Aspects of Inductive Inference in Statistics and Machine Learning

Palash Sarkar

Introduction

Acquisition of knowledge constitutes a fundamental human activity. To address this question, it may appear that at the outset one should put down a definition of knowledge. We will, however, side-step the issue of defining knowledge. Instead, we will assume that people have at least some idea of what constitutes knowledge. The issue that we consider is methods for acquiring knowledge. Even without having a definition of knowledge, one would agree that for the acquired knowledge to be reliable, the methods for acquiring knowledge should also be reliable. In other words, if a method for acquiring knowledge is not reliable, then the acquired knowledge cannot also be considered reliable.[1]

The question of what constitutes reliable methods for acquiring knowledge has been considered by philosophers since ancient times. Typically, discussions in philosophy books talk about valid methods of acquiring knowledge. For the purpose of this article, we will conflate 'reliable' with 'valid'.[2] The methods of acquiring knowledge that have been identified by philosophers are the following: perception; inductive inference; deductive inference; analogy and comparison; and, testimony (of authority).

[1]There are several issues here. For example, is it proper to talk about 'unreliable knowledge'; i.e., if something is unreliable, can it be considered to be knowledge? A related issue is that of quantifying unreliable knowledge by assigning a score of unreliability. Of course, a statement such as 'event X holds with 80% probability' is a definite statement about uncertainty. While these are interesting questions, discussion of these issues are outside the scope of this article.

[2]The fine distinction between a 'reliable method' and a 'valid method' and by implication the fine distinction between 'reliable knowledge' and 'valid knowledge' is again outside the scope of this article.

P. Sarkar (✉)
Applied Statistics Unit, Indian Statistical Institute, 203, B.T. Road, Kolkata 700108, India
e-mail: palash@isical.ac.in

A. Chaudhuri et al. (eds.), *A Tribute to the Legend of Professor C. R. Rao*,
Indian Statistical Institute Series,
https://doi.org/10.1007/978-981-33-6991-7_10

Our target of discussion is inductive inference and its relation to statistics and machine learning. Before moving on to the details of inductive inference, we mention a few words about the other methods. Perception primarily refers to the use of sense organs. A deductive inference essentially refers to a statement that can be derived from a set of statements using the rules of logic. Knowledge acquisition by analogy refers to the translation of ideas from one context to another by an appeal to the similarity of the two contexts. Knowledge from testimony is acquired by accepting a statement from some person who can be trusted.

Statistical inference refers to a wide collection of techniques which aim to derive information from observations. The methods of statistical inference involve mathematical and computational methods. The role of mathematics is to formally prove that statistical inference methods achieve their stated goals. This is essentially an application of the deductive inference method mentioned above. Computational methods are useful in actually implementing a statistical inference technique, or, more theoretically, for understanding the difficulties and limitations for the implementation of various inference techniques. The justification or suitability of a particular statistical method, however, is neither mathematical nor computational. Such justification is usually provided as an appeal to intuition.

Machine learning techniques have a similar goal as statistical inference techniques. In many cases, well-studied statistical techniques are classified as machine learning methods. One way in which machine learning perhaps differs from statistics is in the greater emphasis laid on computational issues in machine learning methods. Similar to statistical inference, justifications for various machine learning methods are also usually stated as being intuitive.

In this paper, we aim to provide perspectives on some statistical and machine learning techniques from the viewpoint of inductive inference. To this end, in the first part, we provide an overview of inductive inference and its various characteristics. This will essentially summarise prior thoughts on inductive inference which are relevant to statistics. In the second part, we will look at a number of techniques from statistics and machine learning. The goal will be to point out why such techniques are essentially inductive inferences and in particular which characteristics of the methodology of inductive inference can be found in these techniques.

One may question the usefulness of the present work, especially since it does not provide any new result in statistics and machine learning. A response would be that explicitly seeing inductive inference at work within statistics and machine learning satisfies a basic intellectual curiosity. Perhaps more importantly, fundamental philosophical issues regarding the validity of inductive inferences can also be seen to apply to statistical inferences and machine learning techniques. This would lead to replacing an aura of definiteness by an umbra of doubt or uncertainty. In more concrete terms, explicitly identifying the connections between the two areas will perhaps lead to a more productive two-way flow of ideas.

