Statistical Learning Theory as a Framework for the Philosophy of Induction

Gilbert Harman and Sanjeev Kulkarni
Princeton University

December 16, 2008

Statistical Learning Theory (e.g., Hastie et al. 2001; Vapnik 1998, 2000, 2006; Devroye, Györfi, Lugosi 1996) is the basic theory behind contemporary machine learning and pattern recognition. We suggest that the theory provides an excellent framework for the philosophy of induction (see also Harman and Kulkarni 2007).

Inductive reasons are often compared with deductive reasons. Deductive reasons for a conclusion guarantee the conclusion in the sense that the truth of the reasons guarantees the truth of the conclusion. Not so for inductive reasons, which typically do not provide the same sort of guarantee. One part of the philosophy of induction is concerned with saying what guarantees there are for various inductive methods.

There are various paradigmatic approaches to specifying the problem of induction.

For example, Reichenbach (1949) argued, roughly, that induction works in the long run if anything works in the long run. His proposal has been followed up in interesting ways in the learning in the limit literature (E.g. Putnam, 1963, Osherson et al., 1982, Kelly, 1996, Schulte, 2002). The paradigm here is to envision a potentially infinite data stream of labeled items, a question Q about that stream, and a method M that proposes an answer to Q given each finite initial sequence of the data. If the data sequence consists of a series of letters of the alphabet, one question might be whether every “A” in the sequence is followed by a “B”; then the issue is whether there is a method for answering that question after each datum, a method that will eventually give the correct answer from that point on.

A second paradigm assumes one has an initial *known subjective* probability distribution satisfying certain more or less weak conditions along with a method for updating one’s probabilities, e.g. by conditionalization, and proves theorems about the results of such a method. (Savage 1954, Jeffrey 2005).

Statistical learning theory, which is our topic, represents a third paradigm which assumes there is an *unknown objective* probability distribution that characterizes the data and the new cases about which inferences are to be made, the goal being to do as well as possible in characterizing the new cases in terms of that unknown objective probability distribution. The basic theory attempts to specify what can be proved about various methods for using data to reach conclusions about new cases.

We will be concerned with what we take to be basic statistical learning theory, which is concerned with what can be proved about various inductive methods, given minimal assumptions about the background probability distribution. We are interested in results that hold no matter what that background probability distribution is (as long as the minimal assumptions are satisfied). In other words, we are interested in “worse case” results: even in the worst case, such and such holds.

We begin by sketching certain aspects of basic statistical learning theory, then comment briefly on philosophical implications for reliability theories of justification, Popper’s (1979, 2002) appeal to falsifiability, simplicity as relevant to inductive inference, and whether direct induction is an instance of inference to the best explanation.

Pattern Recognition

One basic problem discussed in statistical learning theory is the *pattern recognition problem*: "How can data be used to find good rules for classifying new cases on the basis of the values of certain features of those cases?" As we have indicated, the simplest version of the problem presupposes that there is an unknown statistical probability distribution that specifies probabilistic relations between feature values of each possible case and its correct classification and also specifies how likely various cases are to come up either as data or as new cases to be classified. In this simplest version, these probabilities are assumed to be identically distributed and are independent of each other, but no other assumption about the probability distribution is made.

For example, the Post Office wants to use machines to sort envelopes on the basis of handwritten zip codes. The data are samples of actual handwritten zip codes as they have appeared on envelopes that have been received by the Post Office, samples that have been classified by human operators as representing one or another zip code (or as not representing a zip code). In a standard version of the problem the handwritten cases are digitized and presented as an N by M grid of light intensities, so there are $N \times M$ features, the value of each of which is specified as an intensity of light in the corresponding pixel of the grid. A rule for classifying new cases maps each possible value of the $N \times M$ features into a particular zip code (or into a decision that the features shown do not represent a zip code).

The feature values of a possible case can be represented as a vector $\bar{x} = (x_1, x_2, \ldots, x_D)$ or equivalently as a point in a D-dimensional feature space whose coordinates are specified by $\bar{x}$. The ith co-ordinate of a point in the feature space corresponds to the value of the ith feature for an item represented by that point. A pattern recognition rule can be thought of as a rule for assigning labels to each point in the feature space, where the label assigned to the point represents the rule's verdict for a case with those features.

