Introduction to Machine Learning for Newbies

Part I: Conceptual Definitions, Basic Principles, Kinds, and Jargons

Kyungmin Kim, Ph.D.

Created Apr 4, 2016
Last updated Apr 10, 2016

Preface

This material is made for very newbies who are interested in Machine Learning. So, this material only covers conceptual introduction to machine learning and summarizes several terminologies (we call them as jargons.) when you see any materials about machine learning.

Actually, I'm not an expert on machine learning but just an user who have been using it for my researches such as search for gravitational waves or identification of supernovae images. Thus, some of contents or statements may not be general expressions. Instead of using general definitions or expressions, I try to explain them in my own way based on what I've learned from my experiences.

I hope this material may help your basic conceptual understanding about machine learning! Let's start!

What is Machine Learning (ML)?

Wikipedia's the first statement says, "Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.". Following the first statement, wikipedia also show Arthur Samuel's definition in 1959: "A field of study that gives computers the ability to learn without being explicitly programmed".

Can you catch up what is machine learning from above sentences? Well, it's hard to get a direct intuition for me. If you can, you don't need this material at all! Please save your time and find other advanced materials. But if not, let me explain what is it.

Let's suppose you have a huge (scientific, medical, or social) data and you want to find a certain pattern (or meaningful information) from the data without doing it by yourself.

Then, you can let a computer, i.e. machine, learn the characteristics of the data first and you give an order to let the machine give the answer for a new, unknown input data to human.

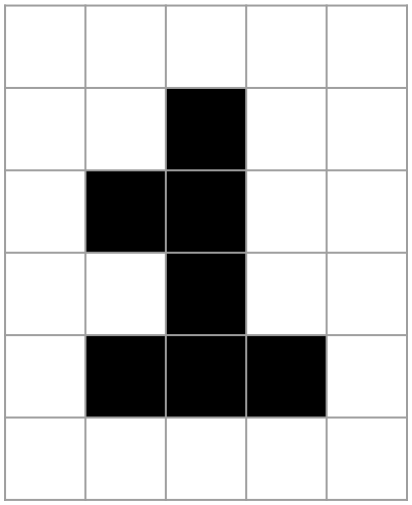
But I think this explanation is still abstract to you. Let's see how these things can be realized.

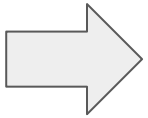
How to teach a machine?

Unfortunately, most machines used for machine learning can't learn human's language directly yet.

So, in practice, teaching a machine means let the machine determine a set of parameters of a mathematical algorithm.

Meanwhile, the textbook for teaching a machine usually written in numbers. For example, when we teach a machine to distinguish a 2-D image of a number among 0 to 9, we need to convert the image of a number to a digitized set.





0	0	0	0	0
0	0	1	0	0
0	1	1	0	0
0	0	1	0	0
0	1	1	1	0
0	0	0	0	0

A break slide:
Difference between ML and ML Algorithm?

ML means the action itself about human let machine study a data and let it return an (expected) answer to human. So, it's rather an abstract concept in my opinion.

While, ML Algorithm (MLA) means a mathematical algorithm or a computational code which describe the details of actual processes of either training a machine or getting an answer from the machine.

Usages of ML

- Regression
  - Finding a line or a curve or a hyperplane which represents the tendency of scattered data points.
  - A simple example is the least squares. Then you can imagine regression of ML is a multidimensional big data version of the least squares.
  - Examples
    - Modeling
    - Stock price expectation

- Classification
  - Classifying a data into a category among possible choices.
  - Examples
    - Image recognition
    - Signal identification
    - AlphaGo

Categories of ML

- Unsupervised Learning
  - Finds a fiducial boundary from a given input data in order to let the boundary divides the data into N different categories.
  - A representative technique for this is ‘clustering’ given input data into several categories.

- Supervised Learning
  - Trains a machine with an example data first to let the machine learn the tendency of a given data and then evaluate a new input data with the parameters determined from the training process.
  - We will focus on the supervised learning for classification only because it is widely used in many scientific problems.

MLAs for Supervised Learning

There are several famous and popular MLAs:

- Decision Tree
  - Performs binary splittings until a termination condition to be satisfied.
  - Alternatives: Random Forest, Boosted Decision Tree, Bagged Decision Tree
- Artificial Neural Network
  - Mimicks biological neural network. It finds the most strong connection weight between two adjacent neurons in order to activate the receiver neuron.
  - Alternative: Deep Neural Network
- Support Vector Machine
  - Finds a marginal hyperplane which can maximally divides a set of data into several categories.

