Induction and Simplicity*

Gilbert Harman† Sanjeev R. Kulkarni‡

March 13, 2006

The following is a draft of Chapter Three of *Reliable Reasoning: Induction and Statistical Learning Theory*, to be published by MIT Press. Basic statistical learning theory is concerned with learning from data—either learning how to classify items or learning how to estimate the value of an unknown function. The basic framework assumes that there is a fixed background probability distribution relating observable features of an item to its classification or to the value of the unknown function, where the same distribution determines the probability that a given item will turn up, either as a datum or as a new case to be classified or evaluated. Apart from assuming that the probabilities are independent and identical, no other assumptions are made about the background probability distribution. (Questions about epistemic reliability appear to require some such assumption about background probability.)

In the previous chapter, we introduced a way of thinking about (one kind of) *enumerative induction*, which chooses a hypothesis from a given class C with minimal error on the data. We described a fundamental result (due to Vapnik and Chervonenkis): enumerative induction uniformly converges to the best rule in C if and only if the "VC dimension" of C is finite. (We there explained uniform convergence and VC-dimension.) The present Chapter Three discusses a somewhat different method of inductive inference.

*For the Formal Epistemology Workshop in Berkeley, California, May 27, 2006.

†Department of Philosophy, Princeton University

‡Department of Electrical Engineering, Princeton University

Chapter 3

Induction and "Simplicity"

3.1 Introduction

We are concerned with the reliability of inductive methods. So far we have discussed versions of enumerative induction. In this chapter, we compare enumerative induction with methods that take into account some ordering of hypotheses, perhaps by simplicity. We compare different methods for balancing data-coverage against an ordering of hypotheses in terms of simplicity or some simplicity substitute. Then we consider how these ideas from statistical learning theory might shed light on some philosophical issues. In particular, we distinguish two ways to respond to Goodman's (1965) "new riddle of induction," corresponding to these two kinds of inductive methods. We discuss some of Karl Popper's ideas about scientific method, trying to distinguish what is right and what is wrong about these ideas. Finally we consider how an appeal to simplicity or some similar ordering might provide a principled way to prefer one hypothesis over another skeptical hypothesis that is empirically equivalent with it.

3.2 Empirical Error Minimization

In Chapter 2 we described an important result (Vapnik and Chervonenkis, 1968) about enumerative induction. In statistical learning theory enumerative induction is called "empirical risk minimization" (although it might be more accurate to call it "empirical error minimization," because its only criterion for choosing a rule from C is that the rule should be one of the rules in C with the least empirical error on the data). Vapnik and Chervonenkis show that the method of empirical risk minimization, when used

to select rules of classification, has the following property. If, and only if, the VC dimension of C is finite, then no matter what the background probability distribution, as more and more data are obtained, with probability approaching 1, enumerative induction leads to the acceptance of rules whose expected error approaches the minimum expected error for rules in C.[1]

Moreover, when C has finite VC dimension V we can specify a function, $m(V, \epsilon, \delta)$, which indicates an upper bound to the amount of data needed to guarantee a certain probability $(1 - \delta)$ of endorsing rules with an expected error that approximates that minimum by coming within ϵ of the minimum.

Now, although this is a very nice result, it is also worrisome, because, if C has finite VC dimension, the best rules in C can have an expected error that is much greater than the best possible rule, the Bayes Rule. For example, if C contains only one rule that is always wrong, the best rule in C has an error rate of 1 even if the Bayes rule has an error rate of 0. Even if C contains many rules and has large VC-dimension, the best rule in C may have an error rate close to $\frac{1}{2}$, which is no better than random guessing, even though the Bayes rule might have an error rate close to 0.

Recall our discussion of linear classification rules, which separate YESes and NOs in a D-dimensional feature space with a line, a plane, or a hyperplane. These rules have VC dimension equal to $D + 1$, which is finite as long as the feature space has finite dimension, which it normally does. But linear rules are by themselves quite limited. Recall, for example, that an XOR classification rule cannot be adequately represented by a classification using a linear separation of YESes and NOs. Indeed, the best linear rule for that classification can have a very high expected error.

To be sure, we can use a class of rules C with many more rules, in addition to or instead of linear rules; we can do so as long as the the VC dimension of C is finite. But no matter how high the VC dimension of C, if it is finite there is no guarantee that the expected error of the best rules in C will be close to the expected error of the Bayes Rule.