In order to evaluate pattern recognition rules, an assumption is needed about the value of getting the right answer and the cost of getting the wrong answer. In many applications, the value of any right answer is taken to be zero, and the cost of any wrong answer is set at one. In other cases, different values and costs might be assigned. (For example, in some medical diagnostic contexts, a false negative verdict might be assigned a greater cost than a false positive verdict.)

Pattern recognition rules are then better or worse depending on the expected value of using them, where expectations are determined by the values and costs of right and wrong answers, the probabilities of the various cases, and the probabilities that the answer given by the rules are correct for each of those cases.

If the value of right answers is set at zero, the best pattern recognition rules have the least expected costs. If the cost of a wrong answer is always 1, then the best pattern recognition rules have the least expected error.

The Post Office may want to assign different values and costs to correct or incorrect decisions about various cases, but on the other hand may be satisfied (at least at the beginning) to try to find rules with the least expected error.

The best pattern recognition rules for a given problem are those rules with the least expected cost. Such a rule is called a *Bayes rule*. There is always at least one such rule. (There is more than one Bayes rule only if for at least one possible case the expected value of two different decisions about the case are tied.)

Without loss of generality, we might assume that we are only concerned with the expected error of a rule. So, in the rest of our discussion we suppose we are interested in minimizing the expected error of decisions about new cases.

If one is concerned with *yes/no* issues, so that each item is either an instance of a certain category or not, we can represent a *yes* classification as 1 and a *no* classification as 0. In this case, a pattern recognition rule is equivalent to a specification of the set of points in the feature space that the rule classifies as *yes* or 1. In the post office example, one might be concerned with whether a given written figure should be classified as a “9”(yes=1) or not (no=0).

Bayes Error Rate R*

To review: We suppose each item has D features $x_1, x_2, \ldots, x_D$, where each feature x_i takes real values . (In the Post Office example, there are $N \times M$ features, each of which might take any of as many as 256 intensity values.) We represent the values of the features of a given item with the feature vector $\bar{x} = (x_1, x_2, \ldots, x_D)$. We can also represent the values of the features of an item as a point in a D-dimensional *feature space*.

In our simplified case, a pattern recognition rule maps each feature vector $\bar{x}$ into a *yes/no* decision. We use 1 for *yes* and 0 for *no*. The rule maps points in the feature space into 1 and 0. We can specify the rule by specifying the set of points that it maps into 1.

We suppose there is a statistical probability distribution P specifying the actual statistical probability relations among feature values and the correct classification of items whose features have those values and also specifying how likely items with those feature values are to come up.

A rule with minimal expected error, i.e. a Bayes rule, maps $\bar{x}$ into 1, if $P(1|\bar{x}) > P(0|\bar{x})$, and maps $\bar{x}$ into 0, if $P(1|\bar{x}) < P(0|\bar{x})$. (It does not matter what a Bayes rule maps $\bar{x}$ into when these conditional probabilities are equal.)

Applying this rule to a fixed observed $\bar{x}$, the probability of error is the smaller of $P(0|\bar{x})$ and $P(1|\bar{x})$, i.e., $\min\{P(0|\bar{x}), P(1|\bar{x})\}$. Therefore, the overall probability of error of a Bayes decision rule, the *Bayes error rate*, denoted by R^* is

$$R^* = \sum_{\bar{x}} P(\bar{x}) \min\{P(0|\bar{x}), P(1|\bar{x})\}$$

(In cases involving probability densities, as specified by a probability density distribution p, the Bayes error rate is given by

$$R^* = \int \min\{P(0|\bar{x}), P(1|\bar{x})\}\, p(\bar{x})\, d\bar{x}$$

)

Using Data to Learn the Statistical Probability Distribution?

The best pattern recognition rule is determined by the statistical probability distribution P. We assume that this distribution is initially unknown. In this case then, we need to use data in order to come up with a good rule.

Suppose that various cases arise with known values of various observable features and that an “expert” tells us the correct classification of these cases. We want to be able to use the data, once we have enough of it, to find a rule that will do well on new cases (as assessed by the same “expert”) given their feature values with performance close to that of the Bayes rule.

