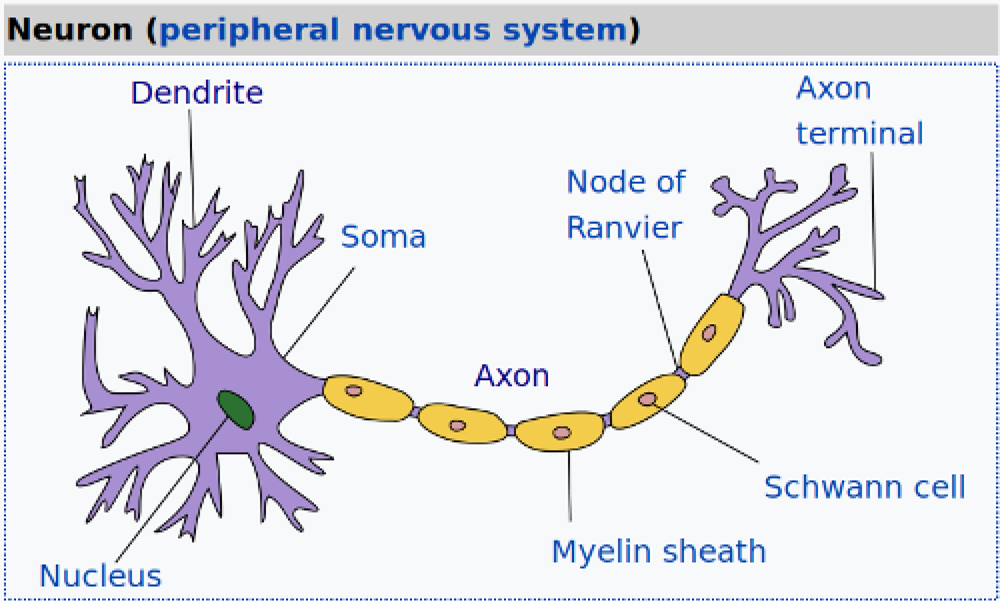
2. **eXtreme Gradient Boosting** (XGBoost) - https://xgboost.readthedocs.io/en/latest/. Winner of Kaggle competitions (https://www.kaggle.com/)

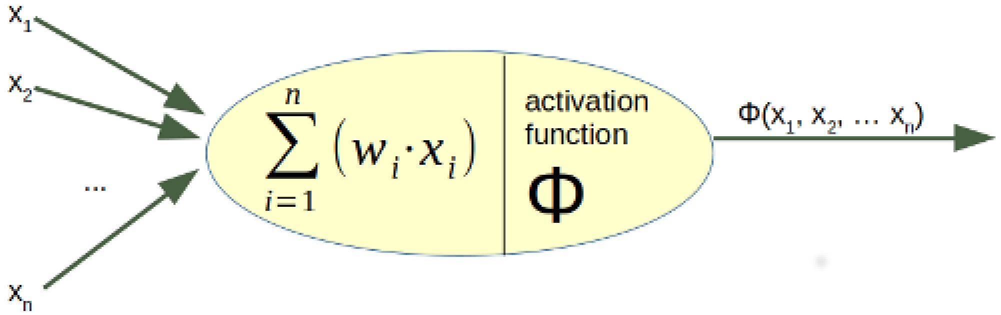
3. **Artificial Neural Networks** (ANN) - Multi-layer Perceptron - http://scikit-learn.org/stable/modules/neural_networks_supervised.html

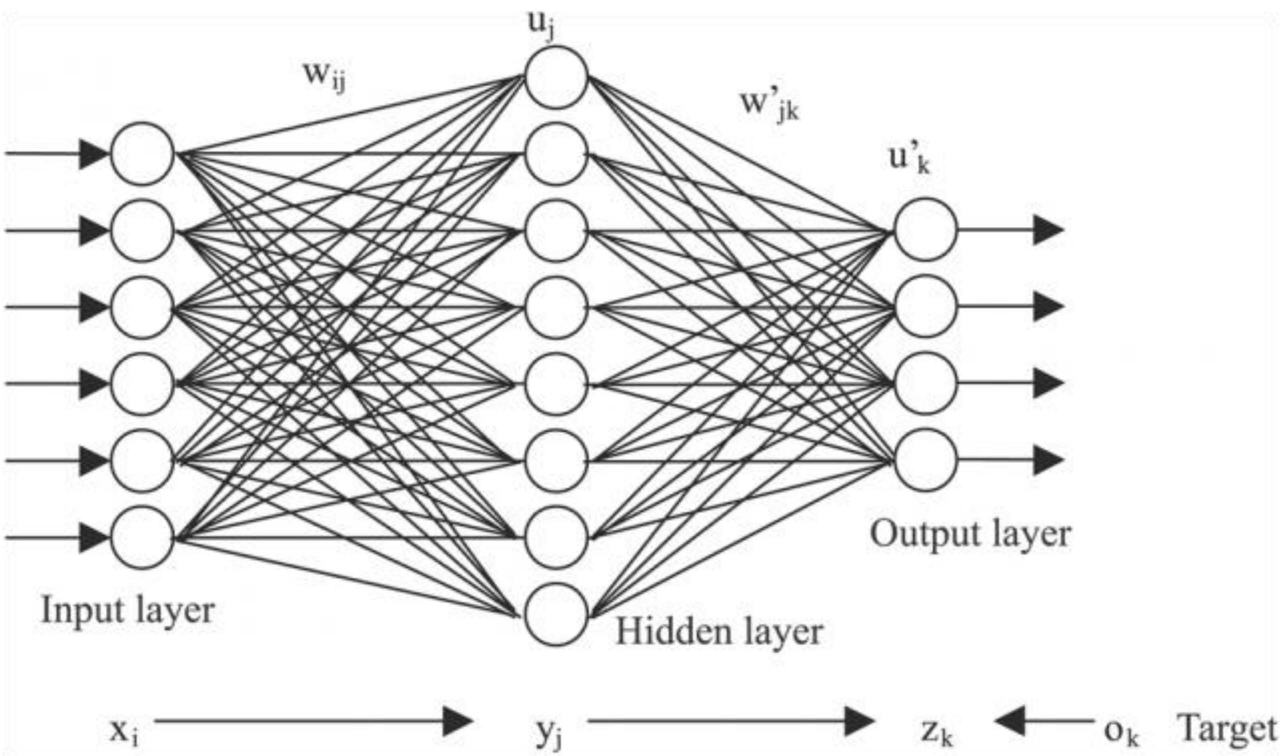
What are Artificial Neural Networks (ANNs)?

The inventor of the first neurocomputer, Dr. Robert Hecht-Nielsen, defines a neural network as:

"...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.”




x1
x2
...
xn
$\sum_{i=1}^{n}(w_i \cdot x_i)$
activation function
Φ
Φ(x1, x2, … xn)


uj
wij
w'jk
u'k
Input layer
Hidden layer
Output layer
xi
yj
zk
ok
Target

Good basic blog to read:

A Beginner’s Guide to Neural Networks in Python and SciKit Learn 0.18 (https://www.springboard.com/blog/beginners-guide-neural-network-in-python-scikit-learn-0-18/)

Exercise: Apply ANN to Iris dataset (iris_data.csv)