

Falsificationism and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions

David Corfield · Bernhard Schölkopf · Vladimir Vapnik

Published online: 20 August 2009
© Springer Science+Business Media B.V. 2009

Abstract We compare Karl Popper's ideas concerning the falsifiability of a theory with similar notions from the part of statistical learning theory known as *VC-theory*. Popper's notion of the dimension of a theory is contrasted with the apparently very similar VC-dimension. Having located some divergences, we discuss how best to view Popper's work from the perspective of statistical learning theory, either as a precursor or as aiming to capture a different learning activity.

Keywords Induction · Popper · Statistical learning theory · Falsification · Dimension of a theory

1 Introduction

Over the past decade and a half, a paradigm shift has occurred in the field of machine learning. Through this period a range of new techniques has been introduced which allow for the inductive treatment of extremely high dimensional data, something traditional statistical methods typically fail to achieve. Data with possibly many thousands of attributes, such as the values of the pixels of a digital photograph, can be handled by powerful classifiers to allow accurate labelling. Rather than forming a model to represent the data, this style of machine learning aims simply to be able to discriminate between inputs in order to label them correctly. An excellent classifier, one able, say, to distinguish well

D. Corfield (✉)
School of European Culture and Languages (SECL), University of Kent,
Canterbury CT2 7NZ, UK
e-mail: d.corfield@kent.ac.uk

B. Schölkopf
Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany
e-mail: bs@tuebingen.mpg.de

V. Vapnik
NEC Laboratories America, Inc., 4 Independence Way, Suite 200, Princeton, NJ 08540, USA
e-mail: vlad@nec-labs.com

between handwritten 7s and 8s, or between male and female faces, or between gene expression microarrays associated with diseases, embodies no theory of the entities it is classifying beyond the fact that each belongs to one of a number of distinct classes. If we consider the generation of such classifiers as cases of *inductive learning*, then the term's scope has been greatly expanded beyond the production of scientific laws and models whose study has for so long been central to its philosophical treatment.

The theoretical basis for these discriminative classifiers is known as *statistical learning theory*. Here, as elsewhere in inductive learning, there is an important balance to be struck between accuracy and overfitting. Overfitting occurs when too rich a space of hypotheses is used to represent a data set. An example familiar to philosophers is the case of regression where a curve of high degree is fitted to a small number of data points. There is then no expectation that this curve will make good predictions on new data points, while a more loosely fitting straight line may be trusted.

Now, in statistical learning theory the richness of the hypothesis space is not controlled by the degree of a curve or the number of parameters of a hypothesis, but by a construction known as the *VC-dimension*. This dimension, as we shall see, can be thought of as measuring a degree of falsifiability. It is not surprising then that Karl Popper's name crops up in discussions of statistical learning theory. (See, for example, the textbook Vapnik 1995). In this paper we shall examine Popper's ideas in the light of statistical learning theory. We shall discuss differences we have found between Popper's idea of falsifiability and the VC-dimension, and explain how Popper might have viewed these differences.

Readers whose interest in statistical learning is piqued by this article may be interested in the philosophical book length study *Reliable Reasoning* (Harman and Kulkarni 2007), which deals with many of the issues presented here, and indeed which discusses the technical report from which this article derives.

2 The VC-Dimension

The simplest setting for statistical learning theory is where we represent a data distribution as a set of points, labelled either '+1' or '-1', in a Euclidean space, and aim to find a classifying hypersurface which separates the two classes as accurately as possible. Imagine, for instance, that the weight, height and waist size of a population of people is represented in a three-dimensional space, and we aim to separate the females from the males. A training set is given, corresponding to a sample of the population. The data from this sample is plotted, and each point labelled '+1' or '-1' according to whether the individual is female or male. We now generate a two-dimensional surface which we hope will separate the whole population accurately. We need to choose a surface which corresponds to the best balance between accurately separating the training sample while avoiding overfitting. Clearly, allowing just any surface, we will be able to find many which classify the training sample perfectly, at least if no two individuals in the sample of opposite gender have precisely the same measurements. Overfitting is avoided by restricting the class of surfaces from which one is to be chosen, or in other words by limiting the *capacity* of the class of hypotheses to discriminate. This capacity is measured by the Vapnik-Chervonenkis, or VC-, dimension (Vapnik 1995).