An idea that doesn’t usually work is to use such data to learn the probability distribution P, which is then used to find the Bayes rule. The idea would be to appeal to the probabilistic “law of large numbers,” which says that in the long run the observed frequency with which an event $\bar{x}$ is correctly classified as 1 will converge to the statistical probability of its being correctly classified 1. In other words, with probability approaching 1, for any ϵ however small, there will be a t such that after t occurrences of the event, the frequency of the event’s being 1 will be within ϵ of its statistical probability of being 1.

If there are finitely many points in the feature space, each representing an event with the features $\bar{x}$, then we might consider labeled items as they come to our attention and note the frequencies with which each given $\bar{x}$ is a 1. In each case, the frequency that a given $\bar{x}$ is found to be a 1 will converge to its statistical probability of being 1. Because there are only finitely many such events, there will be *uniform convergence* of these frequencies. In other words, with probability approaching 1, for any ϵ however small, there will be a t such that after t items of data, the frequency of each $\bar{x}$ being 1 will be within ϵ of its probability of being 1.

Unfortunately, this can take a very long time. For example, in Post Office case, suppose there the grid of pixels were only 10 by 100. Then if there were 256 possible intensity values, there would be 256^{1000} different feature vectors. Even if there were only 2 possible intensity values, *on* and *off*, there would still be $2^{1000} > 10^{300}$ different feature vectors. Given that the current age of the universe is less than 10^{20} seconds, there would not be time for even a tiny fraction of those feature vectors to show up in anyone's lifetime!

We need a method that does not require having to learn the whole probability distribution. Statistical learning theory is the theory of such methods.

Empirical Risk Minimization

What is a good way to choose a rule of pattern recognition that will maximize expected value, or in our special case, to minimize expected error? One initially appealing idea might be simply to choose some rule or other that has the least error on cases that are included in the data. But too many rules have that property. For any given rule, there will be other rules making the same decisions on cases in the data but making all possible decisions on cases not part of the data.

Any workable inductive procedure must therefore have some sort of *inductive bias*. It must favor certain rules of pattern recognition over others that make exactly the same decisions for cases in the data.

The crudest inductive bias simply restricts the relevant pattern recognition rules to a selected class of rules C. Then a policy of *empirical risk minimization*, given some data, selects a rule from C with minimal cost on the data. But of course not all choices of the class of rules C are equally good. As we have just observed, if C contains all possible rules, a policy of empirical risk minimization fails to give any advice at all about new cases (or amounts to choosing classifications of new cases at random).

On the other hand, if C does not contain all rules, it may fail to contain a Bayes rule—a best rule of pattern classification in this instance. Indeed, it may fail to contain anything whose performance is close to a Bayes rule. It might even contain only the worst possible rules!

But let us put this last problem to the side for a moment. We will come back to it shortly.

Consider this question: what has to be true of the set C of rules so that, no matter what the unknown background probability distribution, empirical risk minimization eventually does as well as possible with respect to the rules in C? More precisely, what has to be true of the set C in order to guarantee that, with probability approaching 1, *no matter what the unknown probability distribution*, given more and more data, the expected error for the rules that empirical risk minimization endorses at each stage eventually approaches the minimum value of expected error of rules in C?

A fundamental result of statistical learning theory is that the set of rules in C cannot be too rich, where the richness of C is measured by its *VC-dimension.* Let us explain.

Shattering and VC-dimension

Recall that we are restricting attention to rules that map a set of D feature values into a *yes*/*no* verdict for a particular categorization, where we use 1 for *yes* and 0 for *no.* As we have mentioned, we can represent such a rule in terms of a D dimensional *feature space* in which each point is labeled 1 or 0. Each point in that space represents a possible set of features. The ith co-ordinate of a point in the feature space corresponds to the value of the ith feature for that point. The label attached to the point represents that rule's verdict for a case with those features.

Now consider a set of N points in a given feature space and consider the 2^N possible ways to label those points as *yes* or *no.* If for each possible way of labeling those points there is a rule in C that agrees with that labeling, we say that the rules in C *shatter* those N points.

To say that the rules in C shatter a particular set of points is to say that no possible assignment of verdicts to those points (those possible cases) "falsifies" the claim that some rule in C correctly represents those verdicts, to use terminology from Popper (1979, 2002).

The finite VC-dimension of a set of rules C is the largest finite number N for which some set of N points is shattered by rules in C. If C has no finite VC-dimension, its VC-dimension is infinite.

