Introduction to Machine Learning

Jonathan Wayne Korn @ Columbia University

2021-06-18

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

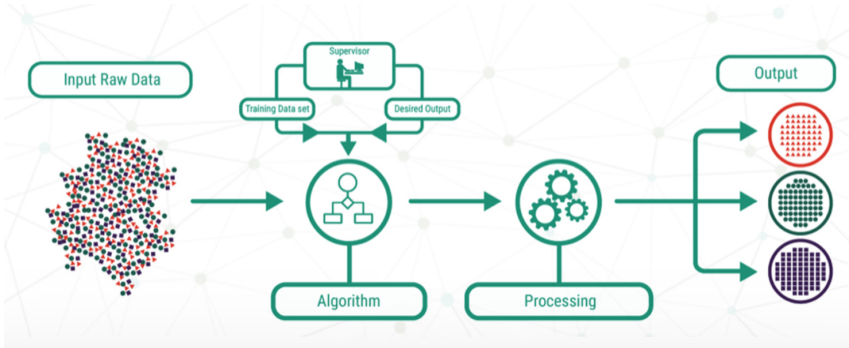
- ***"Field of study that gives computers the ability to learn without being explicitly programmed"*** ~ *Arthur Samuel 1959*
- ***"How to construct programs that automatically learn from experience"*** ~ *Mitchell 1997*
- There are two main types of machine learning including ***supervised and unsupervised 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

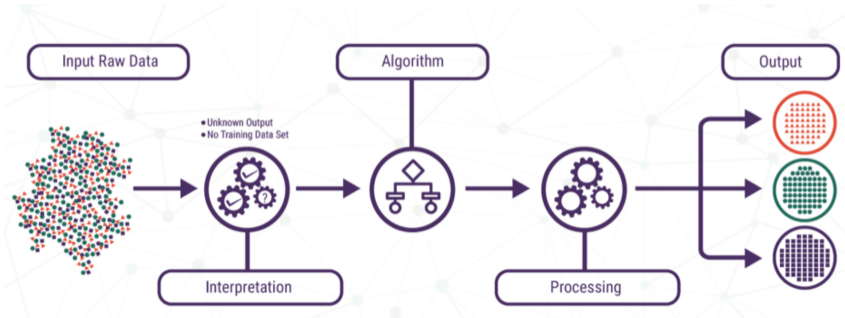
What is Supervised Learning?

- ***"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

- Functions as its name suggests, by classifying data in order to then make predictions.
- Is one of the supervised learning techniques.
- Used for predictive modeling.
- Uses a classification model via a learning algorithm.

Regression

- Regression analysis is a fundamental concept in the field of machine learning.
  - It falls under supervised learning wherein the algorithm is trained with both input features and output labels.
  - It helps in establishing a relationship among the variables by estimating how one variable affects the other.
  - It consists of mathematical methods that allow data scientists to predict a continuous outcome (y) based on the value of one or more predictor variables (x).
- Linear regression is probably the most popular form of regression analysis because of its ease-of-use in predicting and forecasting.

Time Series

- A time series is a sequence of observations taken sequentially in time.
- Time series forecasting involves taking models then fit them on historical data then using them to predict future observations.
- The steps that are considered to shift the data backward in the time(sequence), called lag times or lags.
- Therefore, a time series problem can be transformed into a supervised ML by adding lags of measurements as inputs of the supervised ML.

Clustering

- It is basically a type of unsupervised learning method .
- An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses.
- Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

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 eseentially 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 the 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 Classifcation using Deep Learning

- Image classification is a supervised learning problem: define a set of target classes *(objects to identify in images)*, and train a model to recognize them using labeled example photos.
- Early computer vision models relied on raw pixel data as the input to the model.