The relation between inductive inference and statistics is known (but perhaps not as well known as it should possibly be). In 1956, Mahalanobis had commented that ‘statistics is the universal tool of inductive inference’. A short paper by Fisher [6] had commented on the relation between statistical methods and induction. The preface

of the fascinating book on statistical thought by Chatterjee [3] mentions that the book 'views the problem of statistical induction in a wider philosophical context'. More general connections between statistics and philosophy have been explored by a galaxy of authors in a complied volume edited by Bandyopadhyay and Forster [1]. The connection between machine learning and induction has been explored by Harman and Kulkarni [7]. Finally, we would like to mention the book [11] entitled 'Statistics and Truth: Putting Chance to Work', by C. R. Rao which provides an excellent overview of the nature and role of statistics in various fields of human activity. In particular, we note that the preface of the book suggests that statistics provides a method for codifying inductive reasoning.

In view of the above, our work may be seen as a continuation of the line of thought connecting statistics and machine learning to induction and more generally philosophical issues. We would like to mention that, to the best of our knowledge, our approach of considering specific statistical inference and machine learning techniques to bring out in details the features of inductive inference therein is not present in the above-mentioned works. So, our work does offer something new to a reader interested in the connection between inductive inference and statistics/ machine learning.

Inductive Inference

Inductive inference as a method of acquiring knowledge has been studied in both Western and Indian philosophies. Putting down a precise definition of 'inductive inference' is rather difficult. The term refers to a broad set of inference mechanisms which loosely speaking may be construed as inferring something about unperceived situations from perceived information. We illustrate a few inductive inference methods through examples.

Statements such as 'the Sun rises in the East', or, 'all human beings are mortal' are derived based on observations. These are examples of enumerative induction or universal inference, i.e. inference from particular observations to a universal statement. More generally, these are of the following type. Instances a_1, a_2,…, a_n which are all F's are also observed to be G's; from this a general principle 'all *F's* are *G's* is inferred.

Inductive inference need not only be from particular to the universal. For example, from 'all observed rubies have been red' inferring 'the next yet to be found ruby will also be red' is an example of inductive inference where the premise is general and the conclusion is particular.

It is not necessary that an inductive inference will have a universal statement. For example, from 'Mercury is spherical, Venus is spherical, Earth is spherical, …' inferring 'the next yet to be discovered planet will also be spherical' does not involve any universal step. This is called a singular predictive inference which moves from particular premises to a particular conclusion.

The Problem of Induction

In Western philosophy, Hume made the most influential contribution to the study of inductive inference. The most disturbing question about induction is whether it is justified. For example, what are the justifications for the above examples of inductive inference? Clearly, these inferences cannot be justified using deductive methods. For example, the argument.

((all observed F's have also been G's) and (a is an F)) imply (a is a G).

is invalid; i.e., there exists a model in which both the premises are true, but the conclusion is false.

Hume had identified that any justification of inductive inference must necessarily be inductive leading to a circularity of argument or *petitio principii*. This was summed up in the following famous statement by Hume (1738) in 'A Treatise of Human Nature': 'instances, of which we have had no experience, must resemble those, of which we have had experience, and that the course of nature continues always uniformly the same'. This underlines that inductive inferences assume that nature continues uniformly. The statement that nature continues uniformly itself requires justification, and this justification can only be obtained through induction leading to a circularity of argument.

The problem of induction is the question of how to distinguish reliable from unreliable inductive inferences? This is a conundrum which is yet to be satisfactorily resolved despite efforts by philosophers such as Karl Popper and others. Perhaps the question does not even have a resolution. The inability to resolve the problem of induction has at least two implications. The first is methodological; i.e. there is no method or procedure which may be applied to distinguish good from bad inductive inferences. The second question is more fundamental in that there is possibly no objective difference between reliable and unreliable inductive inferences. While the unresolved problem of induction is a philosophical irritant, in practice, inductive inferences are regularly made. We refer to [9] for description of the problem of induction.

The method of induction has been studied in Indian philosophy. The main criticism against inductive inference is that of circularity. This was identified by the Cārvāka school of thought. Chapter 1 of [2] provides an excellent account of the Cārvāka criticism of induction. We also refer to [12] for a description of how the Cārvāka school of thought anticipated some modern notions.

The Principle of Simplicity

Often called Occam's razor, the principle of simplicity is the idea that the simplest among several available options should be chosen. For example, the simplest among several competing hypotheses suggested by observations should be chosen. Another

example would be to choose the simplest among several models. The simplicity principle is ubiquitous in human reasoning.

