The Jean Nicod Lectures
François Recanati, editor

The Elm and the Expert: Mentalese and Its Semantics, Jerry A. Fodor (1994)

Naturalizing the Mind, Fred Dretske (1995)

Strong Feelings: Emotion, Addiction, and Human Behavior, Jon Elster (1999)

Knowledge, Possibility, and Consciousness, John Perry (2001)

Rationality in Action, John R. Searle (2001)

Varieties of Meaning, Ruth Garrett Millikan (2004)

Sweet Dreams: Philosophical Obstacles to a Science of Consciousness, Daniel C. Dennett (2005)

Things and Places: How the Mind Connects with the World, Zenon W. Pylyshyn (2007)

Reliable Reasoning: Induction and Statistical Learning Theory, Gilbert Harman and Sanjeev Kulkarni (2007)

Reliable Reasoning

Induction and Statistical Learning Theory

Gilbert Harman and Sanjeev Kulkarni

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

2 Induction and VC Dimension

2.1 Pattern Recognition

The problem of inductive reliability can be seen as a problem in learning theory. It is the problem of finding reliable ways to learn from data. For example, how can one find and assess inductive *methods* for using data to arrive at reliable *rules* for classifying new cases or estimating the value of a real variable?

In thinking about this problem, two kinds of methods or rules must be carefully distinguished. Rules of classification or estimation must be carefully distinguished from inductive methods for finding such rules. Rules of classification or estimation are rules for using observed *features* of items to classify them or to estimate the values of a real variable. Inductive methods for finding such rules are methods for using *data* to select such rules of classification or estimation.

In the previous chapter we discussed a particular method, *enumerative induction*. In this chapter, we will say more about using enumerative induction to learn rules of classification and to estimate the values of real variables. In our next chapter we discuss some other methods for using data to arrive at rules of classification or estimation.

In our fourth and final chapter we will go beyond these sorts of *inductive* methods to discuss methods of *transduction* that do not (in a certain sense) first use data to arrive at rules of classification or estimation that are then used to classify new cases or estimate values of a real variable for new cases as they arise. These methods use information about what new cases have actually come up in deciding what to say about the new cases. But in this second chapter and the next third chapter we will be concerned only with inductive methods for coming up with rules of classification or estimation.

An inductive method is a principle for finding a *pattern* in the data that can then be used to classify new cases or to estimate the values of a real variable. So, the problem of finding a good inductive method is sometimes called a *pattern recognition* problem (Bongard 1970; Duda, Hart, and Stork 2001).

2.1.1 Pattern Classification

In a pattern classification problem, we seek to come up with a rule for using observable features of objects in order to classify them into one of a finite number of categories, where each feature can take several possible values, which can be represented by real numbers. In the most common case there are just two categories, so this is the case we consider here. For purposes of medical diagnosis, values of the features could represent the results of certain medical tests. For recognition of written addresses on envelopes, the relevant area of an envelope could be represented by a grid of $W \times H$ pixels, with a feature value for each pixel representing the intensity of light at the pixel, so there would be $W \times H$ different features. For face recognition from color photographs using a grid of $W \times H$ pixels,

feature values could include representations of each of the RGB values of each pixel (the intensities of red, green, and blue components of the color of the pixel), so there would be $3 \times W \times H$ features.

Each observable feature can be treated as a dimension in a D-dimensional *feature space*. If there is a single feature, F, the feature space is one-dimensional, a line. A point in the feature space has a single F coordinate representing the value of that feature. If there are two features, F_1 and F_2, the feature space is the two-dimensional plane and each point has two coordinates, an F_1 coordinate and a F_2 coordinate, indicating the values of those two features. If there are three features, F_1, F_2, and F_3, a point in the three-dimensional feature space has an F_1 coordinate, representing the value of feature F_1, an F_2 coordinate, representing the value of feature F_2, and an F_3 coordinate, representing the value of feature F_3.

In the case of the $H \times W$ color pixels, there are $3 \times H \times W$ dimensions to this space. Each point in this large feature space has $3 \times H \times W$ coordinates. Each such point represents a particular possible color picture, a particular way of assigning feature values to the color pixels.

Data for learning can then be represented by labeled points in the feature space. The coordinates of each such point represent an object with the corresponding feature values. The label indicates a classification of that object, perhaps as provided by an "expert."

A possible new case to be categorized is then represented by an unlabeled point, the inductive task being to interpolate or extrapolate labelings from already labeled data points to the unlabeled point (figure 2.1).

