Data Analysis and Machine Learning: Representing data

Morten Hjorth-Jensen[1,2]

[1]Department of Physics, University of Oslo

[2]Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University

May 22, 2018

Introduction

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. It has become more and more common to see research projects on big data in for example the Social Sciences where extracting patterns from complicated survey data is one of many research directions. Having a solid grasp of data analysis and machine learning is thus becoming central to scientific computing in many fields, and competences and skills within the fields of machine learning and scientific computing are nowadays strongly requested by many potential employers. The latter cannot be overstated, familiarity with machine learning has almost become a prerequisite for many of the most exciting employment opportunities, whether they are in bioinformatics, life science, physics or finance, in the private or the public sector. This author has had several students or met students who have been hired recently based on their skills and competences in scientific computing and data science, often with marginal knowledge of machine learning.

Machine learning is a subfield of computer science, and is closely related to computational statistics. It evolved from the study of pattern recognition in artificial intelligence (AI) research, and has made contributions to AI tasks like computer vision, natural language processing and speech recognition. 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.

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 packages 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. Not all the algorithms and methods can be given a rigorous mathematical justification, 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 algorithms and methods we will discuss.

Learning outcomes

These 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 projects and exercies, 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 large 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, Bayesian statistics, Monte Carlo methods, 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. Gain knowledge of central aspects of Monte Carlo methods, Markov chains, Gibbs samplers and their possible applications, from numerical integration to simulation of stock markets;
5. Understand methods for regression and classification;
6. Learn about neural network, genetic algorithms and Boltzmann machines;

7. Work on numerical projects to illustrate the theory. The projects play a central role and you are expected to know modern programming languages like Python or C++, in addition to a basic knowledge of linear algebra (typically taught during the first one or two years of undergraduate studies).

There are several topics we will cover here, spanning from a statistical data analysis and its basic concepts such expectation values, variance, covariance, correlation functions and errors, via well-known probability distribution functions like uniform distribution, the binomial distribution, the Poisson distribution and simple and multivariate normal distributions to central elements of Bayesian statistics and modeling. We will also remind the reader about central elements from linear algebra and standard methods based on linear algebra used to fit functions such Cubic splines and gradient methods for data optimization and the Singular-value decomposition and least square methods for parameterizing data.

We will also cover Monte Carlo methods, Markov chains, well-known algorithms for sampling stochastic events like the Metropolis-Hastings and Gibbs sampling methods. An important aspect of all our calculations is a proper estimation of errors. Here we will also discuss famous resampling techniques like the blocking, bootstrapping and jackknife methods.

The second part of the material covers several algorithms used in machine learning.

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.

The methods we cover have three main topics in common, irrespective of whether we deal with supervised or unsupervised learning. The first ingredient is normally our data set, the second is a model which is normally a function of some parameters. The last ingredient is a so-called **cost** function which allows us to present an estimate on how good our model is in reproducing the data it is supposed to train.

Here we will build our machine learning approach on elements of the statistical foundation discussed above, with elements from data analysis, stochastic processes etc. We will discuss the following machine learning algorithms

1. Linear regression and its variants, in essence polynomial regression
2. Decision tree algorithms, from simpler to more complex ones
3. Nearest neighbors models
4. Bayesian statistics and regression
5. Support vector machines and finally various variants of
6. Artifical neural networks and deep learning

Why this text?

Choice of programming language

Data handling, machine learning and ethical aspects

Acknowledgements