Fundamental Result

Recall that empirical risk minimization says to select a rule in C with least error on the data. Vapnik and Chervonenkis (1968) show that empirical risk minimization works, no matter what the unknown statistical probability distribution is, if, and only if, C has finite VC dimension. More precisely:

> If and only if C has finite VC dimension: with probability approaching 1, then, no matter what the unknown probability distribution, given more and more data, the expected error for the rules that empirical risk minimization endorses at each stage eventually approaches the minimum value of expected error of rules in C.

The Vapnik Chervonenkis result also provides information about how much data are needed for empirical risk minimization to produce a good result no matter what the unknown statistical probability distribution is. If the rules in C have VC dimension V, there will be a function $m(V, \epsilon, \delta)$ indicating the maximum amount of data needed (no matter what the unknown statistical probability distribution) to ensure that the probability is less than δ that enumerative induction will endorse a hypothesis with an expected error rate that exceeds the minimum expected error rate for rules in C by more than ϵ. (A smaller ϵ indicates a better approximation to the minimum error error for rules in C and a smaller δ indicates a higher probability that the rules endorsed will be within the desired approximation to that minimum expected error.)

Where there is such a function $m(V, \epsilon, \delta)$ there is what has come to be called "Probably Approximately Correct" (or "PAC") learning (terminology due to Valiant, 1984).

Example: Perceptron Learning

A *perceptron* is a simple classifier that takes the weighted sum of the D input feature values (along with an imaginary additional constant input value) and outputs $+1$ for *yes* if the result of the weighted sum is greater than some threshold T and outputs 0 for *no* otherwise. Given data, it is easy to find a threshold and weights for such a perceptron that yield the least error (or cost) on that data.

Any classification rule that can be represented by such a perceptron divides the D dimensional feature space into two regions, the *yes* region and the *no* region, where the regions are separated by a line, plane, or hyper-plane. In other words, a perceptron classifier *linearly separates* the points in the feature space. The class C in this case is the class of linear separation rules in that feature space.

The VC-dimension of such linear separations of a D dimensional feature space is $D + 1$, which is finite, so the the result mentioned in the previous subsection applies. We can know how many items of data are needed in order probably to approximate the best such separation.

Here we can return to the worry we temporarily put aside, because it applies to perceptron learning. The worry is that the class C of rules used for empirical risk minimization may fail to include the best rule, the Bayes rule and may indeed contain no rule whose performance is even close to that of the Bayes rule.

Obviously, many possible classification rules are not and cannot be approximated by linear separation rules. The XOR rule is a well known example. Suppose there are two features, x_1 and x_2, each of which takes a real value between -1 and $+1$, and the correct classification rule, which is also the Bayes rule, takes the value 1 if and only if $x_1 \times x_2 < 0$. Here there is a two dimensional feature space in which the points in the upper left quadrant and the lower right quadrant are to be labeled 1 and the points in the upper right quadrant and lower left quadrant are to be labeled 0. Clearly there is no way to draw a line separating the points to be labeled 1 from those to be labeled 0. Indeed, if the statistical probability density is evenly distributed over the feature space, any linear separation has a significant expected error while the Bayes rule has an expected error of 0!

Example: Feed-forward Neural Network Learning

This last worry can be somewhat alleviated by using a feed-forward neural network with several layers of perceptrons. Inputs go to perceptrons in the first layer whose outputs are inputs to perceptrons in the second layer, and so on to a final perceptron that outputs 0 or 1 depending on whether the weighted sum of its inputs are above or below a certain threshold. A fixed network with fixed connection strengths between nodes and fixed thresholds can be used to classify inputs as 0 or 1, so such a network therefore represents a classification rule. As before, varying connection strengths and thresholds (while retaining all connections) yields other rules.

There are more or less good methods for finding a rule represented by a given structure of perceptrons that has least error on given data. These learning methods typically involve smoothing the threshold functions of all but the final perceptron and use a kind of "gradient descent" that may or may not become stuck in a "local minimum" that is not a global minimum. In any case there will be one or more setting of connection strengths and thresholds that minimizes error on the data.

