Mathematical Foundations of Machine Learning

Hông Vân Lê

October 6, 2020

- Plan of our course.

- Lecture: Basic mathematical problems in machine learning:

1. What are learning, inductive learning and machine learning.
2. History of machine learning and artificial intelligence.
3. Current tasks and main types of machine learning.
4. Basic mathematical problems in machine learning.

Plan of our course.

06.10: Basic mathematical problems of machine learning

13.10: Mathematical models of supervised learning

20.10 : Mathematical models of unsupervised learning

27.10: Mathematical models of reinforcement learning

3.11: Distance measure and maximum likelihood principle

10.11: Consistency of a learning algorithm

17.11: Generalization ability of machine learning

24.11: Support vector machine

01.12 : Kernel based SVMs

08.12: Neural networks and deep learning

15.12: Training neural networks

22.12: Bayesian machine learning and Bayesian networks

12.01: Discussion of exam questions and term papers.

EXAM: 24.01? 31.01?

Recommended literature

1. Lecture notes “Mathematical foundations of machine learning”

2. S. Shalev-Shwart and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

3. M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, MIT Press, 2012.

4. L. Deveroye, L. Györfi and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer 1996.

Suggested topics for term papers

1. Persistent homology (G. Carlsson, Topology and data, Bulletin of the AMS 46(2009), 255-308.)

2. Undecidability of learnability (S. Ben-David, P. Hrubes, S. Moran and A. Yehudayoff,

Learnability can be undecidable, Nature Machine intelligence, 1(2019), 44-48.)

3. Boosting, Decision tree, EM algorithms (see 1.1. (2) and T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd Edition, Springer 2008.)

4. Generative Adversarial Network (Charu C. Aggarwal, Neural Networks and Deep Learning, Springer 2018).

5. Information Theoretic Learning Models (Simons Haykin, Neural Networks and Learning Machines, Pearson Education, 2009)

Advanced topics

Online Seminar on Mathematical Foundations of Data Science (A weekly online seminar on random topics on mathematical foundations of machine learning, statistics and optimization)

1. Learning, inductive learning and machine learning

Inductive learning: from examples of words and phrases children learn the rules of combinations of these words and phrases into meaningful sentences.

Deductive learning: We learn general rules and apply them to particular cases.

- Mixed type of learning: learning in school.

- In mathematical theory of machine learning, or more general, in mathematical theory of learning we consider only inductive learning.

Definition. A learning is a process of gaining new knowledge, more precisely, new correlations of features of observable by examination of empirical data of the observable. Furthermore, a learning is successful if the correlations can be tested in examination of new data and will be more precise with the increase of data.

- Vapnik's mathematical postulation: "Learning is a problem of function estimation on the basis of empirical data".

Example: Learning a physical law by curve fitting to data. In mathematical terms, a physical law is expressed by a function f, and data are the value y_i of f at observable points x_i. The goal of learning in this case is to estimate the unknown f from a set of pairs $(x_1, y_1), \cdots, (x_m, y_m)$. Usually we assume that f belongs to a finite dimensional family of functions of observable points x.

For instance, f is assumed to be a polynomial of degree d over $\mathbf{R}$, i.e., we can write

$$f = f_w(x) := \sum_{j=0}^{d} w_j x^j$$

$$\text{where } w = (w_0, \cdots, w_d) \in \mathbf{R}^{d+1}.$$

In this case, to estimate $f = f_w$ is the same as to estimate the parameter w of f_w, observing the data (x_i, y_i).

The most popular method of curve fitting is the least square method which quantifies the error $R(w)$ of the estimation of the parameter w in terms of the value

$$R(w) := \sum_{i=1}^{m} (f_w(x_i) - y_i)^2 \qquad (1)$$

which the desired function f should minimize. If the measurement generating the data (x_i, y_i) were exact, then $f(x_i)$ would be equal to y_i and the learning problem is an interpolation problem. But in general one expects the values y_i to be affected by noise.

2 A brief history of machine learning

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

- In 1950 Alan Turing asked "Can machines think?" in "Computing Machine and Intelligence" and proposed the famous Turing test.

- In 1956 John McCarthy coined the term "artificial intelligence".

- In 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).

In the early days of AI, statistical and probabilistic methods were employed. Because of many theoretical problems and because of small capacity of hardware memory and slow speed of computers, statistical methods were out of favour. By 1980, expert systems, which were based on knowledge database, and inductive logic programming had come to dominate AI.

Neural networks returned back to machine learning with success in the mid-1980s with the reinvention of a new algorithm, called back-propagation, and thanks to increasing speed of computers and increasing hardware memory.

Machine learning, reorganized as a separate field, started to flourish in the 1990s. 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 robotics are now considered domains of machine learning.

- Representation learning and feature engineering are part of machine learning replacing the field of knowledge representation.

- Pattern detection and recognition become a part of machine learning.

- Robotics becomes a combination of machine learning and mechatronics.

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

Answer. We are able to formalize most concepts and model problems 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.

3a. Main tasks of current machine learning

- Classification task is a construction of a mapping

$$f : \mathcal{X} \to \mathcal{Y}$$

where $\mathcal{X}$ is a set of items we are interested in to a smaller countable set $\mathcal{Y}$ of other items, which are regarded as possible “features” of the items in the domain set, using given data $f(x_i) = y_i \in \mathcal{Y}$.

