

Mathematical Theory of Machine Learning

Lecture course, winter 2007/08, Uni Tübingen

Ulrike von Luxburg

Machine learning deals with the question of how we can "teach" a computer to perform specific tasks, simply by presenting examples of the task to the computer. A well-known example is spam detection: we want to train the spam filter such that it can classify emails into the classes "spam" or "not spam". The main goal is that after training the computer is able to generalize its acquired knowledge to new emails: it should not only memorize the classes of the emails we used to train it, but it should be able to classify previously unseen emails correctly.

The field of machine learning deals with the question of how we can devise algorithms to perform such tasks, and what performance guarantees we can give. Very often, machine learning tasks boil down to cleverly estimating functions based on random points. In this lecture, we will introduce the mathematical framework of machine learning problems and discuss basic tasks and algorithms, with a strong focus on the mathematical side:

- What tasks can be performed in general (positive and negative results)?
- What assumptions do we have to make?
- What key properties does a learning algorithm need to satisfy in order to be "successful"?
- What guarantees can we give on the results of certain learning algorithms?
- How many instances of a problem do we have to show to the computer before it can reliably generalize to new examples?

To answer such questions we will need to use methods from several mathematical disciplines, for example probability theory, functional analysis, and linear algebra. No special knowledge beyond Vordiplom is required by the participants, we will derive our tools from scratch.

Machine learning is a very lively discipline which brings together researchers from many fields. Machine learning algorithms are widely used in practice and are used in many industrial applications. For this reason, this lecture can be interesting for both students who want to get some insight into an active field of research and for students who want to learn some applied mathematics which could be relevant for their professional lives.

Topics we are going to cover

Statistical learning theory framework in general:

- Problems we look at: classification, regression, clustering, density estimation, semi-supervised learning, dimensionality reduction
- Assumptions we make: data sampled iid, no assumptions on probability distribution

What properties should a good learning algorithm have?

- correct "limit behavior": different notions of consistency
- good finite sample behavior
- computationally feasible (e.g. convex, not in NP, ...)

Classification:

Some general terms:

- Overfitting
- Loss functions
- Error functions
- Bias-Variance trade-off

General positive results: there exist universally consistent learning algorithms:

- k-nearest neighbor classifiers and the theorem of Stone
- Perceptron and the theorem of Novikov

General negative results:

- No free lunch theorems
- Arbitrarily slow rates of convergence

Generalization bounds:

- Empirical risk minimization
- Capacity of function classes
- Simple concentration inequalities (Hoeffding, McDiarmid)
- Covering number bounds
- VC dimension and corresponding bounds
- if time allows: other concepts for bounding capacities of function classes: large margin, minimal description length, stability, ...

Weak and strong learners and their equivalence

- Definitions
- Equivalence

Kernel Algorithms:

- theory of kernels, positive definite functions, reproducing kernel Hilbert space
- curse of dimensionality, geometry of high-dimensional spaces
- Support vector machines, kernel principal component analysis

Regularization for classification and regression:

- Tikhonov regularization
- Consistency statements
- applications

Clustering:

What is clustering?

- Different frameworks: axiomatic, density level sets, compression, graph theoretic, model-based, information theoretic
- What can we prove about them?

Consistency statements for clustering algorithms:

- k-means
- single linkage
- spectral clustering

Further topics, depending on how much time is left then

Non-parametric density estimation

Graph-based machine learning algorithms and random walks on graphs

Ranking: PageRank and HITS

Approaches based on information theory

Some literature

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

- R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, U.K., 1999.
- L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of non-parametric regression*. Springer, New York, 2002.
- T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, New York, 2001.
- B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.
- V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.