It can be shown that, the set of all rules represented by possible ways of varying connection strengths and thresholds for any particular feed forward network has a finite VC-dimension. So, the PAC learning criterion is satisfied. Furthermore, any (nonpathological) classification rule can be approximated arbitrarily closely by feed forward neural networks with enough layers by adding more nodes. So it is possible to ensure that the error rate of the best rule that can be represented by such a network is not far from the error rate of the best possible rule, the Bayes rule.

Of course, adding nodes increases the VC dimension of such a network, which means more data will be needed to guarantee a certain level of performance under the PAC criterion. So, there is a trade-off between the amount of data needed to satisfy the PAC criterion and how closely the error rate of the best rule in C is to the Bayes error rate.

Furthermore, no matter how high the VC dimension of C, as long as it is finite, there is no guarantee that, no matter what the background statistical probability distribution, the error rate of the rule selected via empirical risk minimization will converge to the Bayes error.

Data Coverage Balanced Against Something Else

Instead of using pure empirical risk minimization, an alternative learning procedure balances empirical error or cost on the data against something else—often called *simplicity*, although (as we will see) that is not always a good name for the relevant consideration—and then allows C to have infinite VC dimension.

There are versions of this strategy which, with probability approaching 1, will eventually come up with rules from C whose expected error approaches that of the Bayes rule in the limit. However, because the rules in C have infinite VC dimension, the PAC result does not hold. So "eventually" might be a very long time.

One version of this alternative strategy is concerned with rules that can be specified in a particular notation. Each rule in C is assigned a number, its length as measured by the number of symbols used in its shortest description in that notation. Given data, the procedure is to select a rule for which (for example) the sum of its empirical error (or cost) on the data plus its length is minimal.

Another version identifies C with the union of a nested series of classes each of finite VC-dimension $C_1 \subset C_2 \subset \cdots C_n \subset \cdots$, where the VC-dimension of C_i is less than the VC-dimension of C_{i+1}. Given data, the procedure is then to select a rule that minimizes the sum of its empirical error (or cost) on the data plus the number i of the smallest class C_i to which the rule belongs. This version is called *structural risk minimization.*

These two kinds of ordering can be quite different. If rules are ordered by description length, linear separations will be scattered throughout the ordering, because the length of a description of a linear rule will depend on the number of symbols needed to specify various constants in those rules. In the ordering of classes of rules by VC-dimension, linear separations can be put ahead of quadratic separations, for example, because linear separations in a given space have a lower VC-dimension than quadratic separations in the same space.

We return later to the of question whether any approach of this sort is best described as balancing *simplicity* against data-coverage.

Example: Support Vector Machines

The final perceptron in a feed forward neural network makes a linear decision in the space represented by its inputs. The earlier parts of the feed forward network can therefore be thought of as mapping the feature space into the space for which the final perceptron makes a linear decision.

This suggests a more general strategy for solving pattern recognition problems. Map the feature space into another space—perhaps a space with many more dimensions—and then make a linear decision in that space that minimizes error on the data. It is true that the more dimensions to the other space, the higher the VC-dimension of the rules represented by linear separations in that space. However, *support vector machines* get around this problem by using *wide margin separations* instead of just any linear separation.

Linear separations are hyper-planes that have no thickness. Wide margin separations are thick hyperplanes, hyperslabs. Vapnik (2006) observes that, if the relevant points in a space are confined to a hypersphere of a given size, then the VC-dimension of wide margin separations of those points are often much less than the VC-dimension of pure linear separations.

Even if the space in question is an infinite dimensional Hilbert space, the VC-dimension of wide-margin separations of points in that hyper-volume is finite and inversely related to the size of the margin, the thickness of the hyper-slabs.

Support vector machines first map the feature space to a very large or infinite dimensional space in which images of points in the feature space are confined to a hyper volume of fixed radius. Then a wide margin separation of points in the transformed space is selected by trading off empirical error on the data against the width of margins of the best wide margin separations.

Transduction

The learning methods discussed so far use labeled data to find a rule that is then used to classify new cases as they arise. Furthermore, these methods all involve learning total classifications. Nearest neighbor methods, perceptrons, multi-layer feed-forward networks, and standard support vector machines all yield rules that assign a classification to every possible set of features.

We could modify some of these methods to provide only partial classifications. For example, we could modify support vector machines so as not to choose among the various separating hyperplanes internal to the selected separating hyperslab. The points in this between space would be left unclassified. The system would still be an inductive method, since it would classify some, perhaps many, new cases in accordance with a rule derived from labeled data, but the rule would not be a total rule, since it would not characterize all points in the space.