Note: Some of the listed MLAs are also used for the unsupervised learning.

Decision Tree

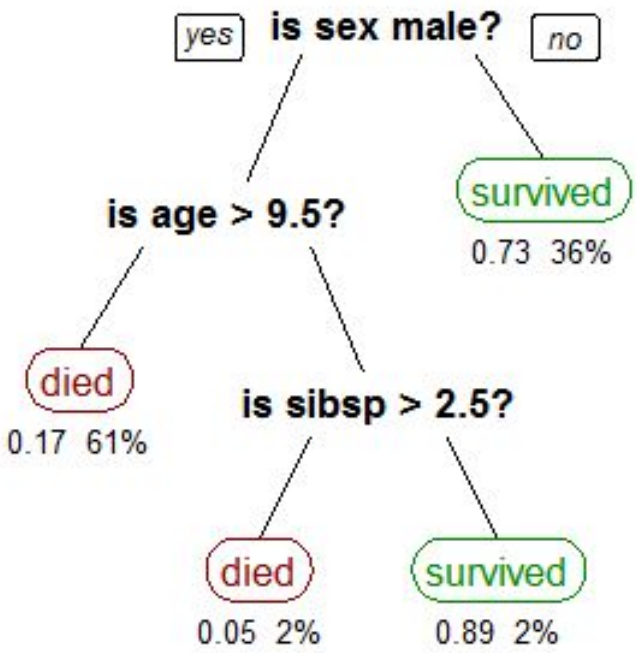
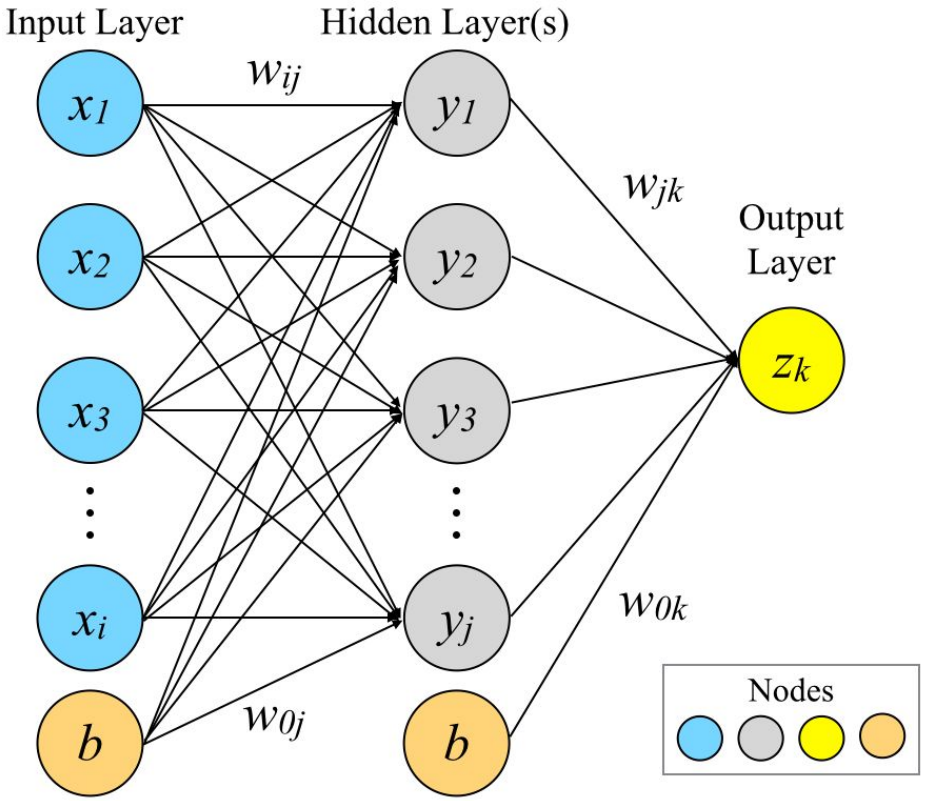