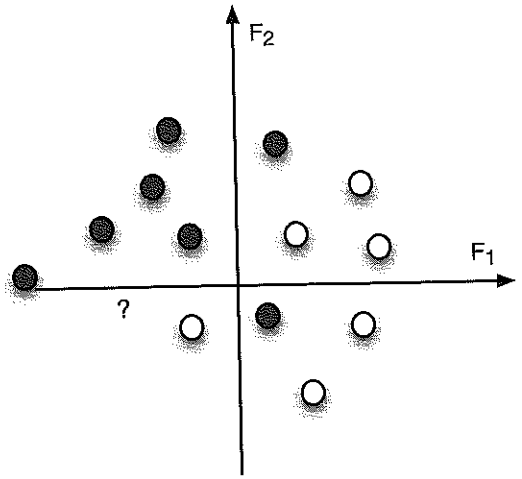

Figure 2.1
Feature space: Gray dots label points that are categorized as YESes; white dots label points that are categorized as NOs. The point at the question mark is unlabeled.

2.1.2 Estimating the Value of a Real Variable

A related problem is the problem of using data in order to estimate the value of a real variable. This problem is like a categorization problem in which the value of the real variable is the correct labeling of a point in feature space. However, there are two important differences between categorization and real variable estimation. One difference is that categorization involves applying one of a small finite number of possible categories (for example, two—YES and NO), while the possible values of a real-valued variable can be nondenumerably infinite. This gives rise to the second difference, which is that in esimation it is not useful to consider the probability of an incorrect estimate. Instead,

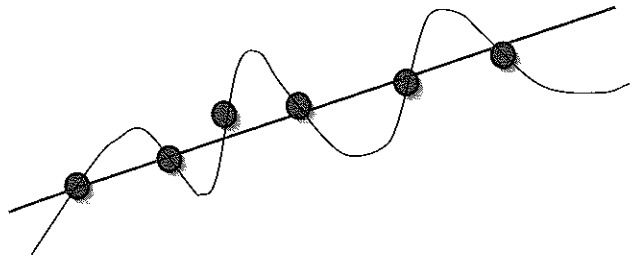

Figure 2.2
Curve fitting.

it is more appropriate to consider how close one real value is to another (rather than whether they are exactly the same).

A variable estimation problem can be considered a "curve fitting problem" if the estimated values of the variable are represented by a "curve" (or hypersurface) in $D+1$ dimensional space. To take a very simple example (figure 2.2), assume that our estimate of the real variable y will be a function $f(x)$ of one argument x and our task is to use data to find a function that provides the best estimate of y. 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 estimation function according to that particular datum. The task is to estimate y by fitting a curve to the data.

2.2 Background Probability Distribution

In general, in classification problems there will not be a perfect correlation between observed features and the best classifications of objects with those features. For one thing, there may be *noise* or errors in measurement in the observed features.

Furthermore, the relation between features and classification may be at best merely probabilistic even apart from issues of noise. For example, suppose the task is to recognize whether a person is currently happy, given only a picture of the expression on his or her face. It may very well be true that a person with a certain visual expression is sometimes happy and sometimes sad, so that the relation between the features revealed in that picture and the correct classification of the person as happy or sad is only probabilistic.

Similarly, estimation of a real valued variable must allow for noise in the data, as well as the possibility that the variable depends on other factors than those we use to make our estimate.

We have already suggested that questions about the reliability of inductive conclusions presuppose that there is a possibly unknown background statistical probability distribution. Discussions of the reliability of a rule of classification presuppose that there is a statistical probabilistic connection between observable features and correct classification. And discussions of the reliability of a rule of estimation of a real valued variable presuppose that there is a statistical probabilistic connection between observable features and the value of the variable given those features.

So, we assume that there is a background probability distribution P which (among other things) defines the conditional probabilities that an item is correctly classified as an A given that it has certain observed features, $P(A|F_1\&F_2\&F_3\&\ldots)$. Or we assume that the background probability P defines in this way the conditional probabilities that the value of a given variable is A given the observation of features whose values are $F_1\&F_2\&F_3\&\ldots$. (In many contexts, conditional probability densities are required, rather than simple conditional probabilities. See, for example, Duda, Hart, and Stork 2001.)

In other words, we assume that the data represent a random sample arising from the background probability distribution, and we assume that new cases that are encountered are also randomly produced by that distribution. We do not assume that we know what that distribution is. We do not assume it is a normal distribution or that its mean, standard deviation, and so on are known. This is a problem in "nonparametric statistics," because nothing is assumed about the parameters of the background probability distribution.

The only assumptions made about the background probability distribution are that (1) the probability of the occurrence of an item with certain features and classification is independent of the occurrence of other items, and (2) the same distribution governs the occurrence of each item. One familiar example of an assumption of probabilistic independence and identical distribution is the assumption that the probability that a tossed coin will come up heads is independent of the results of other tosses and that the probability of heads for each toss is the same. (Given a theory based on an assumption of such probabilistic independence and identical distribution, it may be possible to extend the theory by relaxing the assumptions of independence and identical distribution, but we will not consider such extensions in this book.)