Suppose we are using a method that in this way provides only a partial classification of cases and a case arises to be classified in the intervening space of previously unclassified cases. Vapnik (1998, 2000, 2006) considers certain *transductive* methods for classifying such new cases, methods that use information about what new cases have come up to be classified and then select a subset of separations that (a) correctly classify the data and (b) agree on their classifications of the new cases. In one version, the selected separations also (c) disagree as much as possible on the classifications of other possible cases.

An important related version of transduction uses not only the information that certain new cases have come up to be classified but also the information that there is a certain set U ("universum") of examples that are hard to classify. In this version, transduction selects the subset of linear separations satisfying (a) and (b) but disagreeing as much as possible on the classification of the hard cases in U. Transduction performs considerably better than other methods in certain difficult real-life situations involving high-dimensional feature spaces where there is relatively little data.

Philosophical Implications

We conclude by briefly mentioning a few ways in which statistical learning theory has philosophical implications. It is relevant to reliability theories of epistemic justification, it helps to situate some of Popper's often misunderstood appeals to falsifiability, it casts doubt on claims about the importance of simplicity considerations in inductive reasoning, and it illuminates discussion of direct inductive inference.

Reliability

Many philosophical epistemologists (e.g., Goldman, 1986, 2002; Bishop and Trout, 2005) argue that epistemology should be concerned with the *reliability* of methods of belief formation (or more generally of belief revision). We believe that the relevant notion of reliability makes sense only in relation to an appeal to something like the sort of background statistical probability distribution that figures in the pattern recognition problem studied in statistical learning theory. If so, statistical learning theory provides one sort of foundation for philosophical epistemology.

VC Dimension and Popperian Falsifiability

There is an interesting relation between the role of VC dimension in the PAC result and the emphasis on falsifiability in Karl Popper's writings in the philosophy of science. Popper (1934) famously argues that the difference between scientific hypotheses and metaphysical hypotheses is that scientific hypotheses are "falsifiable" in a way that metaphysical hypotheses are not. To say that a certain hypothesis is falsifiable is to say that there is possible evidence that would not count as consistent with the hypothesis.

According to Popper, evidence cannot establish a scientific hypothesis, it can only "falsify" it. A scientific hypothesis is therefore a falsifiable *conjecture.* A useful scientific hypothesis is a falsifiable hypothesis that has withstood empirical testing.

Recall that enumerative induction requires a choice of a set of rules C. That choice involves a "conjecture" that the relevant rules are the rules in C. If this conjecture is to count as scientific rather than metaphysical, according to Popper, the class of rules C must be appropriately "falsifiable."

Many discussions of Popper treat his notion of falsifiability as an all or nothing matter, not a matter of degree. But in fact Popper does allow for degrees of difficulty of falsifiability (2002, sections 31-40). For example, he asserts that a linear hypothesis is more falsifiable—easier to falsify—than a quadratic hypothesis. This fits with VC theory, because the collection of linear classification rules has a lower VC dimension than the collection of quadratic classification rules.

However, Popper's measure of degree of difficulty of falsifiability of a class of hypotheses does not quite correspond to VC-dimension (Corfield et al, 2005). Where the VC-dimension of a class C of hypotheses is the largest number N such that *some* set of N points is shattered by rules in C, what we might call the "Popper dimension" of the difficulty of falsifiability of a class is the largest number N such that *every* set of N points is shattered by rules in C. This difference between *some* and *every* is important and VC-dimension turns out to be the key notion rather than Popper-dimension.

Popper also assumes that the falsifiability of a class of hypotheses is a function of the number of parameters used to pick out instances of the class. This turns out not to be correct either for Popper dimension or VC dimension, as discussed below.

This suggests that Popper's theory of falsifiability would be improved by adopting VC-dimension as the relevant measure in place of his own measure.

Simplicity

We now want to say something more about Popper's (1972, 2002) discussion of scientific method. Popper argues that there is no justification for any sort of inductive reasoning, but he does think there are justified scientific methods.