One may look for an objective justification of the principle of simplicity. This, however, is hard to find. One possible justification could be that this principle has proved to be correct in the past, and so, it can be used in future. Such a justification is essentially an inductive inference. An eloquent criticism of the justification of the principle of simplicity by appealing to the past has been made by Bertrand Russell in the book 'On Scientific Method in Philosophy' (1914). He remarks: 'But it is just this characteristic of simplicity in the laws of nature hitherto discovered which it would be fallacious to generalise, for it is obvious that simplicity has been a part cause of their discovery, and can, therefore, give no ground for the supposition that other undiscovered laws are equally simple'.

There are several other troublesome issues. The principle of simplicity tacitly assumes that the options are known, that it is possible to compare any two options with regard to simplicity, and that the set of options has a unique simplest member. All of these issues can be stated in a more formal framework though we are not aware of any place where such formalisation has been done. Nevertheless, it is not our purpose here to get into a detailed formal investigation of the principle of simplicity. We note two points. In practice, the principle of simplicity is universally applied and that its only possible justification arises as an inductive inference.

Inference to the Best Explanation

Philosophers distinguish between three types of inferences, namely deductive, inductive and abductive [4]. Abductive inference is also called inference to the best explanation (IBE). A standard example of IBE is the following. Suppose on waking up in the morning, one finds the outside to be wet. From this, one infers that it had rained in the night. This inference is the best one which explains the observation. In theory, it is possible to make other inferences such as water was sprayed from a low flying aircraft, but, would not be considered the best inference. Of course, the inference that it rained in the night is also the simplest of explanations, so this particular example is also an example of the application of the principle of simplicity.

The idea of abductive inference or abductive logic was introduced by Charles Sanders Peirce. He considered abductive inference to be a form of non-deductive inference which is different from inductive inference. The notion of abductive inference has been closely studied. We refer to [4] for an introduction to the various issues.

The validity of IBE can be questioned. As in the case of the principle of simplicity, one may ask for an objective justification of IBE. Such a justification may be provided by considering past applications; i.e., IBE has proved to be true in the past and so it will be true in the future. This is again an appeal to inductive inference. So, while IBE is considered to be different from inductive inference, its justification seems to

rely on inductive inference. So, IBE (and also the principle of simplicity) may be considered to be second-order induction.

Pragmatism

Roughly speaking, pragmatism refers to the idea that among various options, choose the one which is most useful. It may turn out that the most useful is also the simplest or the best. For example, among various techniques that in theory can be employed to analyse a situation, use the one which is the easiest to apply.

As in the case of simplicity and IBE, justification for pragmatism arises from an appeal to induction. Further, the issue of determining the most useful option has difficulties similar to that of determining the simplest or the best option.

Features of Inductive Inference

It has been mentioned that no objective justification of inductive inference has been found till date. Nevertheless, investigations have identified several features that can be seen in various inductive inferences. We briefly discuss these below.

Ampliative: This is intended to mean that the conclusion of the inference has more content that its premise. For example, in universal inference, the premise consists of some observations while the conclusion is a universal statement. This is a distinctive feature of inductive inference as opposed to deductive inference in which there is no amplification of the logical content of the conclusion beyond what is contained in the premises.

Contingent: The conclusion of an inductive inference does not follow as a necessary condition of its premise. In other words, it cannot be logically said that if the premise holds then the conclusion must also hold.

Non-monotonic: Inductive inference is based upon perceived information. An inference which is made from some amount of perceived information may become invalid if additional information becomes available.

Non-preservation of truth: It is possible that the premises of an inductive inference are true, yet the conclusion is false. For example, in an enumerative induction, the individual premises are observations and are true. The universal conclusion, however, could be false since it may not hold for some hitherto unobserved instance.

Statistical Methods and Inductive Inference

In this section, we consider some basic statistical notions and point out aspects of induction that are implicit in such notions. This provides a better understanding of

the link between the more philosophical notion of inductive inference and statistical methods.

Sufficient statistic: The notion of sufficient statistic is a basic notion in statistical inference. Given observations $X_1, \ldots, X_n$ following some known distribution with an unknown parameter θ, a statistic T is a function $T\ (X_1, \ldots, X_n)$ of the observations. In the words of Fisher [5], a statistic is sufficient for an unknown parameter if 'no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter'. The notion of sufficient statistic provides a simple example of non-monotonicity. It is easy to construct examples where a statistic T is sufficient for a parameter using observations $X_1, \ldots, X_n$, but is no longer sufficient if an additional observation X_{n+1} becomes available. So, it is important to use all available data and it is usually assumed that the number of observations is known and fixed.

