Mathematical foundations of machine learning

Hông Vân Lê
Institute of Mathematics, CAS

VIASM, Hà nôi, August 21-24, 2018

Lecture 1: Learning, machine learning and artificial intelligent.

1. What are learning, deductive learning and machine learning.

2. History of machine learning and artificial intelligence.

3. Current tasks and main type of machine learning.

4. Basic questions in mathematical foundation of machine learning.

1. What are learning, deductive learning and machine learning?

(a) Small children learn to speak by observing, repeating and mimicking adults' phrases. Their way of learning is inductive learning.

(b) In school we learn mathematics, physics, biology, chemistry by following the instructions of our teachers and those in textbooks. We learn general rules and apply them to particular cases. This type of learning is deductive learning.

(c) Experimental physicists design experiments and observe the outcomes of the experiments to validate or dispute a conjecture on the nature of the observables. This type of learning is inductive learning.

• A learning is a process of gaining new knowledge by examination of empirical data of the observables. A learning is successful if the knowledge can be tested in examination of new data. A machine learning is an automatized process of learning.

Definition (Russell and Norvig - Artificial Intelligence - A modern Approach) An agent (human, robot, machine) is learning if it improves its performance on future tasks after making observations about the world.

Mathematical definition (Vapnik) Learning is a problem of function estimation on the basis of empirical data.

In mathematical language experience is empirical data and knowledge is function estimation.

• A classical example of learning is that of learning a physical law by curve fitting to data. Assuming the law, an unknown function $f : \mathbf{R} \to \mathbf{R}$, has a specific form and that the space of all functions having this form can be parameterized by N real numbers. For instance, if f is assumed to be a polynomial of degree d, then $N = d + 1$ and the parameters are the coefficients $w_0, \cdots, w_d$ of f. In this case, finding the best fit by the least squares method estimates the unknown f from a set of pairs $(x_1, y_1), \cdots, (x_m, y_m)$.

One computes the vector of coefficients w such that the value

$$\sum_{i=1}^{m} (f_w(x_i) - y_i)^2 \text{ with } f_w(x) = \sum_{j=0}^{d} w_j x^j$$

is minimized where, typically $m > N$. If the measurements generating this set were exact, then $f(x_i)$ would be equal to y_i. But in general one expects the values y_i to be affected by noise. The least square technique, going back to Gauss and Legendre, which is computational efficient and relies on numerical linear algebra.

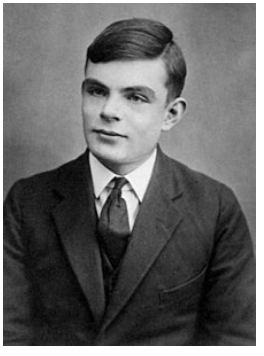
The least-squares method is usually credited to Carl Friedrich Gauss (1809), but it was first published by Adrien-Marie Legendre (1805).

2 History of machine learning and artificial intelligence

• 1945 Vannevar Bush proposed in "As We May Think" published in "The Atlantic", a system which amplifies peoples own knowledge and understanding. Bush's memex was based on what was thought, at the time, to be advanced technology of the future: ultra high resolution microfilm reels, coupled to multiple screen viewers and cameras, by electromechanical controls. Through this machine, Bush hoped to transform an information explosion into a knowledge explosion.

• 1948 John von Neumann suggested that machine can do any thing that peoples are able to do.

• 1950 Alan Turing asked Can machines think? in "Computing Machine and Intelligence" and proposed the famous Turing test. The Turing is carried out as imitation game. On one side of a computer screen sits a human judge, whose job is to chat to an unknown gamer on the other side. Most of those gamers will be humans; one will be a chatbot with the purpose of tricking the judge into thinking that it is the real human.



Alan Turing (1912-1950)

• 1956 John McCarthy coined the term artificial intelligence.

• 1959, Arthur Samuel, the American pioneer in the field of computer gaming and artificial intelligence, defined machine learning as a field of study that gives computers the ability to learn without being explicitly programmed. The Samuel Checkers-playing Program appears to be the world's first self-learning program, and as such a very early demonstration of the fundamental concept of artificial intelligence (AI).

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation, which were unsolvable because of small capacity of hardware memory and slow speed of computers that time. By 1980, expert systems had come to dominate AI, and statistics was out of favor. Expert system uses the idea that "intelligent systems derive their power from the knowledge they possess rather than from the specific formalisms and inference schemes ".

Work on symbolic based learning did continue within AI, leading to inductive logic programming. Neural networks research had been abandoned by AI and computer science around the same time. Their main success came in the mid-1980s with the reinvention of a algorithm in neural network which was able thanks to increasing speed of computers and increasing hardware memory.