3.3 Universal Consistency

In order to guarantee that the expected error of the best classification rules in C will be close to the expected error of the best rule of all, the Bayes Rule, it is necessary that C should have infinite VC dimension. But then

[1] Some very mild measurability conditions are required. And, as we mentioned, a similar result holds for enumerative induction used to select rules of function estimation. For the moment, we concentrate on induction to rules of classification.

the nice result about enumerative induction is not forthcoming. We will not be able to specify a function $m(\infty, \delta, \epsilon)$ that would provide an upper bound to the amount of data needed to guarantee a certain probability $(1 - \delta)$ of endorsing rules whose expected error is within ϵ of the minimum expected error for rules in C, which in this case will be the error rate of the Bayes Rule.

On the other hand there are other inductive methods for finding categorization rules that do not have the sort of guarantee of uniform convergence provided by the Vapnik-Chervonenkis result but do have a different desirable property. In particular, it can be shown that certain methods are *universally consistent*. A universally consistent method is one that, for any background probability distribution, with probability approaching 1, as more and more data are obtained, the expected error of rules endorsed by the method approaches in the limit the expected error of the best rule, the Bayes Rule.

Universal consistency does not imply uniform convergence. There may be no bound on the amount of data needed in order to ensure that (with probability approaching 1) the expected error of the rules endorsed by the method will be within ϵ of the expected error of the Bayes Rule. Nevertheless, universal consistency is clearly a desirable characteristic of a method. It does provide a convergence result, because the error rate of the rule endorsed by a universally consistent method converges to the expected error of the Bayes Rule. Although this does not guarantee a rate of convergence, it can be shown that no method provides such a guarantee.

3.3.1 Nearest Neighbor Rules

There is a kind of nearest neighbor rule that is universally consistent, although the simplest such rule is not universally consistent.

Recall that data can be represented as labeled points in a feature space. Suppose that a distance measure is defined on that space. Then the 1-nearest neighbor method says to classify a new item as having the same category as the nearest datum in the feature space. Any set of N data items then serves to specify the corresponding rule of classification (Figure 3.1). As more and more data are obtained, the corresponding rule changes to adapt to the labels on the new items. The 1-nearest neighbor rule is not universally consistent, but it can be shown that in the limit the expected error of the 1-nearest neighbor rule is no more than *twice* the expected error of the Bayes rule, which is quite good if the Bayes rule has a very small error rate.

It is possible to do better by using a variant of the 1-nearest neighbor

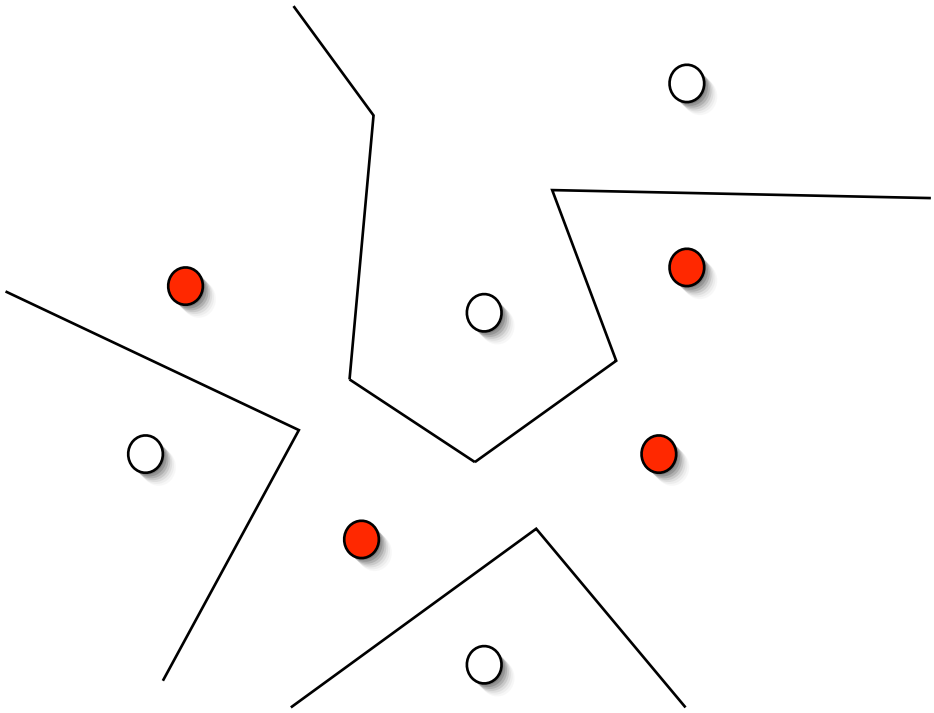

Figure 3.1: Nearest Neighbor Classification

rule. For example, a k-nearest neighbor method says to classify a new item by looking not just at the nearest datum in the feature space but to the k nearest data and assigning to the new item the classification of a majority of those k nearest data. This sometimes (not always) does better than a 1-nearest neighbor rule but is not yet universally consistent.