Maximum likelihood estimate (MLE): Given data $x_1, \ldots, x_n$, the likelihood function $L(\theta; x_1,\ldots,x_n)$ is a function of an unknown parameter θ which gives the probability of obtaining the sample $x_1,\ldots,x_n$ given the value of the parameter. Once the data is available, an estimate of the value of the parameter θ is desired. The MLE $\hat{\theta}$ of θ is the value which maximises the likelihood function. In other words, for $\theta = \hat{\theta}$, the probability of observing the data $x_1, \ldots, x_n$ is maximised. The justification for using MLE is implicitly based on abduction, or inference to the best explanation (IBE). The rationale is that since $\hat{\theta}$ maximises the probability of observing the data, it is the best explanation for observing the data.

Null hypothesis testing: A null hypothesis H_0 to be tested is formulated. This formulation involves defining a test statistic. A number of observations are made which provide the data using which H_0 is to be tested. The p-value of the test is the probability that under H_0 the test statistic equals the observation or more extreme. The null hypothesis H_0 is rejected at α level of significance if the p-value is less than α.

Several features of inductive inference can be identified in the procedure. First, the procedure is ampliative. The premise of the inference mechanism is the data while the conclusion is about the hypothesis. So, the conclusion has more content than the premise. Second, null hypothesis testing is non-monotonic; i.e., a hypothesis which was not priorly rejected can become rejected with the availability of additional data. So, the non-rejection of H_0 does not imply it is established. In the words of Fisher [6], '[i]t is a fallacy, …, to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened'. The third aspect of induction arises in the choice of α. Justification for choosing a value for α is based on this value being used in various other situations. This justification is essentially an inductive inference that the value of α which has been appropriate in other situations will also be appropriate for the situation at hand.

Prediction Error: Let X be a real-valued input random variable, i.e. a predictor or a feature, and let Y be a real-valued output random variable, i.e. the response or the dependent variable. Let Pr[X, Y] be the joint distribution of X and Y. A basic statistical technique is regression, i.e. to obtain a function $f(X)$ of X which can be used to predict Y given X. A loss function is used to measure the efficacy of the prediction.

The most common and convenient loss function is the squared error loss, i.e. $L(Y,f(X)) = (Y - f\ (X))^2$. Using this loss function, the expected (squared) prediction error is defined to be $\text{EPE}(f) = E(Y - f\ (X))^2$. The goal is to choose f such that EPE(f) is minimised. The EPE can be simplified as $EPE(f) = E_X E_{Y|X}\left[\left(Y - f(X)^2\right)|X\right]$, and so to minimise EPE(f) it is sufficient to perform point-wise minimisation, i.e. $f(x) = argmin_c\ E_{Y|X}\left((Y - c)^2|X = x\right)$. The solution is $f(x) = E(Y|X = x)$. So, we have that the conditional expectation, which is also called the regression function, provides the best prediction of Y at a point x.

The notion of 'best' in the above is with respect to the squared error loss. One may ask for a justification of using the squared error loss. For example, the loss function could have been defined as $|Y - f\ (X)|$, in which case the solution would turn out to be the conditional median, i.e. median($Y\ |X = x$). Is there any a priori reason to prefer squared error over absolute error? A sort of justification forwarded in [8] (the descriptions of squared error and median error are also from [8]) is that 'squared error is analytically convenient and the most popular'. The reason for its popularity is perhaps based on analytical convenience, so the main justification for using the squared error is that it is analytically convenient. This is a pragmatic consideration.

We may take a moment to reflect on qualitative aspects of this issue. Regression forms an important technique of statistical decision theory and machine learning. Outputs of a prediction function will conceivably be used to arrive at decisions which can have major social effects. The decisions and their social consequences then depend upon the actual choice of the prediction function. So, for example, using a prediction function based on squared error can lead to a decision which is different from a decision which is arrived at by using a prediction function based on absolute error. The justification for such a difference in decisions would really be the analytical convenience of the squared error. In other words, the comparative simplicity of being able to mathematically handle one expression over another can lead to wholly different social consequences.