The gambler's fallacy, mentioned briefly in the previous chapter, rests on a confusion about probabilistically independent events. After a tossed coin has come up heads four times in a row, the gambler's fallacy leads to the thought that the probability of heads on the next toss is considerably greater than one half "because heads is due."

This thought may rest on the following reasoning:

The coin is fair, so it should come up heads about half the time in a long enough string of tosses. In particular, it is quite probable that heads will come up between four and six times in ten tosses. Since heads has not come up in the first four tosses, it needs to come up at least four times in the next six. So the probability of getting heads on the next toss is at least 4/6.

This reasoning is mistaken. Given that the results of tosses of the coin are probabilistically independent and that the coin is fair, the probability of heads on the next toss is still 1/2. It remains true that in the long run, the frequency of heads will approach 1/2, despite the initial run of four tails. The impact of any finite number of initial results will be dwarfed by the impact of the rest of the idealized long run. The "long run" is infinitely long and thus much longer than any merely finite beginning. Any infinite series in which the frequencey of heads approaches 1/2 will continue to do so with any large finite number of tails added to its beginning.

2.3 Reliability of Rules of Classification and Estimation

2.3.1 Reliability of a Classification Rule

We have discussed the distinction between rules of classification and a method for finding those rules. We have discussed how items to be classified might be represented as points in a feature space and how data might be represented as labeled points in a feature space. We have noted that the reliability of a rule of classification depends on a possibly unknown background statistical probability distribution. And we have noted that we might be able to make only minimal assumptions about that background

probability distribution, namely, the assumption of probabilistic independence and identical distribution (although as we have mentioned, this assumption can be relaxed in various ways).

We can now distinguish two questions.

1. With respect to the (unknown) background probability distribution what is a best rule of classification?

2. If the background probability distribution is unknown, under what conditions can data be used to find a best (or good enough) rule of classification?

One possible answer to the first question is that the best rule is the one that minimizes the expected frequency of error, where the expected frequency of error is determined by the probability (according to the unknown background probability distribution) that a use of the rule will lead to an error.

That answer assumes all errors are equally bad. If certain sorts of errors are worse than others, that can be taken that into account. It could happen, for example, in medical testing, where false positives might be less serious than false negatives. We might then assign different weights or costs to different sorts of errors and then treat the best rule as the one that minimizes expected cost.

The best rule is standardly called the "Bayes Rule" (see, e.g., Hastie et al. 2001, p. 21). Given the (unknown) background probability distribution, the Bayes Rule is the rule that for each set of features chooses the classification with the smallest expected cost, given that set of features. In the special case in which all errors are equally bad, the Bayes Rule is the rule that chooses, for each set of features, the classification with greatest conditional probability given that set of features, which results in the smallest probability of error. (For simplicity in what

follows we will treat all errors as equally bad and take the best rule to be the rule that simply minimizes expected error.)

2.3.2 Reliability of a Rule of Real Variable Estimation

Recall that, in addition to having to allow for noise in the data, estimation of a real valued variable must also allow for the possibility that the variable in question is only probabilistically related to observable features of the data. So, given values of those observable features, there will be various possible values of the real variable, values whose probabilities (or probability densities) are determined by the unknown background probability distribution. On a particular occasion when those are the values of the observable features, the real variable will have a particular value. The amount of error on that occasion of a particular estimate of the value of the function for those values of the arguments might be measured by the absolute value of the difference between the estimate and the value of the variable on that occasion, or by the square of that difference. More generally, the expected error of an estimate with respect to given observable features is the sum of the possible amounts of error of the estimate for those arguments weighted by the probability of those errors (or an integral using probability densities rather than probabilities—we omit details). A rule of estimation of the value of the variable, given all possible observable features, has an *expected error* equal to the sum of its expected errors for various values of observable features weighted by the probability of observing those values of the features. (Again, normally this would be an integral rather than a sum.) In this way, any rule of real variable estimation has an expected error determined by the background probability function. The Bayes Rule for estimating

a variable is then the best rule, that is, the rule for estimating that variable with the lowest expected error in general.

2.4 Inductive Learning

Is there an inductive method that will lead to the selection of the Bayes Rule, given enough data?

One way to proceed would be to try to use data first to discover or at least approximate the background probability distribution and then use that probability distribution to determine the Bayes rule. But as we shall see that turns out to be impractical. Indeed, there is no practical way of exactly finding the Bayes Rule that will work no matter what the background probability distribution given enough data.

Setting our sights somewhat lower, we can consider the following inductive learning question: To what extent can we use data to find a rule of classification or real variable estimation with performance that is as good as (or comparable to) the performance of the Bayes Rule?

The third chapter describes a positive answer to this last question. There is a sense in which we can use data to find a rule with performance that approaches the performance of the Bayes Rule as we get increasing amounts of data.