The key to getting a universally consistent nearest neighbor rule is to let the number of neighbors used grow with N (the amount of data we have) but not too quickly. That is we let k be a function of N, so this is called a k_N-nearest neighbor rule. We let $k_N \to \infty$ so that we use more and more neighbors as the amount of training data increases. But we also make sure that $\frac{k_N}{N} \to 0$, so that asymptotically the number of neighbors we use is a negligible fraction of the total amount of data. This ensures that we use only neighbors that get closer and closer to the point in feature space that we want to categorize. For example, we might let $k_N = \sqrt{N}$ to satisfy both

conditions.

It turns out that with any such k_N (such that $k_N \to \infty$ and $k_N/N \to 0$ are satisfied), in the limit as the amount of training data grows, the performance of the k_N-nearest neighbor rule approaches that of the optimal Bayes decision rule, so this sort of k_N-nearest neighbor rule is universally consistent.

Unfortunately, there will always be probability distributions for which the convergence rate is arbitrarily slow. This is different from enumerative induction using a class of rules C of finite VC dimension, where convergence to the best error rate for classification rules in C is not arbitrarily slow and we can specify a function specifying an upper bound on how much data is needed to achieve a certain convergence, as we have indicated above. On the other hand with enumerative induction the rules in C might not contain the Bayes Rule and might not contain a rule with an error rate that is close to the error rate of the Bayes Rule.

3.4 Structural Risk Minimization

We now want to discuss another kind of universally consistent method for using data to select a rule of classification. This alternative to enumerative induction trades off empirical adequacy with respect to data against another factor, sometimes called "simplicity," although that is not always the best name for the other factor.

One example of this sort of method, "structural risk minimization," (Vapnik and Chervonenkis 1974) is defined in relation to a class of rules that includes an infinite nesting of classes of rules of finite VC dimension. More precisely, $C = C_1 \cup C_2 \cup \cdots \cup C_n \cup \cdots$, where $C_1 \subset C_2 \subset \cdots \subset C_n \subset \cdots$, and where the VC dimension of C_i is strictly less than the VC dimension of C_j when $i < j$. Any class C of this sort has infinite VC dimension.

Structural risk minimization endorses any rule that minimizes some given function of the empirical error of the rule on the data and the VC-dimension of the smallest class containing the rule. It might for example endorse any rule that minimizes the *sum* of these two quantities.

It can be shown that there are many ways to choose these nested classes and the trade-off between fit to data and VC dimension so that structural risk minimization will be universally consistent by endorsing rules that, with probability approaching 1, have expected errors that approach in the limit the expected error of the Bayes Rule.

3.5 Minimum Description Length

Structural risk minimization is one way to balance empirical adequacy with respect to data against some ordering of rules or hypotheses. In that case rules are members of nested classes of finite VC dimension and are ordered by the VC dimension of the smallest class which they belong to.

A different sort of ordering of rules uses the lengths of their shortest representation in some specified system of representation, for example, the shortest computer program of a certain sort that specifies the relevant labeling of points in the feature space (Rissanen 1978, Barron et al. 1998, Chaitin 1974, Akaike 1974, Blum and Blum 1975, Gold 1967, Solomonoff 1964).

The class of rules that are represented in this way can have infinite VC dimension, so enumerative induction with its reliance on empirical risk minimization alone will not be effective. But any such ordering of all representable rules can be used by an inductive method that balances the empirical adequacy of a rule on the data against its place in the ordering. Some methods of this sort will in the limit tend to endorse rules with expected error approaching that of the Bayes Rule.

Notice, by the way, that if rules are ordered by minimum description length, it will not be true for example that all linear rules $y = ax+b$ have the same place in the ordering, because the parameters a and b must be replaced with descriptions of their values and, given a fixed system of representation, different values of the parameters will be represented by longer or shorter representations. For this reason, some linear rules will require considerably longer representations than some quadratic rules, which will by this criterion then be treated as "simpler" than those linear rules.

The kind of ordering involved in structural risk minimization is of a somewhat different sort from any kind of ordering by length of representation. Structural risk minimization identifies rules with mathematical functions and is therefore not limited to considering only rules that are finitely represented in a given system. While the number of linear rules conceived as mathematical functions is uncountably infinite, the number of finitely representable linear rules is only countably infinite.