Model Selection: Suppose $x_1, \ldots, x_n$ constitute the data. Further, suppose that there is a set of models $M_1, \ldots, M_m$ and the goal is to choose one of the models based on the data. This is a typical setting of inductive inference, where from particular observations, one infers a general statement. In this particular setting, the inference is somewhat restricted in the sense that the requirement is to choose one among a finite set of models. One may ask as to how the set of models have been determined? The answer would typically be a combination of the following justifications: from previous experience, usefulness, simplicity. All of these justifications themselves are inductive inferences.

A model is determined by its parameters. Suppose the parameter vector for the ith model is $\boldsymbol{\theta}_i$, $i = 1, \ldots, m$. Further, suppose that the dimension of $\boldsymbol{\theta}_i$ is k_i, i.e. the ith model is determined by k_i parameters. Let $L_i(\boldsymbol{\theta}_i; x_1, \ldots, x_n)$ be the likelihood function for the ith model.

Two standard ways of assigning scores to models are the Akaike information criterion (AIC) and Bayesian information criterion (BIC). For the ith model, these are defined as follows. Let $\hat{\theta}_i$ be the MLE for $\boldsymbol{\theta}_i$.

$$\mathrm{AIC}(M_i) = 2k_i - 2\ln L_i(\hat{\boldsymbol{\theta}}_i; x_1, \ldots, x_n);$$
$$\mathrm{BIC}(M_i) = (\ln n)k_i - 2\ln L_i(\hat{\boldsymbol{\theta}}_i; x_1, \ldots, x_n).$$

The procedure for selecting a model using AIC (resp. BIC) is to compute the AIC (resp. BIC) scores for all the models and choose the one with the minimum score. Both the AIC and the BIC scores use the MLE $\hat{\boldsymbol{\theta}}_i$, and also both penalise models having a large number of parameters. The extent of the penalty is smaller for the AIC score than the BIC score.

Implicit in the above definitions of the AIC and BIC scores are two aspects of inductive inference. The first is the use of MLE. As discussed earlier, the justification for using MLE is inference to the best explanation. The second aspect is that of penalising more complex models,, or in other words preferring simpler models. Again, as discussed earlier, simplicity is justified by induction.

The AIC score for the *i*th model is derived through Taylor series approximations of the Kullback–Leibler (KL) divergence of the density of the *i*th model from that of the correct density. The goal of model selection based on the AIC score is essentially to choose the model for which the KL divergence is the minimum. The choice of KL divergence for use in model selection is itself based on induction; i.e., the KL divergence has proved to be useful in various other settings, and so it should also be useful for model selection.

The derivation of the BIC score is based on Bayes theorem. Suppose a prior probability p_i is assigned to model M_i. Also, consider the observations to be random variables X_1,…,X_n drawn from an unknown distribution. From Bayes theorem, $\Pr(M_i|X_1, \ldots, X_n)$ is proportional to $\Pr(X_1, \ldots, X_n|M_i)p_i$. The BIC score is arrived at through approximations of the last expression. The goal of the BIC score is to maximise the probability $\Pr(M_i|X_1, \ldots, X_n)$, i.e. the probability of M_i given the observations X_1, …,X_n. This is an example of inference to the best explanation.

Machine Learning and Inductive Inference

Machine learning is a broad term used to denote a variety of techniques whose goal is to gather information from data. The data consists of pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are feature vectors and $y_1, \ldots, y_n$ are the labels associated to $\mathbf{x}_1, \ldots, \mathbf{x}_n$, respectively. The feature vectors are drawn from some distribution which is typically unknown, so that the analysis is done in a non-parametric setting. Depending on the nature of the labels, two kinds of problems are identified. If the y_i's are elements of $\{0, 1\}$(or some finite non-empty set), then we have a classification problem, while if the y_i's are elements of R, then we have an estimation problem. Given the pairs (X_i, y_i), $i = 1, \ldots, n$, the goal is to 'learn' some rules so that given a new feature vector x, it is possible to provide the corresponding y. This problem is called supervised learning, since there is a learning phase where the given pairs are used to learn a rule. There is another problem called unsupervised learning which does not have a

learning phase. We do not consider unsupervised learning in this brief discussion. Statistics forms the basic theory for machine learning techniques. We refer to [8, 10, 13] for good introductions to the subject.