In particular, he argues that a version of structural risk minimization best captures actual scientific method (although of course he does not use the term "structural risk minimization"). In his view, scientists accept a certain ordering of classes of hypotheses, an ordering based on the number of *parameters* needing to be specified to be able to pick out a particular member of the class. So, for example, for real value estimation on the basis of one feature, linear hypotheses of the form $y = ax + b$ have two parameters, a and b, quadratic hypotheses of the form $y = ax^2 + bx + c$ have three parameters, a, b, and c, and so forth. So, linear hypotheses are ordered before quadratic hypotheses, and so forth.

Popper takes this ordering to be based on "falsifiability" in the sense at least three data points are needed to "falsify" a claim that the relevant function is linear, at least four are needed to "falsify" the claim that the relevant function is quadratic, and so forth.

In Popper's somewhat misleading terminology, data "falsify" a hypothesis by being inconsistent with it, so that the hypothesis has positive empirical error on the data. He recognizes, however, that actual data do not show that a hypothesis is false, because the data themselves might be noisy and so not strictly speaking correct.

Popper takes the ordering of classes of hypotheses in terms of parameters to be an ordering in terms of "simplicity" in one important sense of that term. So, he takes it that scientists balance data-coverage against simplicity, where simplicity is measured by "falsifiability" (Popper 1934, section 43).

We can distinguish several claims here.

(1) Hypothesis choice requires an ordering of nested classes of hypotheses.

(2) This ordering represents the degree of "falsifiability" of a given class of hypotheses.

(3) Classes are ordered in accordance with the number of parameters whose values need to be specified in order to pick out specific hypotheses.

(4) The ordering ranks *simpler* hypotheses before more *complex* hypotheses.

Claim (1) is also part of structural risk minimization. Claim (2) is similar to the appeal to VC dimension in structural risk minimization, except that Popper's degree of falsifiability does not coincide with VC dimension, as noted in above. As we will see in a moment, claim (3) is inadequate and, interpreted as Popper interprets it, it is incompatible with (2) and with structural risk minimization. Claim (4) is at best terminological and may just be wrong.

Claim (3) is inadequate because there can be many ways to specify the same class of hypotheses, using different numbers of parameters. For example, linear hypotheses in the plane might be represented as instances of $abx + cd$, with four parameters instead of two. Alternatively, notice that it is possible to code a pair of real numbers a, b as a single real number c, so that a and b can be recovered from c. That is, there are functions such that $f(a, b) = c$, where $f_1(c) = a$ and $f_2(c) = b$.[1] Given such a coding, we can represent linear hypotheses as $f_1(c)x + f_2(c)$ using only the one parameter c. In fact, for any class of hypotheses that can be represented using P parameters, there is another way to represent the same class of hypotheses using only 1 parameter.

Perhaps Popper means claim (3) to apply to some ordinary or preferred way of representing classes in terms of parameters, so that the representations using the above coding functions do not count. But even if we use ordinary representations, claim (3) conflicts with claim (2) and with structural risk minimization.

To see this, consider the class of sine curves $y = a + \sin(bx)$ that might be used to separate points in a one dimensional feature space, represented by the points on a line between 0 and 1. Any set of n distinct points in this line segment are shattered by curves from that class. So this class of

[1] For example, f might take the decimal representations of a and b and interleave them to get c.

sine curves has infinite "falsifiability" in Popper's sense (and infinite VC-dimension) even though only two parameters have to be specified to determine a particular member of the set, using the sort of representation Popper envisioned. Popper himself did not realize this and explicitly treats the class of sine curves as relatively simple in the relevant respect (1934, Section 44).

The fact that this class of sine curves has infinite VC dimension (as well as infinite falsifiability in Popper's sense) is some evidence that the relevant ordering of hypotheses for scientific hypothesis acceptance is not a simplicity ordering, at least if sine curves count as "simple".

Transduction as Direct Induction

Vapnik (2000, p. 293) says that transduction does not involve first inferring an *inductive generalization* which is then used for classification. Harman (1965, 1967) argues that any such inference should always be treated as a special case of *inference to the best explanation* where the relevant sort of explanation appeals to a generalization. But the apparent conflict here appears to be merely terminological.

Transduction differs from the other inductive methods we have been discussing in this way: the classification of new cases is not always based on an inductive generalization from labeled data. So, transduction does not involve *that sort of inductive generalization.* That is because transduction makes use of the information that certain new cases have come up to be assessed.

