Nuclear Talent course on Machine Learning in Nuclear Experiment and Theory

Daniel Bazin[1]

Morten Hjorth-Jensen[1]

Michelle Kuchera[2]

Sean Liddick[3]

Raghuram Ramanujan[4]

[1]Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University, East Lansing, Michigan, USA
[2]Physics Department, Davidson College, Davidson, North Carolina, USA
[3]Department of Chemistry and National Superconducting Cyclotron Laboratory, Michigan State University, East Lansing, Michigan, USA
[4]Department of Mathematics and Computer Science, Davidson College, Davidson, North Carolina, USA

Jun 18, 2020

Introduction

During the last two decades there has been a swift and amazing development of Machine Learning techniques and algorithms that impact many areas in not only Science and Technology but also the Humanities, Social Sciences, Medicine, Law, indeed, almost all possible disciplines. The applications are incredibly many, from self-driving cars to solving high-dimensional differential equations or complicated quantum mechanical many-body problems. Machine Learning is perceived by many as one of the main disruptive techniques nowadays.

Statistics, Data science and Machine Learning form important fields of research in modern science. They describe how to learn and make predictions from data, as well as allowing us to extract important correlations about physical process and the underlying laws of motion in large data sets. The latter, big data sets, appear frequently in essentially all disciplines, from the traditional Science, Technology, Mathematics and Engineering fields to Life Science, Law, education research, the Humanities and the Social Sciences.

Machine Learning, short overview

Ideally, machine learning represents the science of giving computers the ability to learn without being explicitly programmed. The idea is that there exist generic algorithms which can be used to find patterns in a broad class of data sets without having to write code specifically for each problem. The algorithm will build its own logic based on the data. You should however always keep in mind that machines and algorithms are to a large extent developed by humans. The insights and knowledge we have about a specific system, play a central role when we develop a specific machine learning algorithm.

Machine learning is an extremely rich field, in spite of its young age. The increases we have seen during the last three decades in computational capabilities have been followed by developments of methods and techniques for analyzing and handling large date sets, relying heavily on statistics, computer science and mathematics. The field is rather new and developing rapidly. Popular software libraries written in Python for machine learning like Scikit-learn, Tensorflow, PyTorch and Keras, all freely available at their respective GitHub sites, encompass communities of developers in the thousands or more. And the number of code developers and contributors keeps increasing.

A multidisciplinary approach

Not all the algorithms and methods can be given a rigorous mathematical justification (for example decision trees and random forests), opening up thereby large rooms for experimenting and trial and error and thereby exciting new developments. However, a solid command of linear algebra, multivariate theory, probability theory, statistical data analysis, understanding errors and Monte Carlo methods are central elements in a proper understanding of many of the algorithms and methods we will discuss.

Learning outcomes

These sets of lectures aim at giving you an overview of central aspects of statistical data analysis as well as some of the central algorithms used in machine learning. We will introduce a variety of central algorithms and methods essential for studies of data analysis and machine learning.

Hands-on projects and experimenting with data and algorithms plays a central role in these lectures, and our hope is, through the various examples discussed in this series of lectures, to expose you to fundamental research problems in these fields, with the aim to reproduce state of the art scientific results. You will learn to develop and structure codes for studying these systems, get acquainted with computing facilities and learn to handle large scientific projects. A good scientific and ethical conduct is emphasized throughout the course. More specifically, you will

1. Learn about basic data analysis, data optimization and machine learning;

2. Be capable of extending the acquired knowledge to other systems and cases;

3. Have an understanding of central algorithms used in data analysis and machine learning;

4. Understand methods for regression and classification;

5. Methods we will focus on are Linear and Logistic Regression, Decision trees, random forests, bagging and boosting and various variants of deep learning methods, from feed forward neural networks to more advanced methods;

6. Work on numerical examples to illustrate the theory;

Types of Machine Learning

The approaches to machine learning are many, but are often split into two main categories. In *supervised learning* we know the answer to a problem, and let the computer deduce the logic behind it. On the other hand, *unsupervised learning* is a method for finding patterns and relationship in data sets without any prior knowledge of the system. Some authours also operate with a third category, namely *reinforcement learning*. This is a paradigm of learning inspired by behavioral psychology, where learning is achieved by trial-and-error, solely from rewards and punishment.

Another way to categorize machine learning tasks is to consider the desired output of a system. Some of the most common tasks are:

- Classification: Outputs are divided into two or more classes. The goal is to produce a model that assigns inputs into one of these classes. An example is to identify digits based on pictures of hand-written ones. Classification is typically supervised learning.

- Regression: Finding a functional relationship between an input data set and a reference data set. The goal is to construct a function that maps input data to continuous output values.

- Clustering: Data are divided into groups with certain common traits, without knowing the different groups beforehand. It is thus a form of unsupervised learning.

Choice of Programming Language

Python plays nowadays a central role in the development of machine learning techniques and tools for data analysis. In particular, seen the wealth of machine learning and data analysis libraries written in Python, easy to use libraries with immediate visualization(and not the least impressive galleries of existing

examples), the popularity of the Jupyter notebook framework with the possibility to run **R** codes or compiled programs written in C++, and much more made our choice of programming language for this series of lectures easy. However, since the focus here is not only on using existing Python libraries such as **Scikit-Learn**, **Tensorflow** and **Pytorch**, but also on developing your own algorithms and codes, we will as far as possible present many of these algorithms either as a Python codes or C++ or Fortran (or other languages) codes.