But in order to explain that answer, it will be useful to spend the rest of this chapter considering the performance of the method of enumerative induction that we began to discuss in chapter 1. There we gave an example of using enumerative induction to find a rule of categorization. Enumerative induction might also be used to find a rule of real variable estimation. Recall that enumerative induction is a method for using data to

choose a rule from a restricted set of rules C: choose a rule from C with minimum error on the data.

The idea behind enumerative induction is, first, to use a rule's "empirical risk," its rate of error on the data as an estimate of its expected error on new cases and then, second, to choose a rule from C whose empirical error on the data is least.

It is possible that several rules from C are tied for having the same minimal error on the data. In that case, we will say that enumerative induction *endorses* all of the tied rules.

As we mentioned in the first chapter, this method is useful only if there are significant limits on the rules included in C. If all possible rules are included, then the rules that minimize error on the data will endorse all possible judgments for items with features that do not show up in the data—all possible interpolations and extrapolations to other cases.

On the other hand, as we also mentioned in chapter 1, if there are significant limits on the rules in C, then C might not contain the Bayes Rule, the rule with the least expected error. In fact, C might not contain any rule with expected error comparable to the minimal expected error of the Bayes Rule. The best rules in C might well have significantly greater expected error than the Bayes Rule.

Still, there will be a certain minimum expected error for rules in C. Then the goal of enumerative induction will be to find a rule with expected error that is near that minimum value. Or, since no method can be expected to find such a rule without a sufficient amount of data, the goal will be to find such a rule given a sufficient amount of data. Actually, even that goal is too ambitious in comparison with the goal of probably finding such a rule. That is to say, a realistic goal is that, with probability approaching 1, given more and more data, the expected error of

a rule endorsed by enumerative induction at each stage will approach the minimum value of expected error for rules in C.

2.4.1 Linear Classification and Estimation Rules

Let us consider an example of enumerative induction to a classification rule. Recall that we are thinking of the observable features of objects as represented in a feature space. Let us suppose that we are interested in a very simple YES/NO classification of some sort. The features might be the results of D different medical tests. The classification of the person with those results might be either YES, has "metrocis" (an imaginary illness) or NO, does not have metrocis. The feature space has D dimensions, one for the result of each test. In this case any classification rule determines a set of points for which the classification is YES according to that rule. The remaining points are classified NO by the rule. So, instead of thinking of rules as linguistic or symbolic expressions, we can think about the corresponding sets of points in feature space (figure 2.3), perhaps certain scattered areas or volumes or hypervolumes of the space—"hypervolumes," because the dimensions of the feature space will typically be greater than three.

Linear classification rules are a very simple case which divide the feature space into two parts separated by a line or hyperplane, with YES on one side and NO on the other. If there are two medical tests with results F_1 and F_2, then one possible classification rule would classify the patient as having metrocis if $F_1 + 2F_2 > 6$ and otherwise classify the patient as not having metrocis. This is a linear classification rule in the sense that the rule distinguishes the YESes from the NOs by the straight line intersecting the F_2 axis at $(0,3)$ and the F_1 axis at $(6,0)$ (figure 2.4).

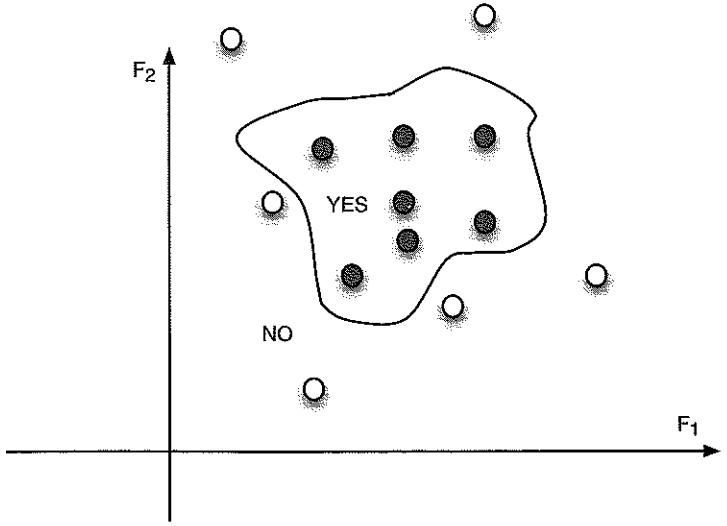

Figure 2.3
Rules as sets of points in feature space.

For any given data, it is easy to find a linear classification rule with minimum error on the data. But of course such rules are limited in what they can represent. They cannot, for example, represent an XOR rule in a two-dimensional feature space, where features can have either positive or negative values. An XOR rule classifies as a YES those and only those points for which the product of F_1 and F_2 is negative. Points classified as NO are those for which the produce is positive (because both F_1 and F_2 are positive or because both are negative). Clearly, it is not possible to separate the YES (gray) and NO (white) points in figure 2.5 using a straight line.