On the other hand, transduction does involve the implicit acceptance of a non-total generalization G, corresponding to the selected subset of separations in the transformed higher dimensional space. So, transduction does involve inductive generalization, even if not inductive generalization from the labeled data alone.

It is true that, although the data include what new cases have come up, the classifications that transduction gives to these new cases are not treated as data. When additional new cases arise, transduction applied to the old plus the new cases can modify the classifications. It might therefore be said that the principle G derived from accepting the new classifications is hostage to the new cases in a way that inductive generalizations from labeled data are not. But transduction treats the fact that certain new cases have come up as data and new data always have the potentiality to change what rule should be accepted.

In other words, there is a sense in which transduction does not involve inductive generalization, because the relevant generalization is not arrived at from the labeled data alone, and there is a sense in which transduction does involve inductive generalization, because it does arrive at a general rule based on labeled data plus information about what new cases have come up.

What is important and not merely terminological is that, under certain conditions, transduction gives considerably better results in practice than those obtained from methods that use labeled data to infer a rule which is then used to classify new cases (Joachims 1999, Vapnik 2000, Weston et al. 2003, Goutte et al. 2004).

Conclusion

To summarize, we began by sketching certain aspects of statistical learning theory and then commented on philosophical implications for reliability theories of justification, on Popper's appeal to falsifiability, on simplicity as relevant to inductive inference, and on whether direct induction is an instance of inference to the best explanation.

Bibliography

Bishop, M. A., and Trout, J. D., (2005). *Epistemology and the Psychology of Human Judgment.* Oxford: Oxford University Press.

Corfield, D., Schölkopf, B., and Vapnik, V., (2005). "Popper, Falsification and the VC-dimension." Max Planck Institute for Biological Cybernetics Technical Report No. 145.

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

Goldman, A., (1986). *Epistemology and Cognition.* Cambridge, MA: Harvard University Press.

Goldman, A., (2002). *Pathways to Knowledge: Private and Public.* Oxford: Oxford University Press.

Goutte, C., Cancedda, N., Gaussier, E., Dèjean, H. (2004) "Generative vs Discriminative Approaches to Entity Extraction from Label Deficient Data." *JADT 2004, 7es Journ'ees internationales d'Analyse statistique des Donn'ees Textuelles*, Louvain-la-Neuve, Belgium, 10-12 mars.

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

Hastie, T., Tibshirani, R., and Friedman, J., (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Jeffrey, R. (2004). *Subjective Probability (The Real Thing).* Cambridge, England: Cambridge University Press.

Joachims, T. (1999) "Transductive Inference for Text Classification Using Support Vector Machines." In I. Bratko and S. Dzeroski, editors, Proceedings of the 16th International Conference on Machine Learning: 200-9. San Francisco: Morgan Kaufmann.

Kelley, K. T., (1996). *The Logic of Reliable Inquiry.* Oxford: Oxford University Press.

Osherson, D., Stob, M., & Weinstein, S., (1982). *Systems That Learn.* Cambridge: MIT Press.

Popper, K., (1979). *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press.

Popper, K., (2002). *The Logic of Scientific Discovery* (London: Routledge, 2002).

Putnam, H., (1963). "Degree of Confirmation and Inductive Logic." In *The Philosophy of Rudolf*

Carnap, ed. A. Schillp. LaSalle, Indiana: Open Court.

Reichenbach, H., (1949). *The Theory of Probability.* Berkeley: University of California Press.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Schulte, O., (2002). "Formal Learning Theory." *Stanford Encyclopedia of Philosophy* Edward N. Zalta, editor. http://plato.stanford.edu/.

Valiant, L. G. (1984) "A Theory of the Learnable", *Communications of the ACM* 27, pp. 1134-1142.

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 as "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Probability and Its Applications* 16 (1971), pp. 264-280.

Vapnik, V., (1998). *Statistical Learning Theory.* New York: Wiley.

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

Vapnik, V., (2006). *Estimation of Dependencies Based on Empirical Data*, 2nd edition. New York: Springer.

Weston, J., Pèrez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., and Schölkopf, B. (2003) "KDD Cup 2001 Data Analysis: Prediction of Molecular Bioactivity for Drug Design-Binding to Thrombin."