The VC-dimension of a set of classifying hypotheses is the largest natural number, n , such that there is a set of n distinct points, for which, however, they are labelled '+' or '-', there is a hypothesis which agrees with this labelling. The set of points is said to be *shattered* by the hypothesis class. For example, the set of half-planes has VC-dimension

three because any set of three non-collinear points is shattered by these classifiers, while any set of four distinct points is not shatterable by the same class. To see this last point, think of four points configured to lie on the corners of a square, such that one pair at opposite ends of a diagonal are marked '+1', while the other pair are marked '-1'. No straight line can separate this data. One dimension higher, as in our weight, height and waist size example, the set of half-spaces has VC-dimension four. If four points are selected which do not lie on the same plane, then whatever the genders of the people corresponding to these points, a plane could be found to separate the males from the females.

Now, the VC-dimension is used in the following kind of way. Someone proposes that there is a linear classifier (a separating plane) which can accurately separate the males from the females within this population, using the body measurements as inputs. Then, if we take our large sample and find an accurate linear classifier, VC-theory gives us results of the form: with probability x % the classifier will be at least y % accurate on future observations if drawn independently from the same distribution, where y depends not only on x , but also on the training error, the sample size and the VC-dimension. If the VC-dimension is not finite, no such guarantee can be given. So, even though a classifier is being found with the help of all of our observations, we may still be able to give ourselves probabilistic assurances of future accuracy rates without needing to consider further data. Higher reaches of statistical learning theory have principled ways of treating more complicated protocols of training and testing with data in the context of both classification and regression tasks.

The idea of a set of points being shattered by a class of hypotheses may bring to mind Karl Popper's notion of non-falsifiability in the following sense. If the class of hypotheses is too rich, in the sense of having too great a capacity to discriminate, then whatever the data, a perfectly accurate classifier could be found. For example, you cannot falsify the theory that a linear classifier can separate the genders by looking at a sample of size four, however they are labelled. It will pay us then to look more closely at Popper's writings to see if he says anything along these lines.

3 The Logic of Scientific Discovery

In his 1934 book *Logik der Forschung*, first translated into English in 1959, Popper marked out his differences from members of the *Vienna Circle* by maintaining that the single factor distinguishing science from other claims to knowledge was its preparedness to risk falsification. Taking Albert Einstein as a paradigmatic scientist, Popper was impressed by Einstein's willingness to make bold conjectures with testable predictions. Starlight would be found to bend as it travelled to our telescopes past the sun, and bend by a specified amount. If, within experimental error, this precise bending was not found to have taken place, Einstein would be prepared to give up on his general theory of relativity, given that he was assured that the observations had been conducted properly.

On the other hand, if the right amount of bending was observed, this would not *confirm* the theory. What was at stake for Popper was simply a point of logic: synthetic universal statements (with infinite domain) can be falsified, but cannot be verified empirically. Popper could make no sense of the idea that a scientific theory becomes more probable when a prediction is verified. Rather, all we can say is that the theory has been severely tested, or *well-corroborated*. The Vienna Circle, and associated logical empiricists, meanwhile, did not agree with this asymmetric treatment of verification and falsification,

and developed a theory of confirmation in a probabilistic framework. Much of *The Logic of Scientific Discovery* is taken up with a critique of the use of probabilistic representations of states of knowledge. Popper has thus become a rallying point for anti-Bayesian philosophers of science.

