Statistical Learning Theory and Induction

Gilbert Harman Department of Philosophy, Princeton University Princeton, NJ USA

Sanjeev Kulkarni Department of Electrical Engineering, Princeton University Princeton, NJ USA

Synonyms

Statistical learning theory: pattern recognition, pattern classification.
Induction: nondeductive reasoning.

Definition

Induction is here taken to be a kind of reasoning from premises that may not guarantee the truth of the conclusion drawn from those premises. It is to be distinguished from "mathematical induction" which is a kind of deductive reasoning. *The philosophical problem of induction* is whether and how inductive reasoning can be justified.

Statistical learning theory (SLT) is a mathematical theory of a certain type of inductive reasoning involving learning from examples. SLT makes relatively minimal assumptions about an assumed background probability distribution responsible for connections between features of examples and their correct classification, the probability that particular examples will occur, etc. The theory seeks to describe various learning methods and say how well they can be expected to do at producing rules with minimum expected error on new cases.

Among the topics discussed in statistical learning theory are various nearest neighbor rules, related kernel rules, feed-forward neural networks, empirical risk minimization as a kind of inductive generalization, ways of balancing data-coverage against "simplicity," PAC learning, support vector machines, and boosting.

Theoretical background

A valid deductive argument has a kind of conditional reliability: the conclusion is guaranteed to be true if the premises are true. Inductive reasoning typically lacks such a guarantee. The philosophical "problem of induction" asks whether any sort of justification is possible for inductive reasoning.

Clearly there cannot be a deductive demonstration that inductive reasoning has the same sort of conditional reliability as deduction, because then there would be no difference between induction and deduction. But there clearly is a difference.

It might be argued that induction is justified because it works. How do we know it works? Because it has worked in the past—not always, but much of the time? It will be objected that this sort of justification appeals to induction and is therefore circular. (Similarly, a deductive justification of deduction would be circular.) But maybe it is possible to give some sort of deductive justification of induction.

Of course, nothing can be deduced about induction in the absence of substantive assumptions. So the only serious philosophical problem of induction asks what sorts of assumptions yield interesting conclusions.

SLT offers a kind of deductive answer to the problem of induction by characterizing various kinds of induction and then deriving theorems about these versions. While basic results in SLT theory go back to the 1960's, this continues to be an active area and further research is ongoing (e.g., see Devroye, Györfi, Lugosi 1996, Vapnik 2000, Hastie, Tibshirani, and Friedman 2009 and references therein). In addition to offering a response to the philosophical problem, SLT has many practical applications.

SLT applies to inductive pattern recognition, function estimation, as well as other inductive issues. We concentrate here on pattern recognition, which involves using data in order to learn to classify examples on the basis of their features. We also suppose for simplicity that the relevant classification is YES/NO: is this character an "e"? does this picture show a face?

The data consist in a set of labeled examples. Each example has a set of its *features* and a label giving the correct classification of the example. In one basic case it is also assumed that there is a single background statistical probability distribution that characterizes the probabilities that particular instances will arise and the probabilistic relations between features and correct classifications. Furthermore, it is assumed that the same probability distribution is responsible both for the data and for new cases that arise. Minimal assumptions are made about this probability distribution; for example, in the basic case it might be assumed that the relevant probabilities of particular instances are independent of the probabilities of other instances.

Given the unknown probability distribution as the background for a pattern recognition problem, there is a minimum expected error R^* for rules associating labels with features. (A rule with this minimum expected error rate R^* is called a *Bayes Rule*.) It follows immediately that induction cannot arrive at a rule with a smaller expected error than R^* and the question is then how close can one or another inductive rule come to that minimal error rate.

Items are assumed to have some number d of features, each of which can take several, perhaps infinitely many, values. We can identify the possible values of features with real numbers. Then to specify the values of all the relevant features of a particular item is to specify a point in a d dimensional *feature space*.