From the overview of machine learning stated above, the entire field essentially amounts to inductive inference. A particular machine learning method is a particular kind of inductive inference. Machine learning procedures measure performance by bounds on the error in estimation or classification. This often involves sophisticated mathematical machinery. Such mathematical rigour may lead one to believe that the problem of induction (i.e. providing justification for employing a particular induction procedure) discussed earlier has been addressed, if not fully at least partially. Such an assumption, however, would be incorrect. In the various machine learning procedures, aspects of inductive inference are implicitly used in such a manner that any justification of the learning procedure would amount to circular reasoning. Below we consider examples of implicit inductive inferences in two well-known machine learning procedures.

Neural Networks

Neural networks are complex models for learning from data. They come in various forms. Our brief consideration of neural networks is based on the treatment in [10].

Consider the problem of classification. A multilayer feed forward neural network for this problem can be visualised in the following manner. The network consists of computation units. The units are organised into layers. There is an input layer and an output layer. The output layer has a single unit. Each unit receives some inputs and provides a single output. The single unit in the output layer provides the output of the network which is a binary value. The units in the input layer receive the input to the network which is a feature vector. Connections between the units are as follows. Suppose there are $k + 1$ layers L_0,…,L_k, where L_0 is the input layer and L_k is the output layer. The output of any unit in layer L_i is provided as input to one or more units in layer L_{i+1}, for $i = 0, \ldots, k-1$. These connections are considered to be directed arcs from one unit to another. So, information flows from the input of the network to the output of the network, i.e. the network maps a feature vector to a binary value.

Weights are associated to each of the arcs in the network. Suppose u is a unit in a layer L_i other than the input layer, i.e. $i \geq 1$. Further, suppose that outputs of the units $u_1, \ldots, u_k$ of the previous layer are provided as input to u. So, there are k arcs connecting $u_1, \ldots, u_k$ to u. Let the weights associated to these arcs be $w_1, \ldots, w_k$. On a particular input x to the entire network, suppose that the outputs of u_1, u_k are $b_1, \ldots, b_k$, respectively. Then, the consolidated input to u is the weighted sum $a = w_1 b_1 + \ldots + w_k b_k$. The computation done by u is on the value a to produce an output b; if L_i is not the output layer, i.e., $i < k$, then the value b propagates to units in layer L_{i+1} to which u is connected.

The computation done by all the units in the network are the same. Each unit computes a sigmoidal function, i.e., a function from R to R which is bounded and

has non-negative derivative at all real values. Various sigmoidal functions are known and used in neural network computations. The output unit additionally applies a thresholding at the mid-point of the two bounds of the sigmoidal function to convert the real value to a binary value.

The network 'learns' by modifying the weights associated to the arcs. Suppose the set of arcs is A. Initially, the learning process starts out with a weight assignment $\left\{w_a^{(0)}\right\}_{a\in A}$. After the first pair (X_1, y_1) is processed, the weights are updated to $\left\{w_a^{(1)}\right\}_{a\in A}$; in general, after the ith pair (X_i, y_i) is processed, the weights are updated to $\left\{w_a^{(i)}\right\}_{a\in A}$, $i = 1, \ldots, n$. So, at the end of the learning procedure, the final weights are $\left\{w_a^{(n)}\right\}_{a\in A}$. At this point, the network is ready to perform classification of new inputs. On any new feature vector X, the corresponding classification value y is the output of the network when fed with X.

The goal of learning is to minimise misclassification error. This implies that learning is not perfect; i.e., from a finite number of samples, it is not possible to predict all future outputs in an error-free manner. Suppose the training data $(X_1, y_1), \ldots, (X_{i-1}, y_{i-1})$ have already been processed and the arc weights are $\left\{w_a^{(i-1)}\right\}_{a\in A}$. The next training data is (X_i, y_i). The vector X_i is provided as input to the network with arc weights $\left\{w_a^{(i-1)}\right\}_{a\in A}$, and the output $y_i^{'}$ of the network is computed. If $y_i = y_i^{'}$, then there is no error and the arc weights $\left\{w_a^{(i)}\right\}_{a\in A}$ are taken to be the arc weights $\left\{w_a^{(i-1)}\right\}_{a\in A}$. On the other hand, if $y_i \neq y_i^{'}$,, then the network has made an error. This necessitates updating the arc weights $\left\{w_a^{(i-1)}\right\}_{a\in A}$ to obtain the arc weights $\left\{w_a^{(i)}\right\}_{a\in A}$.