In the development of his conception of scientific learning, Popper needed a way to compare theories as to their riskiness, their potential to be falsified. He did so by describing two methods for comparing theories:

3.1 The containment relation between classes of falsifiers

This is determined by the degree of universality and the degree of precision (of predicate and of measurement) of the theory. So “all planets move in ellipses” is less universal and less precise than “all heavenly bodies move in circles”. The collection of falsifiers of the first theory is contained within the collection of falsifiers of the second theory, so the latter is more falsifiable.

3.2 The dimension of a theory

If there exists, for a theory t , a field of singular (but not necessarily basic) statements such that, for some number d , the theory cannot be falsified by any d -tuple of the field, although it can be falsified by certain $(d + 1)$ -tuples, then we call d the characteristic number of the theory with respect to that field. All statements of the field whose degree of composition is less than d , or equal to d , are then compatible with the theory, and permitted by it, irrespective of their content. (Popper 1959, 130)

At first glance Popper’s dimension of a theory seems to be a precursor of the VC-dimension. But let us look more closely. The point about statements being “singular (but not necessarily basic)” is intended in the sense that a recording of the position of an event in d -dimensional Euclidean space would count as one statement, rather than as d statements concerning its co-ordinates. This agrees with the VC-dimension set-up, so let us calculate the dimension of the theory which states that labelled points in the plane can be separated by a line. Clearly any two distinct points can be separated. However, there are configurations of three points and labellings which cannot be separated, for example, three collinear points labelled ‘+’ ‘−’ ‘+’. The existence of this non-generic configuration means that the Popper dimension of the theory is two. And indeed the Popper dimension of hyperplane classifiers in any d -dimensional space is two, where the VC-dimension can be shown to be $(d + 1)$. The VC-dimension is the largest number of points one can shatter, the Popper dimension is one less than the smallest number of points one can not shatter.

4 What to make of Popper

There are two ways to respond to this discrepancy we have observed:

- (a) First, we might say that Popper simply made a mistake.
- (b) Second, we might say that Popper is simply not interested in the same problems as Vapnik and Chervonenkis.

(a) One could argue that Popper was just being imprecise in his definition. If he had insisted that points could not be specified with complete accuracy, then one could not

observe perfectly collinear points, and the Popper-dimension for separating half-spaces would equal the VC-dimension.¹ However, Popper's treatment of a different situation suggests he had not thought through his definition. We find that he gives the dimension of the theory that "All planets move in ellipses" as five, although, strictly speaking, the observation of three collinear points on the orbit would falsify the theory. Allowing for a margin of error in the observations, however, three points could be accommodated.² But still, very generous margins of error would be necessary to allow some configurations of four points to be situated on an ellipse. So strictly speaking, the Popper-dimension of this theory is two, and disallowing perfect precision of observations it is three.³

This relates to a more general mistake concerning his treatment of simplicity and dimension. Popper claims: "In algebraic representation, the dimension of a set of curves depends upon the number of parameters whose values we may freely choose. We can therefore say that the number of freely determinable parameters of a set of curves by which a theory is represented is characteristic for the degree of falsifiability (or testability) of that theory." (Popper 1959, 131). So he might identify simplicity either with degree of falsifiability or number of parameters. But he comes down in the end in favour of the former:

The epistemological questions which arise in connection with the concept of simplicity can all be answered if we equate this concept with degree of falsifiability. (Popper 1959, 140)

Above all, our theory explains why simplicity is so highly desirable. To understand this there is no need for us to assume a 'principle of economy of thought' or anything of the kind. Simple statements, if knowledge is our object, are to be prized more highly than the less simple ones because they tell us more; because their empirical content is greater; and because they are better testable. (Popper 1959, 142)