Even apart from that consideration, the ordering that results from structural risk minimization need not be a well-ordering, because it might not have the property that every rule in the ordering has at most only finitely many rules ordered before it. In a typical application of structural risk minimization infinitely many linear rules are ordered before any nondegenerate quadratic rule. But an ordering of rules by description length can be con-

verted into a well-ordering of rules (by ordering "alphabetically" all rules whose shortest representations have the same length).

3.6 Simplicity

If the ordering against which empirical fit is balanced is supposed to be an ordering in terms of simplicity, one might object that this wrongly assumes that the world is simple. But to use simplicity in this way in inductive reasoning is not to assume the world is simple. What is at issue is comparative simplicity. Induction favors a simpler hypothesis over a less simple hypothesis that fits the data equally well. Given enough data, that preference can lead to the acceptance of very unsimple hypotheses.

3.7 Function Estimation and Curve Fitting

We have discussed these two sorts of induction as aimed at coming up with rules of classification. Similar results apply to function estimation or curve fitting. Here we review our earlier discussion of "function estimation" and note how structural risk minimization applies.

In function estimation, the task is to estimate the value of a function given the values of each of D "arguments". The function in question may or may not depend on all of the arguments and may depend on other arguments as well. We assume that there is a background probability distribution that specifies the probability relationship between values of the "arguments" and possible observed values of the function. We represent each of the D "arguments" of the function as features in a D dimensional feature space, where the value of the function for specified arguments on a particular occasion is the correct labeling of the relevant point in feature space on that occasion. (We have to mention particular occasions, because the function may depend on more arguments than we have specified.) A possible rule for estimating the function can be represented as a curve in a $D + 1$ space.

We mentioned a very simple example where $D = 1$ and we are trying to estimate an unknown function f that we take to depend on a single argument x. The task is to use data to estimate the function, where the possibly noisy data provide values of the function on certain occasions for certain values of the argument. We have already discussed how any estimate of the function has an expected error determined by the background probability distribution.

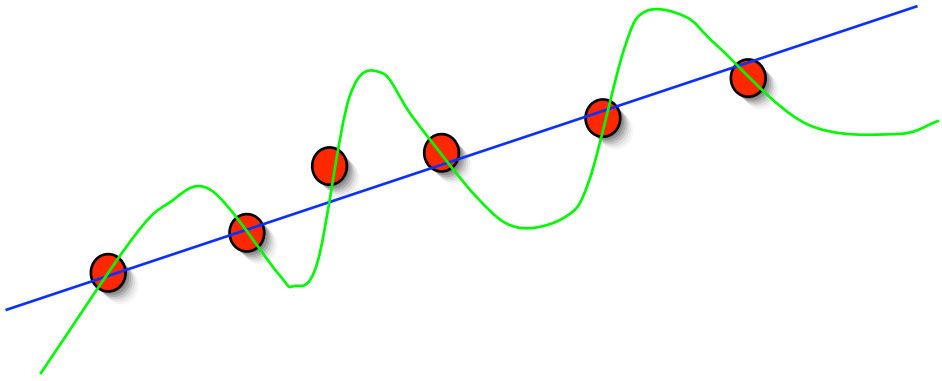

Figure 3.2: Curve Fitting

Each datum can be represented as a point in the plane, where the x coordinate represents the value of the argument and the y coordinate represents the value of the function the datum provides for that value of the argument. The task is to estimate the function for other points by fitting a curve to the data.

Obviously, infinitely many curves go through all the data (Figure 3.2). So there are at least two possible strategies. We can limit the curves to a certain set C, such as the set of straight lines and choose that curve in C with the least error on the data. Or we can allow many more curves in C and use something like structural risk minimization to select a curve, trying to minimize some function of the empirical error on the data and the complexity of the curve.

We might measure complexity by the VC dimension of the class C, thinking of these curves as the border between YES, too high, and NO, too low.

One might use simple enumerative induction to fit a curve to data points, for example, a linear equation. Or one might balance empirical fit to data against something else, as in structural risk minimization.

3.8 Goodman's New Riddle

The distinction between empirical risk minimization and structural risk minimization sheds light on certain philosophical issues. For one thing, it sheds light on different ways some philosophers have reacted to Nelson Goodman's

"new riddle of induction" (Goodman, 1967).

As formulated, Goodman's "New Riddle" doesn't fit into the standard statistical learning theory paradigm. But there is a reformulation of it that does fit.

We might formulate the original version as follows. The problem is to predict whether a given item is green or not, given when it is first observed. In other words, there is a single feature, representing time of first observation and the feature space is therefore one-dimensional. The data consist in labeled points in this one-dimensional feature space, where each label is either "green" or "not-green". We want to use the data to select a function that assigns labels to all points in the feature space. Our goal is to minimize expected error in our predictions about cases as they arise.