Of course, there are other sorts of classification rules besides linear rules. For example, there are inner circular rules as repre-

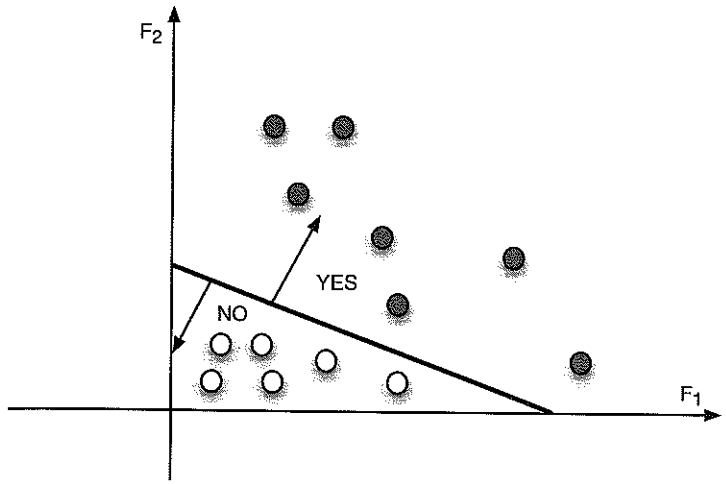

Figure 2.4
Linear classification: Metrocis.

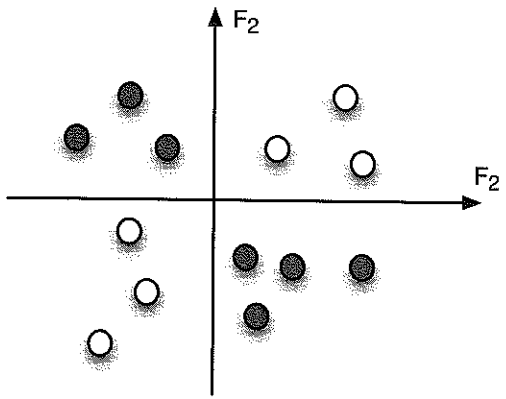

Figure 2.5
XOR representation problem for linear rules.

sented by the insides of circles or hyperspheres in the space. A rule of this sort categorizes all points inside a particular circle or hypersphere as YES and all other points as NO. There are outer circular rules, represented by the outsides of circles or hyperspheres. There are circular rules consisting in both inner and outer circular rules. There are box rules that include both inner box rules and outer box rules. There are quadrant rules that include the rule for XOR.

For any set of sets of points in feature space, there is a corresponding set of classification rules. So, there are many more classification rules than there are linguistic or symbolic representations of classification rules.

It may seem that linear categorization rules will rarely be useful.[1] But linear estimation rules are often quite useful. We noted in our first chapter a number of areas in which linear rules provide better estimates than people can, even experts—predicting the success of medical interventions, predicting criminal recidivism, predicting tomorrow's weather, predicting academic performance, predicting loan and credit risk, predicting the quality of a French wine vintage, to mention only a few (from Bishop and Trout 2005, pp. 13–14).

2.5 Conditions for Satisfactory Enumerative Induction

As we have emphasized, enumerative induction only works given a limited set C of rules. What we would like to know is what has to be true of the set C of rules if enumerative induction is to work no matter what the unknown background probability distribution.

1. Linear categorization rules do play an important role in support vector machines, as is explained in chapter 4, below.

In other words, what has to be true of the set C in order to guarantee that, with probability approaching 1, given more and more data, the expected error for the rules that enumerative induction endorses at each stage will approach the minimum value of expected error for rules in C?

You might wonder whether this sort of convergence isn't guaranteed by the statistical law of large numbers. That principle implies that with probability approaching 1, the empirical error of any particular rule will approach the expected error of that rule, given more and more data. But this is not the same as what is wanted. The trouble is that, given infinitely many rules, as more and more data are taken into account, the rules endorsed by enumerative induction can change infinitely often. Even if the empirical error for each rule approaches a limit, that does not imply anything about the limit of the empirical error of the varying rule endorsed by enumerative induction at each stage.

For example, C could contain a rule c_0 whose expected error is 0.1 and, in addition, an infinite series of rules $c_1, c_2, \ldots, c_n, \ldots$, each of whose expected error is 0.5. There could be possible data so that the following happens. The empirical error of the rule c_i is 0 until the number of data points n exceeds i; thereafter, the empirical error of the rule c_i approaches 0.5. In that case, the empirical error of the varying rule endorsed by enumerative induction at each stage will be 0, but the expected error of the rules made available will always be 0.5. So, the expected error of the rules endorsed at each stage will *not* approach the minimum value of expected error for rules in C, namely 0.1.