So he appears to have decided in favour of the degree of falsifiability as to the strength of a theory, although confusingly he equates this to simplicity. However, he then continues to describe sine oscillations as simple (Popper 1959, 143), giving no thought to the number of points needed to falsify this class. The use of the VC-dimension in statistical learning theory runs against the idea that the generalisability of a theory goes along with its simplicity, calculated in terms of the number of its parameters. Standard examples exist of hypothesis classes with low VC-dimension and a very large number of parameters (e.g., support vector machines), and vice versa. The sine curves Popper saw as easily falsifiable are often given as a paradigmatic example of the latter case. Indeed, the set of classifiers $\{\text{sign}(\sin bx) : b > 0\}$ shatters any finite set of distinct points distributed along the x -axis, although they are governed by just one parameter.⁴

(b) Let us turn now to the more charitable reading of Popper—that his concerns are not those of statistical learning theory. Where statistical learning theory operates in situations where the data we receive is recorded passively, with a view to giving us confidence that we can expect classification or regression to continue to work accurately if further incoming data is independent and from the same distribution, Popper is interested in the

¹ On the other hand, these dimensions would diverge if a minimum margin were required.

² Even here, though, configurations of three very nearly collinear points might require an ellipse with a very large major axis relative to the distances between the points.

³ Peter Turney (1991) has also located this confusion on Popper's part.

⁴ One might try to save Popper again by requiring a minimum margin to a sinusoidal classifier, or to limit the precision with which a parameter could be specified.

situation confronting the bold research scientist, who will leave no stone unturned in their quest for truth:

...our methodological decision—sometimes metaphysically interpreted as the principle of causality—is to leave nothing unexplained, i.e. always to try to deduce statements from others of higher universality. This decision is derived from the demand for the highest attainable degree of universality and precision, and it can be reduced to the demand, or rule, that preference should be given to those theories which can be most severely tested. (Popper 1959, 123)

Rather than happily anticipating future accurate predictions on new data, Popperian scientists go out of their way to find the weaknesses of their theory. In other words, Popper appears to be more interested in what is called active learning, and indeed the active learning of generative models. This isn't about settling for a classifier for which we have probabilistic guarantees that its future error rate will be less than some figure. Rather, theories are there to be shot down. Let's illustrate this point by returning to our example of the weight, height and waist size of a population. We mentioned above that the VC-dimension of the class of linear classifiers is four. The Popper-dimension for this class by contrast is only two according to his definition if infinite accuracy is allowed. Where VC-theory gave results about our expectation of an accurate classifier continuing to perform well on unseen data, Popperian scientists dealing with the proposed hypothesis are looking to falsify it. Having found a woman with measurements (60 kg, 160 cm, 80 cm) and a man with measurements (70 kg, 170 cm, 90 cm), if they can find a woman whose measurements are, for instance, (80 kg, 180 cm, 100 cm), they will have succeeded in falsifying the hypothesis.

Reading Popper as interested in 'active learning' makes sense of a further discrepancy between Popper and VC-theory. Recall that Popper's first criterion for comparative falsifiability states that if the falsifiers of one theory are contained within those of another, the latter is more falsifiable. Increasing the domain of a theory allows further opportunities for falsification to occur. In VC-theory, on the other hand, when the domain of a set of functions is increased, its VC-dimension can only increase since there are more opportunities to find shatterable configurations, hence the set becomes less falsifiable.⁵

A neat way of summarising the situation, then, might be to congratulate Popper for his original idea of falsifiability, then to say that it acts as a *motif*, appearing in different guises in the kind of generative learning that is science and the kind that is discriminative machine learning. The question then rises as to how closely related these varieties of learning are. If we could imagine Popper being dismissive about the latter, it should be remembered that we don't usually opt for it without good reason. In many situations, and it looks as though these can only become more common, we simply have no choice but to use discriminative techniques if we wish to gain any kind of capacity to find patterns in data sets.