This version of the problem does not fit the basic statistical learning theory paradigm in which data are assumed to arise from the same probability distribution as new cases to be predicted. In this first version of Goodman's problem, the relevant feature, time of first observation, is not randomly distributed because there is no chance that the data will assign labels to items first examined later than the current time.

But we can easily modify the problem by taking the relevant feature to be some property of items that we can assume to have the same random distribution in the data and in new cases, for example, the weight or *mass* of the item. Then the data consist in certain pairings of values of measured mass and labels, "green" and "not green". Again we want to use the data to select a function that assigns labels to all possible values for mass, where our goal is to minimize expected error in our predictions about cases as they arise.

Suppose that we want to use enumerative induction with no limit on the hypotheses in C. Of course, if all the data points are labeled "green" and none are labeled "not green", it seems we would want to adopt the hypothesis that all points are to be labeled "green", because that hypothesis has no error on the data. This would lead us to predict that the next item, no matter what its mass, will be correctly labeled "green". However, to adapt Goodman's point in his original formulation of the "riddle," there are other hypotheses that will correctly fit the data that will give different predictions about new items. For example, there will always be a possible hypothesis that says assigns the label "green" to all the actual data points and "not green" to all other points. So, the rule of enumerative induction does not give useful advice about cases whose values of the relevant feature differ from any data points.

From this, Goodman concludes that we cannot allow enumerative induc-

tion to treat equally all possible hypotheses. In our terms, there must be limits on C. Furthermore, Goodman assumes that there is a unique class of hypotheses C, consisting in those hypotheses that are "confirmed" by their instances. The "new riddle of induction" is then the problem of characterizing the relevant class of hypotheses, C, the confirmable or law-like hypotheses. Goodman attempts to advance a solution to this problem (a) by characterizing a class of "projectible" predicates in terms of the extent to which these predicates have been used to make successful predictions in the past and (b) by giving principles that explain the confirmability of a hypothesis in terms of the projectibility of the predicates in which it is expressed.

Goodman argues that projectible predicates cannot be identified with those predicates for which we have a single word, like "green" as opposed to "green if mass of 15, 24, 33, ... and not green otherwise," because we could use a single word "grue" for the latter predicate. He argues that projectible predicates cannot be identified with directly observational predicates, like "green", because we can envision a machine that can directly observe whether something is "grue". Goodman himself suggests that the projectible predicates can be characterized in terms of the extent to which these predicates have been used to make successful predictions in the past

Statistical learning theory takes a very different approach. It does not attempt to solve this "new riddle of induction". It does not attempt to distinguish those predicates that are really projectible from those that are not and it does not attempt to distinguish those hypotheses that are really confirmable from their instances from those that are not.

Of course, statistical learning theory does accept the moral that induction requires inductive bias among hypotheses. But it does not attempt to specify a unique class C of confirmable hypotheses. In the case of enumerative induction, statistical learning theory says only that the set C of hypotheses to be considered must have finite VC-dimension. In the case of structural risk minimization, statistical learning theory requires a certain structure on the set of hypotheses being considered. Statistical learning theory does not attempt to specify which particular hypotheses are to be included in the set C, nor where particular hypotheses appear in the structures needed for structural risk minimization.

Goodman's riddle has received extensive discussion by philosophers (some collected in Stalker 1994, and Elgin 1997). While many authors suppose that the solution to the new riddle of induction requires specifying some relevant class of projectible hypotheses, others have argued instead that what is needed is an account of "degrees of projectibility," where for example in-

tuitively simpler hypotheses count as more projectible than intuitively more complex hypotheses.

One observation about these two interpretations of the riddle is that the first, with its emphasis on restricting induction to a special class of projectible hypotheses, involves identifying induction with enumerative induction, conceived as empirical risk minimization, with the advantages and disadvantages of considering only rules from a class of rules with finite VC dimension. The second interpretation, with its emphasis on degrees of projectibility, can allow consideration of rules from a class of rules with infinite VC dimension. It can do this by abandoning simple enumerative induction in favor of structural risk minimization or some other way of balancing data-coverage against simplicity or projectibility.

Philosophers discussing Goodman's new riddle have not fully appreciated that these two ways of approaching the new riddle of induction involve different kinds of inductive methods, empirical risk minimization on the one hand and methods that balance fit to data against something else on the other hand.

One philosophically useful thing about the analysis of inductive reasoning in statistical learning theory is the way it sheds light on the difference between these two interpretations of Goodman's new riddle.