What is needed, then, is not just that the empirical error of each rule should converge to its expected error but also that the empirical error of the varying rules endorsed by enumerative induction should approach the value of the expected error of

that rule in the limit. If c_n is a rule endorsed by enumerative induction after n data points, then what is needed is that the empirical error of the rule c_n after n data points should approach the expected error of c_n in the limit. In that case, with probability approaching 1, given more and more data, the expected error of the varying rules endorsed by enumerative induction will approach in the limit the minimum value of expected error for rules in C.

This will happen if (with probability approaching 1) the empirical error of the rules in C converge *uniformly* to their expected error. Let R_c be the expected error of the rule c. Let $\hat{R}_c^n$ be the empirical error of the rule c after n data points. Let $R^n = \max_{c \in C}(|\hat{R}_c^n - R_c|)$ be the maximum value of the absolute difference between the empirical error of a rule in C and its expected error.[2] Then the empirical error of the rules in C converges uniformly to their expected error just in case R^n converges to 0 as $n \to \infty$.

What has to be true of the set of rules C for such uniform convergence? Vapnik and Chervonenkis (1968) show (in effect) that this condition is met for classification rules if and only if the set of classification rules C is not too rich, where the richness of the set is measured by what has come to be called its "VC dimension." (Results with a similar flavor hold for real variable estimation rules with suitably modified notions of dimension, but here we will discuss the result only for classification rules.)

Suppose that some set of N points in the feature space is *shattered* by rules in C in the sense that, for any possible labeling of

2. Strictly speaking, we should use the supremum (sup), or least upper bound, rather than the maximum (max) here, because with infinitely many rules in C the maximum value of the difference may not be defined.

those points, some rule in C perfectly fits the points so labeled. Then the VC dimension of the set of rules C is at least N. More specifically, the VC dimension of a set of rules C is the largest number N such that some set of N points in feature space is shattered by rules in C. If a set of rules does not have a finite VC dimension—because for any number N there is a set of N points shattered by rules in C—then the set of rules C has infinite VC dimension.

Notice that the definition of VC dimension refers to *some* set of N points being shattered, not to *all* sets of N points being shattered. Consider the set of all linear classifications of points in the plane where the YESes and NOs are separated by a straight line. The VC dimension of this set of classification rules is 3, because some set of three points in the plane can be shattered by this class of rules and no set of four points can be shattered. Three collinear points (i.e., three points on the same straight line) cannot be shattered by this class of rules, because there is no such rule that can classify the middle point as a YES and the outer points as NOs (figure 2.6). But three points that are not collinear can be shattered because, for example, any two can be separated from the third by a straight line (figure 2.7). So, the VC dimension of these linear separations is at least 3. And no four points can be shattered by this class of rules, so the VC dimension of these linear rules is exactly 3. (If any three of the

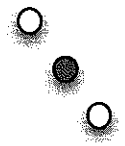

Figure 2.6
Three collinear points cannot be shattered.

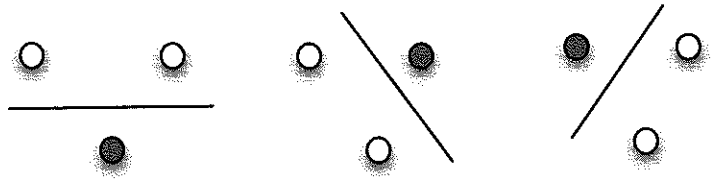

Figure 2.7
Shattering three noncollinear points in the plane.

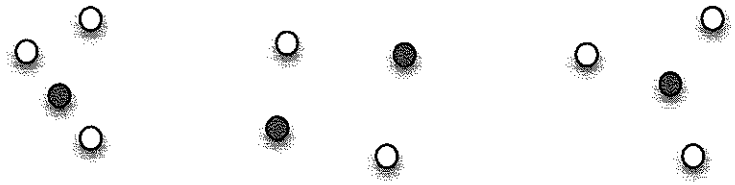

Figure 2.8
No set of four points can be shattered.

four points are collinear, the four points cannot be shattered. Otherwise, either none of the points is within the triangle defined by the other three or one of them is. Figure 2.8 indicates labelings that cannot be separated in those cases by linear rules.)

Some other examples: The VC dimension of the set of all linear separations in D-dimensional spaces is $D + 1$. The VC dimension of the set of all inner rectangles in the plane is 4. The VC dimension of the set of all unions of rectangles in the plane is infinite.

So, that is what the VC dimension comes to. Vapnik and Chervonenkis (1968) show, roughly, that enumerative induction is guaranteed to work no matter what the background probability distribution if and only if the classification rules in C have a finite VC dimension. More precisely (subject to some very mild but technical regularity conditions):

misrepresentation — focus on convergence in part misses point that method is aiming at small samples

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
the rules in C have a finite VC dimension.