Image from Wikipedia

Artificial Neural Network



Support Vector Machine

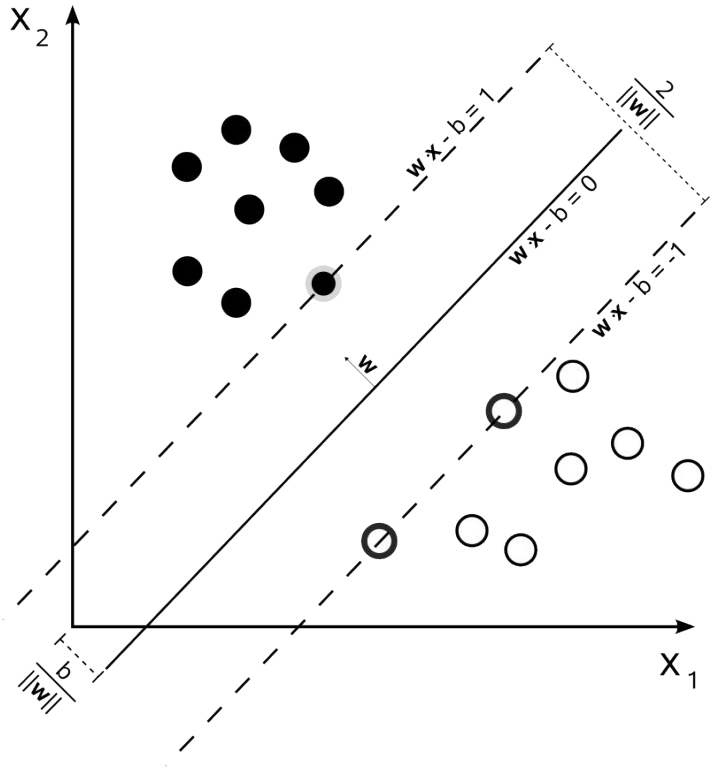


Image from Wikipedia

How do Supervised MLAs its work?

1. A machine learns a certain pattern from a set of example (training) data.

2. The machine decides an optimal criterion based on what it has learned.
   a. The determined criterion is nothing but a set of optimal parameters of an algorithm.
   b. Usually, when an observed error between the pre-assigned class and the obtained value from the algorithm is less than a certain tolerance, the training process is finished. But, the error is larger than the tolerance, the training process iterates the same process by updating the parameters until the error to be less than the tolerance. When the error becomes smaller than the tolerance, we say that “The iteration reaches at the global minima.”.

3. When a new, i.e., an unknown input data is provided, the machine judges which category is the most probable one for this unknown data.
   a. Usually, the judgement is achieved based not on a deterministic value but on a probabilistic value because we expect there is hardly the same data in the training data with the unknown data. So, you may understand that the reason of why the machine gives that answer is because it judges the unknown input data corresponds the most likely to that certain category among possible choices.

Basic but very important key issues you should consider

- In Data Preparation,
  - Proper selection for features: Some features may be useful and others may be useless.
  - Sufficient and appropriate training data: You need to prepare training data as many as possible in order to avoid either biased classification or over/under estimation. Also, if possible, the same number of samples for each class is recommended. In addition, if possible, configure the set of each class data to have less overlap with other class data at most.

- In Choosing a MLA,
  - Each MLA may show different performance even on the same data. So, you may need to prepare similar but different data sets, pairs of train data and test data, and test multiple times in order to find the best MLA for your problem. But if it is hard to prepare sufficiently many data, there is a Plan B: you may divide your data into several pairs then perform test with them.
  - Also, the data type, e.g., raw data or normalized data, depends on the chosen MLA. If your data is not normalized one but your chosen MLA need normalized data, you need to a preprocess for the data preparation.

Jargons

- Features (or Input variables): Variables which characterize a data.
  - E.g. age, gender, race, height, weight, and so on of a patient in a medical data
- Input data (or Input samples): The data which is used for either training or evaluation.
- Category (or Class): Pre-assigned quantity for each data.
  - Usually it is either 1 or 0 for two class problem.
  - In some cases, strings, e.g., foreground or background, are used directly.
- Train: It refers the training process.
- Test (or Evaluation): A process that examing the trained machine whether a machine properly determines a set of optimal parameters or not. So, we usually test the trained machine with some known data such as a part of training data because the train data have pre-assigned classes, i.e., we know the answer for chosen test data already.

Jargons (cont'd)

- Performance Test: It refers analyzing the result of test process in order to look whether the trained machine classify the (known) test data properly or not in probabilistic point of view.
  - Also, run time for training and testing is also quoted for summarizing a MLA's performance.
- True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), etc.: Usually these statistical quantities are used in the performance test.
  - https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Have fun
with
Machine Learning! ;)