The goal of the updation procedure is to minimise the training error. Doing this in an absolute sense would require knowledge of the entire error surface. Since the error surface can be complex, it is not feasible to minimise over the entire surface. Instead, the updation procedure attempts to minimise the error using a procedure called gradient descent. This results in local minima which may be different from the global minima. The algorithm resulting from the application of gradient descent to update the arc weights is called the back propagation algorithm.

Let us now consider whether a neural network provides a formal justification of inductive inference. It indeed provides a formal description of a method for obtaining a general rule from available information. One may argue that since the prediction of a neural network is not guaranteed to be correct, this itself shows that it is not a reliable induction. A response to this argument would be that being able to predict an outcome with a guaranteed bound on error itself counts as knowledge; i.e., knowledge may be probabilistic in nature.

We do, however, think that the neural network methodology does not provide a justification for the method of induction. For the sake of concreteness, we focus on the multi-layer feed forward network with weights computed using the back propagation algorithm as outlined above. The back propagation algorithm is one particular embodiedment of the gradient descent methodology. So, one may question the justification for using gradient descent to minimise error. As mentioned above, this does not guarantee that the error obtained is the global minimum. It only guarantees that the error is locally minimum. So, even assuming knowledge to be probabilistic in

nature, how does one become sure that the correct probabilistic knowledge has been attained? The justification provided in [10] is that 'the algorithm has been found to work well in practice and is the most widely used training algorithm for multilayer networks'. So, the argument being made here is that the algorithm has worked well on previous occasions and hence is expected to work well in the future. This, of course, is an inductive inference. So, if we consider multilayer neural network as a method for induction, then to justify it we need to take recourse to induction, leading to circularity in argument.

Support Vector Machines

We consider the support vector machine (SVM) method for the classification problem. Our discussion is based on the description given in [10]. There are two key ideas in SVM. The first is that of using a linear separator with the maximum margin, and the second is that of mapping the feature space to a high-dimensional space and using a linear separator in the high-dimensional space.

Suppose the training data $(X_1, y_1), \ldots, (X_n, y_n)$ is linearly separable. If the data is not linearly separable, then the idea mentioned below is augmented using the notion of slack variables. Linear separability of the training data means that there is a hyperplane such that all training data having label 0 falls on one side of the hyperplane while all training data having label 1 falls on the other side. There is no unique hyperplane which separates the data. In fact, there will be inifinitely many hyperplanes any of which can act as a separator for the data. The first key idea of SVM is to choose a hyperplane having the maximum margin. Given a hyperplane h, let d_i be the minimum distance of h from all points labelled i, for $i = 0, 1$. Then, the margin of h is $d_0 + d_1$. The goal of SVM is to choose h such that $d_0 + d_1$ is maximised. The maximisation problem is formulated in terms of the training data and the solution to the maximisation problem constitutes the learning phase of the method.

The second key idea is to map to a high-dimensional space and apply the maximum margin separation in that space. This is done by applying a nonlinear function to the feature vector. The actual mappings that are used are in terms of the so-called kernel functions. The optimisation problem in the high-dimensional space can be expressed in terms of the kernel functions. There are several possibilities for choosing the kernel function. Each choice gives rise to one particular SVM method.

From the point of view of induction, one would look for justifications of two issues. First would be the rationale for obtaining a maximum margin linear separator. While the idea is intuitively appealing, it is not clear that such a separator is necessary for achieving minimum error. Like other inductive methods, justification for using a maximum margin separator is based on the idea being useful in previous cases and from that inferring that it is likely to be useful for future applications. This is again an inductive inference in itself. The other issue would be the choice of kernel function. Since there is no clear cut choice of a particular kernel, any justification for choosing

a kernel would be based on appeal to other successful applications sharing similar characteristics and inferring that it is likely to be useful for the application at hand. Once again, this is an inductive inference.

Concluding Remarks

In the preceding sections, we have tried to make explicit certain inductive inferences which are implicit in various statistical and machine learning methods. The problem of induction or, more specifically, the problem of justifying inductive inferences is a question of major philosophical interest. As mentioned earlier, it may appear that the formalism introduced by statistics and machine learning provides justification for inductive inference.