Half of this result is that, if the classification rules in C *do not* have a finite VC dimension, then no matter how many data points are provided, there will be probability distributions for which enumerative induction will not select only rules with expected error close to the minimum for rules in C. To see this, consider what can be expected after obtaining n items of data and let $K = 1{,}000{,}000 \times n$. Since the rules in C do not have a finite VC dimension, there is a set of K points in the feature space that is shattered by rules in C. Consider some probability distribution that assigns probability $1/K$ to each member of some such set of K points shattered by rules in C. Any subset of those points will of course also be shattered by those rules.

So, if C does not have a finite VC dimension, then for any n items of data, there are probability distributions that guarantee that there are rules in C fitting whatever data are obtained but giving all possible verdicts on all other points that might come up, where the probability that one of these other points comes up in any given case is very close to 1. (The probability that one of the data points comes up again in any given case is 1/1,000,000.)

This is true no matter how large n is. So it is not true that, with probability approaching 1, the expected error of the rules that enumerative induction leads to will approach the minimum

expected error of rules in C *no matter what the background probability distribution.*

The other half of Vapnik and Chervonenkis' (1968) result is that if the rules in C *do* have a finite VC dimension, then, with probability approaching 1, the expected error of the rules endorsed by enumerative induction will approach the minimum expected error of rules in C no matter what the background probability distribution. If the rules in C have VC dimension V, there is a function $m(V, \epsilon, \delta)$ that indicates the maximum amount of data needed (no matter what the unknown background 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.

Where there is such a function $m(V, \epsilon, \delta)$ there is "probably approximately correct" learning, or PAC learning (terminology due to Valiant 1984). Here a smaller ϵ indicates a better approximation to the minimum expected 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.

2.6 Popper

Vapnik (2000) sees an interesting relation between the role of VC dimension in this 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 be inconsistent 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 Corfield, Schölkopf, and Vapnik (2005) observe that Popper's measure of degree of difficulty of falsifiability of a class of hypotheses does not correspond to VC dimension. 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 for VC dimension, as discussed in the next chapter.

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.

2.7 Summary

In this chapter we have continued our treatment of the problem of induction as a problem in statistical learning theory. We have distinguished inductive classification from inductive real variable estimation. The inductive classification problem is that of assessing inductive methods for using data to arrive at a reliable rule for classifying new cases on the basis of certain values of features of those new cases. We introduced the notion of a D-dimensional feature space, each point in the feature space representing a certain set of feature values. We assumed an unknown probability distribution that is responsible for our encounter with objects and for the correlations between feature values of objects and their correct classifications. The probability distribution determines the best rule of classification, namely the Bayes Rule that minimizes expected error.

For the special case of a YES/NO classification, we can identify a classification rule with a set of points in feature space, perhaps certain scattered areas or hypervolumes. For example, linear rules divide the space into two regions separated by a line or plane or hyperplane.

The real variable estimation problem is that of assessing inductive methods for using data about the value of a real variable

given certain observed features to arrive at a reliable estimate of the value of the real variable.

Enumerative induction endorses that rule or those rules from a certain set C of rules that minimize error on the data. If enumerative induction is to be useful at all, there have to be significant limits on the rules included in C. So C may fail to contain any rule with expected error comparable to the Bayes Rule. So, we cannot expect enumerative induction to endorse a rule with expected error close to the Bayes Rule. At best it will endorse a rule with expected error close to the minimum for rules in C. And, in fact, we have to settle for its probably endorsing a rule close to the minimum for rules in C.

Vapnik and Chervonenkis (1968) show that for inductive classification, 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* the rules in C have a finite VC dimension. (A similar result holds for inductive real variable estimation.)

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. This points to a possible relationship between the role of VC dimension in learning by enumerative induction and the role of falsifiability in Karl Popper's methodology, a relationship to be discussed further in the next 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 (1953) "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." In a context in which all errors are equally bad, 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.

Although this is a very nice result, it is also worrisome, because if C has a 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

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

even if the Bayes Rule has an error rate of 0. Even if C contains many rules and has a large VC dimension, the best rule in C may have an error rate close to .5, 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 a 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 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 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

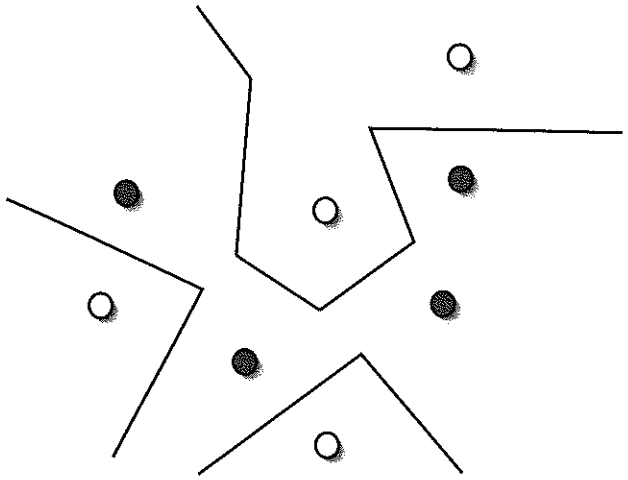

Figure 3.1
Nearest neighbor classification.