Machine learning, reorganized as a separate field, started to flourish in the 1990s.

• AI $\rightsquigarrow$ ML: tackling solvable problems of a practical nature.

• Methods and models borrowed from statistics and probability theory. Laplacian determinism $\rightsquigarrow$ probabilistic modeling of random observables - new paradigm shift in sciences.

• The current trend is benefited from Internet.

In the book by Russel and Norvig "Artificial Intelligence a modern Approach" (2010) AI encompass the following domains:

- natural language processing,
- knowledge representation,
- automated reasoning to use the stored information to answer questions and to draw new conclusions;
- machine learning to adapt to new circumstances and to detect and extrapolate patterns,
- computer vision to perceive objects,
- robotics.

All the listed above domains of artificial intelligence except knowledge representation and robotics are now considered domains of machine learning. Pattern detection and recognition were and are still considered to be domain of data mining but they become more and more part of machine learning. Thus AI = knowledge representation + ML + robotics.

• representation learning, a new word for knowledge representation but with a different flavor, is a part of ML.

- Robotics = ML + hardware.

Why did such a move from artificial intelligence to machine learning happen?

The answer is that we are able to formalize most concepts and model problem of artificial intelligence using mathematical language and represent as well as unify them in such a way that we can apply mathematical methods to solve many problems in terms of algorithms that machine are able to perform.

3. Current tasks and types of ML.

Main tasks

• Classification task assigns a category to each item. For example, document classification may assign items with categories such as politics, email spam, sports, or weather, image classification may assign items with landscape, portrait, or animal. A classification task is a construction of a function on the set of items that takes value in a countable set of categories.

• As we have remarked in the mathematical example of learning (p. 6) usually we have ambiguous/incorrect measurement and we have to add a "noise" to our measurement. If every thing would be exact, the classification task is the classical interpolation function problem in mathematics. In real life and for machine learning we need to model the noise using probability theory. Therefore machine learning is based on statistical learning theory, which we shall learn in tomorrow lecture.

• Regression task predicts a real value for each item. Examples of regression tasks include prediction of stock values or variations of economic variables. In this problem, the penalty for an incorrect prediction depends on the magnitude of the distance between the true and predicted values, in contrast with the classification problem, where there is typically no notion of closeness between various categories. A regression task is a (construction of a) function, on the set of items that takes value in $\mathbb{R}$, taking into account a "noise" from incorrect measurement.

The term regression was coined by Francis Galton in the 19 century to describe a biological phenomenon that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean of population). For Galton, regression had only this biological meaning, but his work was later extended to a more general statistical context. Galton method of investigation is non-standard at that time: first he collected the data, then he guessed the relationship model of the events.

• Density estimation task finds the distribution of inputs in some distribution space. Karl Pearson (1857-1936) proposed that all observations come from some probability distribution and the purpose of sciences is to estimate the parameter of these distributions. Density estimation problem has been proposed by Ronald Fisher (1980-1962) as a key element of his simplification of statistical theory, namely he assumed the existence of a density function $p(\xi)$ that defines the randomness (noise) of a problem of interest.

The measure ν is called dominated by μ (or absolutely continuous with respect to μ), if $\nu(A) = 0$ for every set A with $\mu(A) = 0$. Notation: $\nu << \mu$. By Radon-Nykodym theorem, we can write

$$\nu = f \cdot \mu$$

and f is the density function of ν w.r.t. μ.

For example, the Gaussian distribution on the real line is dominated by the canonical measure dx and we express the standard normal distribution in terms of its density

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2).$$

The classical problem of density estimation is formulated as follows. Let a statistical model A be a class of densities subjected to a given dominant measure. Let the unknown density $p(x, \xi)$, where $\xi \in A$. The problem is to estimate the parameter ξ using i.i.d. data $X_1, \cdots, X_l$ distributed according to this unknown density $p(x, \xi)$.

Karl Pearson (1857-1936) Ronald Fisher (1890-1962)

• Ranking task orders items according to some criterion. Web search, e.g., returning web pages relevant to a search query, is the canonical ranking example. If the number of ranking is finite, then this task is close to the classification problem, but not the same, since in the ranking task we need to specify each rank during the task and not before the task as in the classification problem.

• Clustering task partitions items into (homogeneous) regions. Clustering is often performed to analyze very large data sets. Clustering is one of the most widely used techniques for exploratory data analysis. For example, computational biologists cluster genes on the basis of similarities in their expression in different experiments; retailers cluster customers, on the basis of their customer profiles, for the purpose of targeted marketing; and astronomers cluster stars on the basis of their spacial proximity.