Example: document classification.

- Usually we have ambiguous/incorrect measurement and we call uncertainty of the exactness of the measurement a noise to our measurement. If every thing would be exact, the classification task is the classical interpolation function problem in mathematics.

- Regression task is a construction of a function

$$f : \mathcal{X} \to \mathbf{R}$$

using given data of the values of the desired function at given items. Examples of regression tasks include learning physical law by curve fitting to data with application to predictions of stock values or variations of economic variables.

- The error of f depends the distance between the true and predicted values, in contrast with the classification problem.

- Density estimation task is a construction (or estimation) of the distribution of observed items. In other words, given a sample space $(\mathcal{X}, \Sigma_{\mathcal{X}})$, where $\Sigma_{\mathcal{X}}$ denotes the σ-algebra of a measurable space $\mathcal{X}$, we need to find a map

$$\sigma : \mathcal{X} \to \mathcal{P}(\mathcal{X})$$

where $\mathcal{P}(\mathcal{X})$ - the set of all probability measures on $\mathcal{X}$.

This is one of basic problems in statistics. Over one hundred year ago Karl Pearson proposed that all observations come from some probability distribution and the purpose of sciences is to estimate the parameter of these distributions, assuming that they belong to a known finite dimensional family S of probability measures on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. Density estimation problem has been proposed by Ronald Fisher as a key element of his simplification of statistical theory, namely he assumed the existence of a density function $p(\xi)$ that governs the randomness of a problem of interest.

- Ranking task orders observed items according to some criterion. Web search, e.g. returning web pages relevant to a search query, is a typical 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.

- Dimensionality reduction, representation learning/manifold learning transforms an initial representation of items in high dimensional space into a representation of the items in a space of lower dimension while preserving some properties of the initial representation. A common example involves preprocessing digital images in computer vision tasks.

Many of dimensional reduction techniques are linear. When the technique is non-linear we speak about manifold learning technique. We can regard clustering as dimension reduction too.

3b. Main types of machine learning The type of a machine learning task is defined by the type of interaction between the learner and the environment. In particular, we consider types of training data, i.e., the data available to the learner before making decision and prediction, and the type of the outcomes.

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

- In supervised learning training data are labelled, i.e., each item in the training data set consists of the pair of a known instance and its feature, also called label. Examples 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. In these cases the learners output would be a spam filter and a diagnostic program, respectively. Most of classification and regression problems of machine learning belong to supervised learning.

We also interpret a learning machine in supervised learning as a student who gives his supervisor a known instance and the supervisor answers with the known feature.

- In unsupervised learning there is no additional label attached to the training data and the task is to describe the structure of data. Density estimation, clustering and dimensionality reduction are examples of unsupervised learning problems.

Regarding a learning machine in unsupervised learning as a student, then the student has to learn by himself without teacher. This learning is harder but happens more often in life. At the current time, unsupervised learning is most exciting part of machine learning where many new deep mathematical methods, e.g., persistent homology and other topological methods, are invented to described the structure of data. Furthermore, many tasks in supervised learning can be reduced to the density estimation problem, which is an unsupervised learning.

- A machine learning task can be performed using supervised learning or unsupervised learning or some intermediate type between supervised learning and unsupervised learning. In anomaly detection, i.e., finding samples that are inconsistent with the rest of the data, unsupervised techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal. Supervised anomaly detection techniques require a data set that has been labeled as normal and abnormal and involves training a classifier.

- Reinforcement learning is the type of machine learning where a learner actively interacts with the environment to achieve maximal reward. In reinforcement learning the learner an agent collects information through a course of actions by interacting with the environment. In reinforcement learning the training data are not labeled, i.e., the supervisor does not directly give answers to the students questions. Instead, the supervisor evaluates the students behavior and gives feedback about it in terms of rewards.

4. Basic mathematical problems in machine learning

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

For this purpose we need to quantify the notion of "success of a learning prediction".

Question 1 How to construct a mathematical model of learning?

In general, this question is as large as mathematics and sciences. In many concrete cases, we have a family of candidate models and a search for the most suitable model is called a model selection.

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

We quantify the difficulty of a problem in terms of its time complexity, which is the minimum time needed for performing computer program to solve a problem, and in terms of its resource complexity which measures the capacity of data storage and energy resource needed to solve the 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?"

Question 3 How to choose a learning algorithm?

Clearly we want to have a best learning algorithm, once we know a model of a machine learning which specifies the set of possible predictors (decisions) and the associated error/reward function .

- It is an optimization problem.

- The notion of success of learning process requires a mathematical treatment of the asymptotic rate of the error/reward.

Conclusion. Machine learning is learning with computer performed algorithms, whose performance improves with increasing volume of empirical data. In machine learning we use probability theory and mathematical statistics to model incomplete information and the random nature of the observed data. To build mathematical models for problems in machine learning we need to specify the type of the problem we want to solve and the type of available data, which result in the type of machine learning model, which we also call the type of machine learning.

One of the main problems of mathematical foundations of machine learning concerns the quantification of the difficulty/complexity of a learning problem. This problem is ultimately related to the question of learnability of a learning problem, i.e., if a learning problem admits a successful learning algorithm.

Other important problems are the problem of formalizing the “meaning” or “information content” of a scientific data, and the challenge of mathematical optimization in the presence of large input sizes.