

Introduction to Machine Learning

Jonathan Wayne Korn @ Columbia University

2021-04-30

Introduction to Machine Learning

We have to ask ourselves a fundamental question regarding the idea of a machine learning, which is:



Definition of Learning

- ▶ **One Definition:** Learning is optimizing performance (*based on some criterion*) using example data or past experience.
 - ▶ In that case, the goal of machine learning is to build algorithms enabled by constraints exposed by representations that support models targeted at:
 - ▶ Thinking,
 - ▶ Perception,
 - ▶ Action.
- ▶ **Another Definition:** Statistical learning refers to a vast set of tools for understanding data.
 - ▶ These tools can be classified as supervised or unsupervised machine learning models.

Defintion of Machine Learning

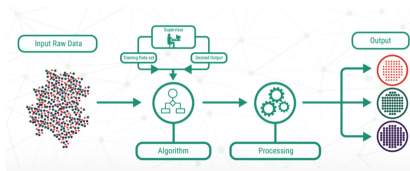
- ▶ ***“Field of study that gives computers the ability to learn without being explicitly programmed” ~ Arthur Samuel 1959***
- ▶ ***“How to construct programs that automatically learn from experience” ~ Mitchell 1997***
- ▶ There are two main types of machine learning including ***supervised and unsupervised learning***.
- ▶ We'll talk about supervised learning more than unsupervised learning.

Other Machine Learning Perspectives

- ▶ ***Information based Learning:*** We use information to guide or inform our decisions or lead us to the next question. When we automate this we will optimize this by minimizing errors or entropy or some parameter. Example: decision tree.
- ▶ ***Similarity based Learning:*** What we are doing is putting together things that are similar. Example: regression.
- ▶ ***Probability based Learning:*** Find the probability of belonging to different classes rather than exactly identifying the class. Example: NB
- ▶ ***Error Based Learning:*** What we're doing is developing a model, running it and checking our output against something then tweaking the model to reduce the error we find when we run our check. Example: KNN

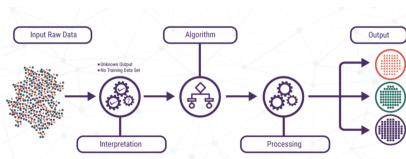
What is Supervised Learning?

- ▶ ***“Supervised learning is the machine learning task of inferring a function from labeled training data.” ~ Mohri 2012***
- ▶ The training data consist of a set of training examples In supervised learning, each example is a pair consisting of an input object (*typically a vector*) and a desired output value (*also called the supervisory signal*).
- ▶ The goal is to find that algorithm with which you can infer the given output from the given input.



What is Unsupervised Learning?

- ▶ In unsupervised learning, we are encountering data which is not labeled like it was in supervised learning.
- ▶ Usually means finding an algorithm that can be used to produce output given some un described input.
- ▶ For example, given some data we might try to find a way to cluster it in order to formulate an algorithm that describes the data in terms of that clustering.



Supervised Learning

- ▶ Supervised Learning involves:

- (1) Training to learn.
- (2) A limited predictive capacity by historical nature of training data, etc.

- ▶ Techniques include:

- ▶ statistical learning,
- ▶ decision tree,
- ▶ naives baye,
- ▶ nearest neighbors,
- ▶ support vector machines,
- ▶ ensembles of models.

Unsupervised Learning

- ▶ Unsupervised learning involves:
 - (1) Inferring based on finding structures in unlabeled data.
 - (2) A limited predictive capacity by the lack of reinforcement, such as an error signal.
- ▶ Techniques include:
 - ▶ clustering,
 - ▶ anomaly detection,
 - ▶ ann's.

Classical Machine Learning

The following types of problems are useful :

- ▶ Classification
- ▶ Regression
- ▶ Time Series
- ▶ Clustering

Classification

A classification problem is defined by the target or target variables. If the target variable is categorical in nature than it is a classification problem. Think of categorical as groupings and classification as the act of assigning the correct grouping. In these cases, you could face either a binary or multiple classification problems. Classification models are very useful to use on problems that involve a decision based on a certain set of input(s). Refer to section titled ***Example of Classification Problem using EzML*** for an example of how to conduct a classification model.

