PRECIS OF "RELIABLE REASONING: INDUCTION AND STATISTICAL LEARNING THEORY" (MIT Press, 2007)

Gilbert Harman & Sanjeev Kulkarni

Reliable Reasoning seeks to show how results in basic statistical learning theory bear on issues in philosophy and psychology.

One is the ancient philosophical problem of induction. The problem is sometimes (misleadingly) motivated through a comparison of induction with deduction. Deduction is said to be perfectly reliable in the sense that the truth of the premises in a deduction guarantees the truth of the conclusion. This is said typically not to be the case for induction. Induction can lead from truths to falsehoods.

The comparison with deduction is of course misleading because a deductive relation can hold between premises and a conclusion even though one cannot reasonably infer that conclusion from those premises, for various reasons. Sometimes good reasoning leads one to abandon a premise rather than accept a conclusion. More generally, logic—the theory of deduction—is not a theory of inference.

Can any inductive methods be justified? It is sometimes said that the justification of methods of inference can only consist in adjusting those methods and one's various beliefs with each other until one reaches a *reflective equilibrium*. Although there is evidence that people do adjust methods and opinions in this way, there is also considerable evidence that the results are fragile and unreliable.

A better idea is that the justification of inductive methods must involve finding a way to assess the reliability of inductive methods, where reliability has to do with the statistical likelihood that the conclusions these methods lead to are correct. (It is hard to be in reflective equilibrium if one cannot believe one's methods of reasoning are reliable in this sense.)

Statistical learning theory is precisely concerned with assessing the reliability of inductive methods. How can data be used so as to arrive at a reliable rule for classifying new cases on the basis of certain values of observable *features* of those new cases?

In order to provide a formal mathematical theory, statistical learning theory appeals to the notion of a *D*-dimensional *feature space* in which each point represents a possible set of values of observable features. This framework assumes that an unknown probability distribution characterizes encounters with objects and the correlations between feature values of objects and their correct classifications. The unknown probability distribution determines the unknown best rule of classification, namely the Bayes Rule that minimizes expected error. In the simplest case, the same unknown probability distribution applies to the data as well as the new cases to be classified. It is also assumed in the simplest case that the probability of getting a particular object with such and such features and such and such a classification is independent of what other features and objects have or will occur.

For the case of a YES/NO classification, a classification rule can be identified with a set of points in feature space, perhaps certain disjoint areas or hyper-volumes, indicating which points in feature space are to be labeled YES and which are to be labeled NO. For example, linear rules divide the space into two regions separated by a line or plane or hyperplane.

Given a set *C* of rules, *enumerative induction* endorses a rule from *C* that minimizes error on the data. Enumerative induction makes sense only if there are significant limits on the rules included in *C*. Without such *inductive bias*, enumerative induction allows all possible inferences about new cases. On the other hand, if there are significant limits on the rules included in *C*, then, given enough data, it is highly probable that enumerative induction will endorse a rule whose expected error is close to the minimum error for rules in *C* (Vapnik and Chervonenkis, 1968).

Vapnik and Chervonenkis (1968) show that (subject to some very mild conditions) *no matter what the background probability distribution*, with probability approaching 1, as more and more data are considered, the expected error of the rules that enumerative induction endorses will approach the minimum expected error of rules in *C*, *if and only if C* has a finite index that has acquired the name, *VC dimension*.

VC dimension is explained in terms of *shattering*. Rules in *C* shatter a set of *N* data points if and only if, for every possible labeling of the *N* points with YESes and NOs, there is a rule in *C* that perfectly fits that labeling. In other words, there is no way to label those *N* points in a way that would falsify the claim that the rules in *C* are perfectly adequate.

The VC dimension of *C* is the largest number *N* such that some set of *N* points in the feature space can be shattered by rules in *C*. If for any *N*, there is a set of points that is shattered by rules in *C*, the VC dimension of *C* is infinite.

Vapnik and Chervonenkis show (again under very mild conditions) that, if and only if the set of rules *C* has finite VC dimension, expected error from the use of enumerative induction *uniformly converges* to the minimum expected error of rules in *C*. This means that it is possible to calculate for any given ε and δ, how much data are needed (no matter what the probability distribution) so that, with probability $1 - \varepsilon$, the difference between the expected error and the minimum expected error of rules in *C* is less than δ.

This does not mean that the expected error of enumerative induction converges to that of a best possible rule, a Bayes rule, because the least expected error of rules in *C* may be greater than the expected error of a Bayes rule, perhaps much greater.

Enumerative induction uses data to choose a rule from *C* entirely on the basis of the empirical adequacy of the rule with respect to that data. There are alternative inductive methods that in choosing a rule from *C* balance such empirical adequacy against something else—some sort of simplicity perhaps.

What Vapnik calls *structural risk minimization* is an example of this second sort of inductive method. In structural risk minimization, the class *C* of rules is an infinite union of subclasses of increasing VC dimension. (The VC dimension of *C* is therefore infinite, so that enumerative induction may perform poorly.) For each rule in *C* let *m* be the number of the smallest subclass of *C* to which the rule belongs. Then structural risk minimization chooses a rule by balancing that number *m* against the rule's fit to the data.

Structural risk minimization allows *C* to be chosen in such a way that it contains rules whose expected error is arbitrarily close to the expected error of a Bayes rule. Furthermore, it can be shown that the method of structural risk minimization is *universally consistent* in the sense that the expected error of rules endorsed by such methods does approach in the limit the expected error of a Bayes rule. But since the class *C* used in structural risk minimization has infinite VC dimension, we get no guarantee of uniform convergence to the expected error of the Bayes rule.

VC dimension is defined in terms of falsifiability in a way that is reminiscent of Popper. The VC dimension of a class of rules is at least *n* if there is a set of *n* points with the property that no assignment of labels to those points could falsify the hypothesis that some rule in *C* fits the data. Popper similarly measures the simplicity of a class of rules in terms of the amount of data needed to falsify the hypothesis that the correct rule is in that class. But there is a difference. And VC dimension turns out to be the more important characteristic of a class of rules than what might be called its "Popper dimension."

There are other universally consistent inductive methods. Certain *nearest neighbor* methods provide one example. There are also inductive methods that take into account not just the empirical adequacy of a rule but also the place of the rule in a well-ordering of the rules in *C*, e.g. by the length of its minimal description in a specified format.

In a final chapter, we explain how the ideas we have taken from statistical learning theory apply to the analysis of perceptrons and feed-forward neural networks that philosophers and psychologists often discuss as providing approximate models of aspects of the brain. We then go on to discuss a different model involving *support vector machines* (SVMs).

In effect, SVMs map the original feature space into a typically much higher dimensional space (sometimes even an infinite dimensional space) in which images of the data are linearly separated. We consider the possibility that SVMs might provide a useful model of psychological categorization that would compete with currently standard models.

We also discuss what Vapnik calls "transduction," a learning method that uses additional information beyond the labeled data—information about hard cases and about what cases have come up to be classified. The theory of transduction suggests new models of how people sometimes reason. The hypothesis that people sometimes reason transductively provides a possible explanation of some psychological data that have been interpreted as showing that people are inconsistent or irrational. It also provides a possible account of a certain sort of "moral particularism."

Gilbert Harman & Sanjeev Kulkarni

Princeton University

harman@princeton.edu / kulkarni@princeton.edu