The simplest classification rule is the 1-nearest neighbor rule. Given n labeled examples, the rule assigns to any new instance the same label as the nearest labeled example. There are also k-nearest neighbor rules which assign to any new instance the label that that a majority of the k nearest neighbors have (assuming k is an odd number). And there are k_n nearest neighbor rules, where the number of nearest neighbors considered is a function of n.

It can be shown that no matter what the background probability distribution, the limit of the the expected error of the 1-nearest rule as $n \to \infty$ is no more than $2R^*$. And, it can be shown that the expected error of a k_n nearest neighbor rule approaches R^* as $n \to \infty$ if $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, no matter what the background probability distribution. So that rule is said to be *universally consistent*. However, there is not *uniform convergence* to R^*, since there will be

probability distributions for which convergence is arbitrarily slow. Here we have two interesting results about this sort of inductive method.

A different method of reasoning, which philosophers sometimes call *enumerative induction*, starts with the selection of a class C of rules and chooses from C a rule with minimum error on the data. (This learning method is also called *empirical risk minimization*.) Vapnik and Chervonenkis (1968) prove that if and only if the rules in C satisfy a certain condition, namely that they have *finite VC-dimension* (which we explain in a moment), then with probability approaching 1 the expected error of the rules endorsed by enumerative induction will converge uniformly to the minimum expected error of rules in C. In other words if, and only if, the VC-dimension of C is finite, this sort of enumerative induction satisfies the PAC (probably approximately correct) criterion.

The *VC dimension* of the collection of rules C is the largest number N such that some set of N points in feature space is *shattered* by rules in C, so that for any possible labeling of those N points, some rule in C fits perfectly the points so labeled. If for any finite number N there exists some set of N points that is shattered by the rules C, then the VC dimension of C is infinite. Subject to some very mild measurability conditions on the rules in C, enumerative induction is an instance of PAC learning if and only if the VC dimension of rules in C is finite.

Learning using standard feed-forward artificial neural networks satisfies the PAC criterion, since the rules representable by such a network with fixed architecture and variable weights have finite VC-dimension.

A different inductive learning method, sometimes called *structural risk minimization* balances empirical adequacy of the data against some other feature of rules in C, sometimes mistakenly identified with *simplicity*. Let $C = C_1 \cup C_2 \cup \cdots C_n \cup \cdots$, where $C_i \in C_{i+1}$, where the VC dimension of each C_i is finite and the VC dimension of C is infinite. A particular version of structural risk minimization selects a rule r in C that minimizes a function $f(e, i)$ of the error e on the data and the least i such that r is in C_i. Something like this fits much scientific practice but cannot satisfy the PAC criterion (since the VC dimension of the rules C is infinite). Under certain conditions it does allow universal consistency.

So, SLT sheds light on a variety of inductive methods by specifying various assumption different kinds and proving that particular inductive methods have different desirable properties under the appropriate assumptions. This is discussed further in Harman and Kulkarni (2007).

Important Scientific Research and Open Questions

The field of statistical learning theory is quite active. Some fairly recent developments include support vector machines (Vapnik 2006) and Boosting (Schapire 1999). There are many extensions and generalizations of the basic results in which the different assumptions of the theory are relaxed, stronger results are obtained, more detailed analysis is provided, or different settings are considered.

Cross-References Bayesian learning, Connectionist theories of learning, Learning in artificial neural networks, Nonparametric statistics, PAC learning, Probability theory in machine learning, Statistical learning, Statistical learning techniques.

References

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

Hastie, T., Tibshirani, R., and Friedman, J., (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, New York: Springer.

Harman, G., and Kulkarni, S., (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge, MA: MIT Press.

Schapire, R. E., (1999) "A brief introduction to boosting," *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*.

Vapnik, V., (2000). *The Nature of Statistical Learning Theory*, 2nd edition, New York: Springer.

Vapnik, V., (2006). "Empirical Inference Science," in *Estimation of Dependencies Based on Empirical Data*, second edition, Springer.

Vapnik, V., and Chervonenkis, A. Ja., (1968). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." (In Russian.) *Doklady Akademii Nauk USSR* 181, translated into English in *Theory of Probability and Its Applications* 16 (1971): 264-280.