Regression

In a regression problem the target variable is linear in nature. The predictive purpose of regression is to predict the linear trend of a target variable from an input or inputs. Refer to section titled ***Example of Regression Problem using EzML*** for an example of how to conduct a regression model.

Time Series

Times series problems are regression problems with a time element. However, it should be known that times series models may be classification type problems if each time stamp requires a time based prediction or even a combination of the two. Many time series solutions actually use a distributed machine learning system that provides both a classification and regression prediction for the future time period. Refer to section titled ***Example of Time Series Problem using EzML*** for an example of how to conduct a time series model.

Clustering

Clustering problems are classification problems, however, there is no target variable hence we are dealing with unsupervised technique. The model will attempt to use optimized linear techniques to cluster inputs into groupings based on their relationships. Different models utilize various optimized techniques and linear methods to group the inputs and associate the relationships.

Supervised Deep Learning

We will focus on supervised learning techniques including:

- (1) Deep Learning using Structured Data (*Classification and Regression*)
- (2) Deep Learning using Unstructured Data
 - ▶ Text Classification
 - ▶ Text Generation
 - ▶ Image Classification

We will be discussing the various types of deep learning models that there are and how to construct them including:

- ▶ Recurrent Neural Networks (*RNN*)
- ▶ Long Short Term Memory Networks (*LSTM*)
- ▶ Convolutional Neural Networks (*CNN*)
- ▶ Hybrid Neural Networks (*i.e. Convolutional Long Short Term Memory Network (C-LSTM)*)

Deep Learning using Structured Data

We will discuss the proper steps to modeling with structured data using deep learning for classification and regression problems.

(Regression) using Deep Learning

If the problem involves data that is linear in nature and you need to predict the future value or linear **target variable** you most likely need a regression model.

Deep Learning is very effective at regression models and provides reliable results, but may be difficult to train effectively.

We will discuss some best practices when using deep learning to address a regression problem (*i.e. predicting the future value of the close price of an asset in the stock market.*)

(Classification) using Deep Learning

A deep learning model designed for classification is necessary when the problem/data involves predicting the likely outcome of a ***target variables*** that is categorical in nature.

We will be discussing some best practices to design a classifier using deep learning. (*i.e. predicting the class of a type of flower from measurements of the different varieties.*)

Deep Learning using Unstructured Data: (*Text*)

There are a few options when it comes to using supervised deep learning techniques for text problems.

In this course we will focused on two methods ***text classification and text generation***.

Each are basically a form of classification as the classifiers will simple input a text and attempt to infer the class from the recognized features, and the text generation will esentially predict the next character of a text from the features it learns during training.

Text Classification using Deep Learning

One of the most important steps to modeling with text is the actual reshaping of the data for the model to ingest.

We are going to discuss the appropriate steps to process your text data.

We will then discuss the steps to architect a deep learning model to classify texts (*... the purpose of the classification that a model is targeted to address is all determined by the data.*), train the model, and test the model's performance.

Text Generation using Deep Learning

We will discuss how to construct a deep text generation model that can train on samples texts and attempt to write its own text outputs.

We will discuss how the samples of text and how they are pre-processed and shaped may effect the overall results.

We will need to be creative in the training approach if we expect the model to become a writer like a human.

Deep Learning using Unstructured Data (*Images*)

One of the most important steps to modeling with images is the actual reshaping of the data for the model to ingest.

We are going to discuss the appropriate steps to process your text data.

We will then discuss the steps to architect a deep learning model to classify images, train the model, and test the model's performance.

Image Classification using Deep Learning

We will discuss how to construct a deep image classifier model that can train on samples images and attempt to classify its own outputs.

We will discuss how the samples of images and how they are pre-processed and shaped may effect the overall results.

We will need to be creative in the training approach if we expect the model to become a successful image classifier.