Popper saw science as a quest for truth, its theories not subject to static appraisal. As we mentioned above, Popper was strongly opposed to any notion of confirmation:

Like inductive logic in general, the theory of the probability of hypotheses seems to have arisen through a confusion of psychological with logical questions. Admittedly, our subjective feelings of conviction are of different intensities, and the degree of confidence with which we await the fulfilment of a prediction and the further

⁵ This difference is made less decisive if one considers instead the VC-entropy, a more subtle measure of capacity than the VC-dimension (Vapnik 1995). Since the VC-entropy depends on the distribution, it is not possible to say in general how it is affected by changes of the domain.

corroboration of a hypothesis is likely to depend, among other things, upon the way in which this hypothesis has stood up to tests so far—upon its past corroboration. But that these psychological questions do not belong to epistemology or methodology is pretty well acknowledged even by believers in probability logic. (Note: I am alluding here to the school of Reichenbach rather than to Keynes.) (Popper 1959, 255)

And this isn't to make a distinction between the probable truth of a hypothesis and its probable accuracy. Nowhere in *The Logic of Scientific Discovery* does he say anything about how reliable our well-corroborated theories are as we travel the path towards truth, i.e., how much we can trust their predictions.

Now this is rather peculiar. Can machine learning and scientific learning be so very different? Surely we rely on what our scientists tell us are our best confirmed theories. For example, we believe their predictions of the next solar eclipse. This isn't just blind faith. We seem here to be treating scientific theories as we might the hypotheses in the output of a machine learning algorithm, i.e., as reliable predictors. Whence does this confidence arise?

In 1963 Popper introduced the concept of *verisimilitude* in his book *Conjectures and Refutations* (Popper 1963). With this concept he was attempting to measure the truth-likeness of a theory by measuring the difference between its truth content and its falsity content. We need not enter into details here as this idea has almost universally been seen to have failed,⁶ but it does point to an interest on Popper's part in relations between the successes and failures of a theory. In view of the fact that statistical learning theory has developed such a relation in the form of the *generalization error*, we might speculate that something akin to this could have filled the lacuna in Popper's theories in a way which might have satisfied him, avoiding as it does the notion of the probable truth of a theory. What we have instead is a theory's probable accuracy. Although, as elsewhere in statistics, there are Bayesian versions of probable accuracy, the idea can be treated in a frequentist way, and, to move closer to Popper, there would seem to be no obstacles to a propensity theoretic interpretation.⁷ Furthermore, active learning is a branch of statistical learning theory, although there have been far fewer theoretical achievements here.

5 Conclusion

It seems clear that Popper can only be rescued from the charge that he made serious mistakes in *The Logic of Scientific Discovery* by generous interpretation, evidence that he had any inkling of the need for which is lacking. However, we do not think it right merely to say that Popper had a partial insight into an important component of statistical learning theory. His first concern was with a different kind of learning, that conducted by the 'bold and daring' research scientist. This having been said, however, there appears to be no way for Popper to speak about the reliability of well-tested theories, and yet one surely needs to be able to speak of one's confidence in the future predictions of such theories. It is precisely on this issue that we should expect the similarities between scientific and machine learning to emerge. We suggest that statistical learning theory might have satisfied him on this score.

⁶ See Tichy (1974) and subsequent articles in the same edition of the BJPS.

⁷ See Gillies (2000) for a comparison of the different interpretations of probability theory, including Popper's propensity interpretation.

Acknowledgments David Corfield would like to thank the Max Planck Society for financial support during the writing of this article.

References

- Gillies, D. (2000). *Philosophical theories of probability*. London: Routledge.
- Harman, G., & Kulkarni, S. (2007). *Reliable reasoning: Induction and statistical learning theory*. Cambridge: MIT Press.
- Popper, K. (1959). *The logic of scientific discovery*, Hutchinson, translation of *Logik der Forschung*, 1934.
- Popper, K. (1963). *Conjectures and refutations*. London: Routledge.
- Tichy, P. (1974). On Popper's definitions of verisimilitude. *British Journal of Philosophy of Science*, 25, 155–160.
- Turney, P. (1991). A note on Popper's equation of simplicity with falsifiability. *British Journal for the Philosophy of Science*, 42, 105–109.
- Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.