3.9 Popper on Simplicity

We now want to say something more about Popper's (1934, 1979) discussion of scientific method. We noted earlier that 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 function estimation with one argument, 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.

As explained in Chapter 2, 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 Chapter 2 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 $f_2(c) = b$.[2] Given such a coding, we can

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

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.

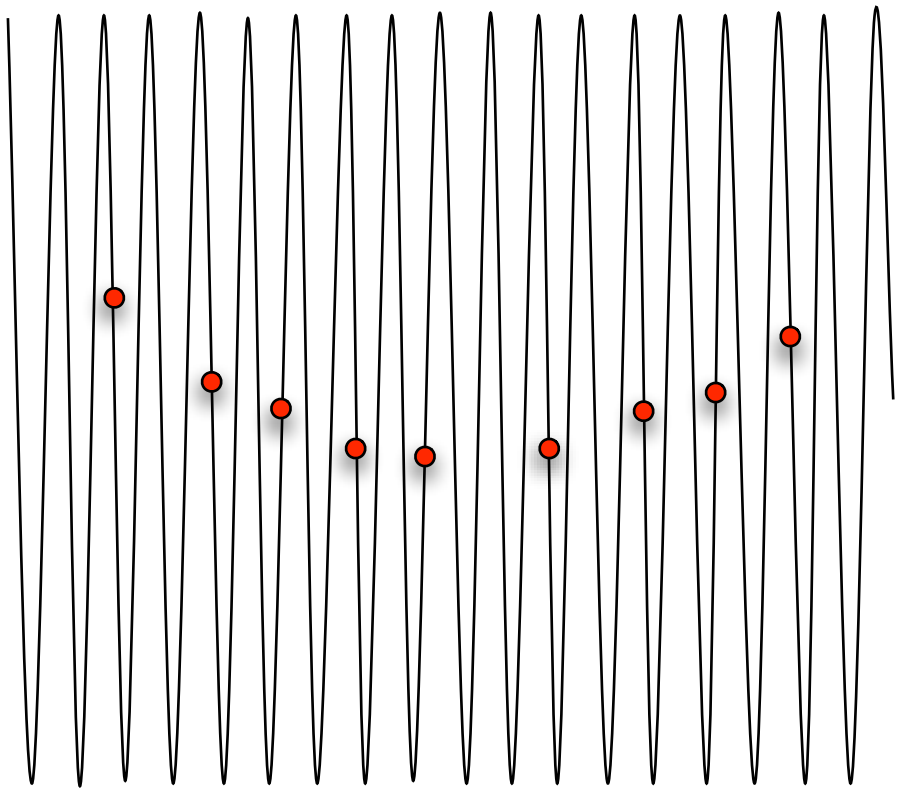

Figure 3.3: Function Estimation using Sine Curves

To see this, consider the class of sine curves $y = a\sin(bx)$. For any set of n consistent data points (which do not assign different y values to the same x value) there will be sine curves coming arbitrarily close to those

points(Figure 3.3). In that sense, the class of sine curves has infinite "falsifiability" in Popper's sense 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 class of sine curves can also be seen to have infinite VC dimension if we think of the curves as rules for classifying points as "too high" or "not too high," because for any n there will be a set of n points that is shattered by the class of sine curves. That is, members of that class can provide the 2^n possible classifications of the n points.

The fact that the 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".

3.10 Empirically Equivalent Rules

Finally, we consider whether empirically equivalent hypotheses must always be treated in the same way in statistical learning theory. In particular, what about scientific hypotheses in comparison with empirically equivalent skeptical hypotheses?

Suppose two hypotheses, H and D, are empirically equivalent. For example, where H is some highly regarded scientific hypothesis, let D be the corresponding demonic hypothesis that a powerful god-like demon has arranged that the data you get will be exactly as expected if H were true. Could simplicity as analyzed in statistical learning theory provide a reason to accept H rather than D?

One might suppose that the answer is "no", because the kinds of analyses provided by statistical learning theory concern how to minimize expected errors and these hypotheses make exactly the same predictions. Indeed, if we identify the hypotheses with their predictions, they are the same hypothesis.

But it isn't obvious that hypotheses that make the same predictions should be identified. The way a hypothesis is represented suggests what class of hypotheses it belongs to for purposes of assessing simplicity. Different representations suggest different classes. Even mathematically equivalent hypotheses might be treated differently within statistical learning theory. The class of linear hypotheses, $f(x) = ax + b$, is simpler than the class of quadratic hypotheses, $f(x) = ax^2 + bx + c$, on various measures—number of

parameters, VC-dimension, etc. If the first parameter of a quadratic hypothesis is 0, the hypothesis is mathematically equivalent to a linear hypothesis. But its linear representation belongs to a simpler class than the quadratic representation. So for purposes of choice of rule, there is reason to count the linear representation as simpler than the quadratic representation.