• Dimensionality reduction or manifold learning transforms an initial representation of items in high dimensional space into a space of lower dimension while preserving some properties of the initial representation. A common example involves pre-processing digital images in computer vision tasks. We can regard clustering as dimension reduction too.

Main types of ML

The type of ML is defined by the type of interaction between the learner and the environment: the type of training data, i.e., the data available to the learner before making decision and prediction; and the type of the test data that are used to evaluate and apply the learning algorithm.

Main types of ML are supervised, unsupervised and reinforcement learning.

• In supervised learning a learning machine is a device that receives labeled training data, i.e, the pair of a known instance and its feature, also called label. In computer sciences language, a known instance is an input and and its feature is the output of a program that predicts the label for unseen instances. Examples of sets of labeled data are emails that are labeled "spam" or "no spam" and medical histories that are labeled with the occurrence or absence of a certain disease.

• Most of classification and regression problems of machine learning belong to supervised learning.

• In unsupervised learning there is no additional label attached to the data and the task is to describe structure of data. Since the examples (the avalable data) given to the learning algorithm are unlabeled, there is no straightforward way to evaluate the accuracy of the structure that is produced by the algorithm. Density estimation, clustering and dimensionality reduction are examples of unsupervised learning problems.

Most important applications of unsupervised learning are finding association rules that are important in market analysis, banking security and consists of important part of pattern recognition, which is important for understand advanced AI.

At the current time, unsupervised learning is primarily descriptive and experimental whereas supervised learning is more predictive (and has deeper theoretical foundation).

• Reinforcement learning is the type of machine learning where a learner actively interacts with the environment to achieve a certain goal. More precisely, the learner collects information through a course of actions by interacting with the environment. This active interaction justifies the terminology of an agent used to refer to the learner. The achievement of the agent's goal is typically measured by the reward he receives from the environment and which he seeks to maximize. For examples, reinforcement learning is used in self-driving car.

Reinforcement learning is aimed at acquiring the generalization ability in the same way as supervised learning, but the supervisor does not directly give answers to the students questions. Instead, the supervisor evaluates the students behavior and gives feedback about it.

Basic questions in mathematical foundations of ML

A learning is a process of gaining knowledge on a feature of observables by examination of partially available data. The learning is successful if we can make a "good" prediction on unseen data, which improves when we have more data.

Mathematical foundations of machine learning aim to answer the following questions
How and why do machine learn successfully?

1. What is the mathematical model of learning?

To answer Question 1 we need to specify our definition of learning in a mathematical language which can be used to build instructions for machines.

2. How to quantify the difficulty/complexity of a learning problem?

The difficulty of a problem shall be defined in terms of complexity: time complexity to solve a problem, resource complexity of a problem to have enough data/space/energy to solve a problem. If the complexity of a problem is very large then we cannot not learn it. So Question 2 contains the sub-question why can we learn a problem?

3. How to choose a learning algorithm?

Clearly we want to have the best learning algorithm, once we know a model of a machine learning which contains all possible learning algorithms. To answer Question 3 we need to measure success of a learning algorithm/a learning machine, e.g. we quantify the success in the rate/number of mistakes, which can be linked to the complexity of a learning problem. Thus Question 3 is related to the first and second Question.

4. Is there a mathematical theory underlying intelligence?

I shall discuss the last Question in the last lecture.

Future of machine learning and AI

Nowadays many machine learning systems can automate things that humans do well. Examples include image recognition, speech recognition, and email spam classification which are mostly supervised learning.

We are now surpassing human-level performance on more and more of the tasks where we can get easily labeled training data. Unsupervised learning currently is mostly experimental, since we cannot quantify the notion of success for unsupervised learning, e.g. for clustering. For example, it is not clear what is the "correct" clustering for given data or how to evaluate a proposed clustering. If we can quantify the "success" in an unsupervised learning problem then we can make a mathematical model for this problem.

Conclusion Machine learning is automatized learning, whose performance is improves with increasing volume of empirical data. Machine learning uses mathematical statistics to model incomplete information and the random nature of the observed data. Machine learning is the core part of artificial intelligence. Machine learning is very successful experimentally and there are many open questions concerning its mathematical foundations. Mathematical foundations of machine learning is important for building general purpose artificial intelligence, also called AGI, or UAI.

Recommended literature for the first two lectures

• F. Cucker and S. Smale, On mathematical foundations of learning, Bulletin of AMS, 39(2001), 1-49.

• B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, Building machines that learn and think like people. Behavioral and Brain Sciences,(2016) 24:1-101, arXiv:1604.00289.

• Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature, 521(2015), 452-459.

• S. J. Russell and P. Norvig, Artificial Intelligence A Modern Approach, Prentice Hall, 2010.

• V. Vapnik, The nature of statistical learning theory, Springer, 2000.