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 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 $k_n/n \to 0$), 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 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 that sets an upper bound on how much data is needed to achieve a certain performance, 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 this 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. Various other sorts of ordering have been proposed (e.g., Rissanen 1978; Barron et al. 1998; Chaitin 1974; Akaike 1974; Blum and Blum 1975; Gold 1967; Solomonoff 1964).

One alternative type 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. 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. Whereas 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 converted 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 Estimating a Real Variable 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 estimating the value of a real variable and note how structural risk minimization applies.

In real value estimation, the task is to estimate the value of a variable given the values of each of D observed features. The variable in question may or may not depend on all of the features and may depend on other quantities as well. We assume that there is a background probability distribution that specifies the probability relationship between values of the features and possible observed values of the function. We represent each of the D observable features using a D-dimensional feature space. A possible rule for estimating the value of the variable 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 variable using a single feature. As we have already discussed, any function estimating the variable has an error determined by the background probability distribution.

Each datum can be represented as a point in the plane, where the x coordinate represents the value of the observable feature and the y coordinate represents the value of the variable the datum provides for that observed feature. The task is to estimate the value of the variable for other values of the feature 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

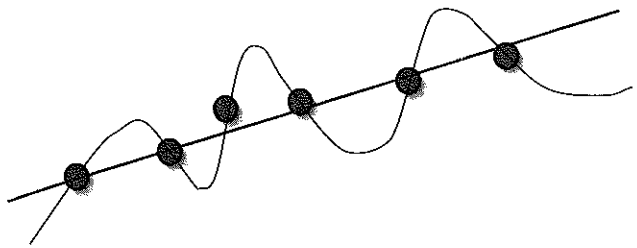

Figure 3.2
Curve fitting.

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 1953).

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, 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 is 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 correctly fit the data but 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 induction to treat all possible hypotheses equally. 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 intuitively 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 that 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 that 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 does, it is incompatible with (2) and with structural risk minimization. Claim (4) is at best terminological and may be just 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 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 one 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

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

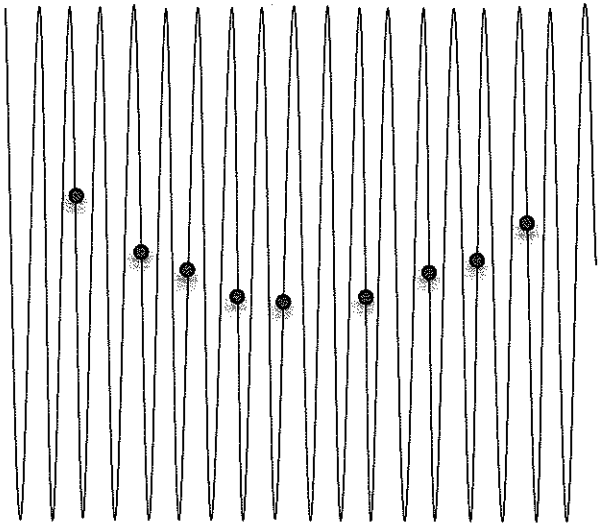

Figure 3.3
Real variable estimation using sine curves.

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 $y = a \sin(bx)$. For almost every 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 G, are empirically equivalent. For example, where H is some highly regarded scientific hypothesis, let G be the corresponding demonic hypothesis that a powerful godlike 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 G?

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 two 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, and so on. 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 G 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 G, perhaps because the class containing H has a lower VC dimension than the class containing G. The relevant class containing G 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 G. From this perspective, there is reason to prefer H over G even though they are empirically equivalent.

So, in fact we may have reason to think that we are not 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 which 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.

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

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 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 to real variable 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.

4 Neural Networks, Support Vector Machines, and Transduction

4.1 Introduction

In our three previous chapters we discussed methods of induction that arrive at general rules of classification on the basis of empirical data. We contrasted enumerative induction with nearest neighbor induction and with methods of induction that balance empirical risk against some sort of ordering of hypotheses, including structural risk minimization in which classes of hypotheses are ordered by their VC dimension. We compared results about these methods with philosophical discussions by Nelson Goodman and Karl Popper.

In this final chapter, we briefly sketch some applications of statistical learning theory to machine learning, including perceptrons, feed-forward neural networks, and support vector machines. We consider whether support vector machines might provide a useful psychological model for human categorization. We describe recent research on "transduction." Where induction uses labeled data to come up with rules of classification, transduction also uses the information that certain new unlabeled cases have come up.