Similarly, although H and D yield the same predictions, there is a sense in which they are not contained in the same hypothesis classes. We might say that H falls into a class of hypotheses with a better simplicity ranking than D, perhaps because the class containing H has a lower VC-dimension than the class containing D. The relevant class containing D might contain any hypothesis of the form, "The data will be exactly as expected as if ϕ were true," where ϕ ranges over all possible scientific hypothesis. Since ϕ has infinite VC-dimension, so does this class containing D. From this perspective, there is reason to prefer H over D even though they are empirically equivalent.

So, we may have reason to think that we are not just living in the Matrix (Wachowski and Wachowski 1999)!

3.11 Important Ideas from Statistical Learning Theory

Here are some of the ideas from statistical learning theory that we have discussed so far that we believe are philosophically and methodologically important.

Statistical learning theory provides a way of thinking about the reliability of a rule of classification in terms of expected cost or expected error, where that presupposes a background statistical probability distribution.

With respect to rules of classification, there is the notion of the Bayes Rule, the most reliable rule, the rule with the least expected error or expected cost.

There is the idea that the goodness of an inductive method is to be measured in terms of the reliability of the classification rules the method comes up with.

There is the point that useful inductive methods require some inductive bias, either as reflected in a restriction in the rules in C or as a preference for some rules in C over others.

There is the idea of shattering, as capturing a kind of notion of falsifiability, and the corresponding notion of VC dimension.

There is the contrast between uniform convergence of error rates and universal consistency.

In the next chapter we will discuss some additional ideas from statistical learning theory and will consider their significance for psychology and cognitive science as well as for philosophy.

3.12 Summary

In this chapter, we compared enumerative induction with methods that also take into account some ordering of hypotheses. We discussed how these methods apply to classification and function-estimation or curve fitting. We compared two different methods for balancing data-coverage against an ordering of hypotheses in terms of simplicity or some simplicity substitute. We noted that there are two ways to respond to Goodman's (1965) new riddle of induction, corresponding to these two kinds of inductive method. We also discussed some of Karl Popper's ideas about scientific method, trying to distinguish what is right and what is wrong about these ideas. Finally, we considered how appeal to simplicity or some similar ordering might provide a principled way to prefer one hypothesis over another skeptical hypothesis that is empirically equivalent with it.

Bibliography

Akaike, H., (1974). "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.* AC-19: 716-723.

Barron, A., Rissanen, J., and Yu, B., (1998). "The Minimum Description Length Principle in Coding and Modeling." *IEEE Trans. Information Theory* 44: 2743-2760.

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

Blum, L., and Blum, M., (1975). "Toward a Mathematical Theory of Inductive Inference," *Information and Control* 28: 125-55.

Bongard, M., (1970) *Pattern Recognition* (Spartan Books, Washington, D.C.

Burge, T., (1993). "Content Preservation," *Philosophical Review*.

Chaitin, G. J., (1974). "Information-Theoretic Computational Complexity," *IEEE Transactions on Information Theory* IT-20: 10-15.

Chomsky, N., (1968). *Language and Mind*. New York: Harcourt, Brace & World.

Chomsky, N., (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Cullicover, P. W., (1997). *Principles and Parameters: An Introduction to Syntatic Theory*. Oxford: Oxford University Press.

Dancy, J., (1993). *Moral Reasons*. Oxford, Blackwell.

Daniels, N., (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics." *Journal of Philosophy* 76: 256-82.

Descartes, R., (1641). *Meditationes de Prima Philosophia*. Paris.

Duda, R. O., Hart, P. E., and Stork, D. G (2001) *Pattern Classification*, Second Edition (Wiley, New York), Chapters 1, 2.

Elgin, C., (1997). *Nelson Goodman's New Riddle of Induction. The Philosophy of Nelson Goodman, Volume 2*. New York: Garland.

Feldman, J. A., (1981). "A Connectionist Model of Visual Memory." In G. E. Hinton and J. A. Anderson (Eds.). *Parallel Models of Associative Memory*, (Hillsdale, NJ.: Erlbaum), 49-81.

Foley, R., (1994). "Egoism in Epistemology." In F. Schmitt, ed., *Socializing Epistemology*. Lanham: Rowman and Littlefield.

Gladwell, M., (2005). *Blink: The Power of Thinking Without Thinking*, New York: Little Brown.