This idea has been explicitly mentioned on Page 7 by Kulkarni and Harman [10] where they comment: '… statistical learning theory provides partial deductive mathematical justifications for certain inductive methods, given certain assumptions'. This is a carefully crafted sentence which suggests that specific inductive inferences can be justified by statistical learning theory, but also adds the safeguard of 'given certain assumptions'. In the previous section, we have argued that neural networks and support vector machines are inductive inference mechanisms which cannot be justified without getting into a circular argument. If one were to question the 'given certain assumptions' clause in the assertion by Kulkarni and Harman, then we would be led to ask for justifications of such assumptions. It is very likely that such justifications will involve employing an inductive inference, so that we get back to a circular reasoning. Since, Kulkarni and Harman do not specify assumptions for particular learning methods, we are unable to pinpoint the circularity that would arise from such reasoning.

So, it is our case that looking at some specific examples of the use of inductive inference in statistics and machine learning, we are emboldened to state that almost all of statistics and machine learning essentially consist of inductive inference. A discerning reader may immediately note that in making such an assertion, we have ourselves made an inductive inference. So, whether our inference is justified is a specific case of the problem of induction. Consequently, this leaves open the possibility that there is indeed some statistical and/or machine learning method which justifies induction. At present, all we can say is that we are unaware of any such method.

We have repeatedly mentioned that trying to justify induction leads to a circular argument. In other words, if one tries to justify a particular inductive inference, then one is led to assuming that another inductive inference is valid, leading to a circularity of reasoning. Let us look a little more closely at this. Suppose, we are looking for a justification of a particular inductive inference, say I_1. In the process, suppose we are led to assuming the validity of a certain other inductive inference, say I_2. So, if

I_2 is valid, then so is $\boldsymbol{I_1}$. In other words, we have reduced the problem of justifying $\boldsymbol{I_1}$ to that of justifying I_2. If we denote the problem of 'justifying $\boldsymbol{I_1}$' by $\boldsymbol{P_1}$ and that of 'justifying I_2' by P_2, then we have been able to reduce problem $\boldsymbol{P_1}$ to problem $\boldsymbol{P_2}$.

The notion of reducing one problem to another is present in mathematics and many areas of computer science. Since computer science deals with problems of varying complexity, it is of interest to identify classes of problems of similar complexity and study separation between such complexity classes. The fundamental technique in such study is that of reduction. In fact, for each class, one tries to identify one problem (or a set of problems) which capture the complexity of the class in the sense that being able to solve this problem will lead to a solution to all problems in the class. Such a problem is called complete for the class.

With this background in mind, it may be of interest to form classes of inductive inferences, where all inductive inferences in a particular class have similar characteristics. The problem of justifying an inductive inference can then be reduced to one inference (or a set of inferences) in the class. The ultimate aim of such an exercise would be to try and identify one or more inductive inferences such that if it is possible to find justifications for such inductive inferences, then the entire problem of justifying inductive inference would be solved. This would not solve the problem of inductive per se, but it would focus attention on a few kinds of inferences which are of some fundamental nature. Hopefully, this would lead to a better understanding of the problem of induction.

References

1. Bandyopadhyay, P. S., & Forster, M. R. (eds.). (2011). *Philosophy of statistics*. Elsevier.
2. Chakrabarti, K. K. (2010). *Classical Indian philosophy of induction: The Nyaya viewpoint*. Lexington Books.
3. Chatterjee, K. S. (2012). *Statistical thought: A perspective and history*. Oxford University Press.
4. Douven, I. (2017). *Abduction*. Stanford Encyclopedia of Philosophy (Summer 2017 edition), Edward
5. Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A, CCXXII*, 309–368.
6. Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B (Methodological), 17*(1), 69–78.
7. Harman, G., & Kulkarni, S. (2007). *Reliable reasoning: Induction and statistical learning theory*. MIT Press.
8. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd Edn. Springer.
9. Henderson, L. *The problem of induction.* The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), archives/win2019/entries/induction-problem/.
10. Kulkarni, S., & Harman, G. (2011). *An elementary introduction to statistical learning theory*, Wiley.
11. Rao, C. R. (1997). *Statistics and truth: Putting chance to work*, 2nd Edn. World Scientific.
12. Sarkar, P. (2018). C$\bar{a}$rv$\bar{a}$kism Redivivus. *Newsletter of the American Philosophical Association on Asian and Asian-American Philosophers and Philosophies*, pages 26–31,

Fall, 2018. https://cdn.ymaws.com/www.apaonline.org/resource/collection/2EAF6689-4B0D-4CCB-9DC6-FB926D8FF530/AsianV18n1.pdf.

13. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.