January 15-19: Advanced machine learning and data analysis for the physical sciences

Morten Hjorth-Jensen[1,2]

[1]Department of Physics and Center for Computing in Science Education, University of Oslo, Norway
[2]Department of Physics and Astronomy and Facility for Rare Isotope Beams, Michigan State University, East Lansing, Michigan, USA

Dec 21, 2023

Overview of week first week

1. Presentation of course and participants
2. Discussion of possible projects
3. Discussion of evaluation forms
4. Eventual start with theory discussions on deep learning methods

Practicalities and possible projects

1. Although the course is defined as a self-study course, we can have weekly lectures with small weekly exercise assignments
2. We plan to work on two projects which will define the content of the course, the format can be agreed upon by the participants but the following topics could define an outline for possible projects and machine learning topics
   - Deep learning with the aim to develop a code for CNNs and/or RNNs and study data of relevance for own research (Higgs challenge for example)
   - Study autoencoders and variational autoencoders with application to own data
   - GANs and applications to own data

- Solve quantum/or classical many-body problems with deep learning methods (overlaps with FYS4411)
- Physics informed Machine Learning, applications to for example solution of Navier-Stokes equations
- Bayesian Machine Learning and Gaussian processes
- and many other research paths and topics

3. No exam, only two projects whoch count 1/2 of the final grade each
4. All info at the GitHub address `https://github.com/CompPhysics/AdvancedMachineLearning`

Deep learning methods covered

1. Deep learning, classics
   (a) Feed forward neural networks and its mathematics (NNs)
   (b) Convolutional neural networks (CNNs)
   (c) Recurrent neural networks (RNNs)
   (d) Autoencoders and principal component analysis
   (e) Physics informed neural networks
2. Deeep learning, generative methods
   (a) Boltzmann machines and energy based methods
   (b) Autoencodervariational autoencoders (VAEe)
   (c) Generative Adversarial Networks (GANs)

The lecture notes contain a more in depth discussion of these methods, in particular on neural networks, CNNs and RNNs.

Autoencoders and Variational Autoencoders

Autoencoders are artificial neural networks capable of learning efficient representations of the input data (these representations are called codings) without any supervision (i.e., the training set is unlabeled). These codings typically have a much lower dimensionality than the input data, making autoencoders useful for dimensionality reduction.

More importantly, autoencoders act as powerful feature detectors, and they can be used for unsupervised pretraining of deep neural networks.

GANs

Generative modeling is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

Generative models describe a class of statistical models that are a contrast to **discriminative models**. Informally we say that generative models can generate new data instances while discriminative models discriminate between different kinds of data instances. A generative model could generate new photos of animals that look like 'real' animals while a discriminative model could tell a dog from a cat. More formally, given a data set x and a set of labels / targets y. Generative models capture the joint probability $p(x, y)$, or just $p(x)$ if there are no labels, while discriminative models capture the conditional probability $p(y|x)$. Discriminative models generally try to draw boundaries in the data space (often high dimensional), while generative models try to model how data is placed throughout the space.

Kernel regression (Gaussian processes) and Bayesian statistics

Kernel machine regression (KMR), also called Gaussian process regression, is a popular tool in the machine learning literature. The main idea behind KMR is to flexibly model the relationship between a large number of variables and a particular outcome (dependent variable).

Physics informed machine learning

Here we can discuss neural networks that are trained to solve supervised learning tasks while respecting any given law of physics described by general nonlinear partial differential equations.

See also `https://arxiv.org/abs/2211.08064`.