Gold, E. M., (1967). "Language Identification in the Limit," *Information and Control* 10: 447-74.

Goodman, N., (1953). *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard 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.

Graff, D., (2000). "Shifting Sands: An Interest-Relative Theory of Vagueness," *Philosophical Topics* 28: 45-81.

Hacking, I., (1965). *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Harman, G., (1965). "The inference to the best explanation," *Philosophical Review* 74: 88-95.

Harman, G., (1967). "Enumerative induction as inference to the best explanation," *Journal of Philosophy* 64: 529-33.

Harman, G., (2005). "Moral Particularism and Transduction." *Philosophical Issues* 15.

Harnad, S., (1987). "Psychophysical and cognitive aspects of categorical perception: A critical overview." Harnad, S., ed., (1987) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

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

Holyoak, K. J. and Simon, D., (1999). "Bidirectional Reasoning in Decision Making by Constraint Satisfaction," *Journal of Experimental Psychology: General*, 128: 3-31.

Hooker, B. and Little, M., (2000). *Moral Particularism*. New York, Oxford University Press.

Iyengar, S. S. and Lepper, M. R., (2000). "When Choice Is Demotivating: Can One Desire Too Much of a Good Thing?" *Journal of Personality and Social Psychology* 79: 995-1006.

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.

Kihlbom, U., (2002). *Ethical Particularism*. Stockhold Studies in Philosophy 23. Stockholm: Almqvist and Wiksell.

Kulkarni, S. R., Lugosi, G., and Venkatesh, L. S., (1998). "Learning Pattern Classification: A Survey," *IEEE Transactions On Information Theory* 44: 2178-2206.

McDowell, J., (1998). *Mind, Value, and Reality*. Cambridge, MA: Harvard University Press.

Popper, K., (1934). *Logik der Forschung*. Vienna: Springer. Specific page and section references are to the English translation, *The Logic of Scientific Discovery* (London: Routledge, 2002).

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

Rawls, J., (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Read, S. J., Snow, C. J., and Simon, D., (2003). "Constraint Satisfaction Processes in Social Reasoning." *Proceedings of the 25th Annual Conference of the Cognitive Science Society*: 964-969.

Redelmeier, D. A. and Shafir, E., (1995). "Medical Decision Making in Situations that Offer Multiple alternatives," *Journal of the American Medical Association* 273: 302-5.

Rissanen, J., (1978). Modeling by shortest data description. *Automatica* 14, 465-471.

Schwartz, B., (2004). *The Paradox of Choice: Why More Is Less*. New York: HarperCollins.

Simon, D., (2004). "A Third View of the Black Box," *University of Chicago Law Review*, 71, 511-586.

Simon, D. and Holyoak, K. J., (2002). "Structural Dynamics of Cognition: From Consistency Theories to Constraint Satisfaction," *Personality and Social Psychology Review*, 6: 283-294.

Simon, D., Pham, L. B., Le, Q. A., and Holyoak, K. J., (2001). "The Emergence of Coherence over the Course of Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27: 1250-1260.

Sinnott-Armstrong, W., (1999). "Varieties of Particularism," *Metaphilosophy* 30: 1-12.

Solomonoff, R. J., (1964). "A Formal Theory of Inductive Inference," *Information and Control* 7: 1-22, 224-54.

Stalker, D., editor, (1994). *Grue!: The New Riddle of Induction* Peru, Illinois: Open Court.

Stich, S. and Nisbett, R., (1980). "Justification and the Psychology of Human Reasoning," *Philosophy of Science* 47: 188-202.

Thagard, P., (1988). *Computational Philosophy of Science*. Cambridge, MA: MIT Press.

Thagard, P., (1989). "Explanatory Coherence." *Brain and Behavioral Sciences*, 12: 435-467.

Thagard, P., (2000). *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Tversky, A. and Kahneman, D., (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 185: 1124-1131.

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

Vapnik, V., (1979). *Estimation of Dependencies Based on Empirical Data* (in Russian), Moskow: Nauka. English translation (1982) New York: Springer.

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., 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., and Chervonenkis, A. Ja., (1947). *Theory of Pattern Recognition* (in Russian), Nauka: Moscow.

Väyrynen, P., (2004). "Particularism and Default Reasons," *Ethical Theory and Moral Practice* 7: 53-79.

Wachowski, A., and Wachowski, L., (1999). *The Matrix*. Warner Brothers.

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." *Bioinformatics*.

Wiggins, D., (1998). *Needs, Values, and Truth*, 3rd edition. Oxford, Oxford University Press.