The induction problem: a machine learning vindication argument

Gianluca Bontempi

Machine Learning Group
Computer Science Department, Faculty of Sciences
ULB, Université Libre de Bruxelles, Belgium
mlg.ulb.ac.be

Abstract. The problem of induction is a central problem in philosophy of science and concerns whether it is sound or not to extract laws from observational data. Nowadays, this issue is more relevant than ever given the pervasive and growing role of the data discovery process in all sciences. If on one hand induction is routinely employed by automatic machine learning techniques, on the other most of the philosophical work criticises induction as if an alternative could exist. But is there indeed a reliable alternative to induction? Is it possible to discover or predict something in a non inductive manner?

This paper formalises the question on the basis of statistical notions (bias, variance, mean squared error) borrowed from estimation theory and statistical machine learning. The result is a justification of induction as rational behaviour. In a decision-making process a behaviour is rational if it is based on making choices that result in the most optimal level of benefit or utility. If we measure utility in a prediction context in terms of expected accuracy, it follows that induction is the rational way of conduct.

1 Introduction

The process of extraction of scientific laws from observational data has interested the philosophy of science during the last two centuries. Though not definitely settled, the debate is more relevant than ever given the pervasive and growing role of data discovery in all sciences. The entire science, if not the entire human intellectual activity, is becoming data driven. Data science, or the procedure of extracting regularities from data on the basis of inductive machine learning procedures [3], is nowadays a key ingredient of the most successful research and applied technological enteprises.

This may appear paradoxical if we consider that the process of induction, from Hume's [4] work on, has been accused of having no rational foundation. Induction is the inferential process in which one takes the past as grounds for beliefs about the future, or the observed data as grounds for beliefs about the unobserved. In an inductive inference, where the premises are the data (or observations) and the conclusions are referred to as hypothesis (or models), three main properties hold [1]:

1. The conclusions follow *non-monotonically* from the premises. The addition of an extra premise (i.e. more data) might change the conclusion even when the extra premise does not contradict any of the other premises. In other terms $D_1 \subset D_2 \not\Rightarrow (h(D_2) \Rightarrow h(D_1))$ where $h(D)$ is the inductive consequence of D.
2. The truth of the premises is not enough to guarantee the truth of the conclusion as there is no correspondent to the notion of deductive validity.
3. There is an information gain in induction since a hypothesis asserts more than data alone.

In other words an inductive inference is *ampliative*, i.e. it has more content in the conclusion than in the premises, and contrasts with the mathematical and logical reasoning which is deductive and non ampliative. As a consequence, inductive inference is unsafe: no conclusion is a guaranteed truth and so it can dissolve even if no premise is removed.

According to Hume *all reasonings concerning nature are founded on experience, and all reasonings from experience are founded on the supposition that the course of nature will continue uniformly the same* or in other words that the future will be like the past. Any attempt to show, based on experience, that a regularity that has held in the past will or must continue to hold in the future will be circular. It follows that the entire knowledge discovery process from data is relying on shaky foundations. This is well known as the Hume's problem and the philosopher C. D. Broad's defined induction as "the glory of science and the scandal of philosophy". The serious consequences of such result was clear to Hume himself who never discouraged scientists from inductive practices. In fact, in absence of a justification, he provided an explanation which was more psychological than methodological. According to Hume, we, humans, expect the future will be like the past since this is part of our nature: we have inborn inductive habits (or instincts) but we cannot justify them. The *principle of uniformity of nature* is not a priori true, nor it can be proved empirically, and there is no reason beyond induction to justify inductive reasoning. Thus, Hume offers a naturalistic explanation of the psychological mechanism by which empirical predictions are made without any rational justification for this practice.

The discussion about the justification of induction started in the late 19th century. A detailed analysis of the responses to the Hume problem are contained in the Howson book [4]. Let us review some of the most interesting arguments.

First, the Darwinian argument which claims that the inductive habit was inherited as a product of evolutionary pressures. This explanation is a sort of "naturalized epistemology" but can be accused of circularity since we assume inductive science to justify induction in science.

Bertrand Russell suggests instead to accept the insolubility of the problem and proposes to create an "inductive principle" that should be accepted as sound ground for justifying induction.

Another well known argument is "Rule-circularity" which states: "Most of the inductive inferences humans made in the past were reliable. Therefore the majority of inductive inferences are right". Cleve insisted that the use of induction to prove induction is rule circular but not premise circular and, as such, acceptable. Criticisms about this argument, considered to be anyway question-begging, are detailed in [4].

”The No-Miracles” argument states: ”If an hypothesis h predicts independently a sufficiently large and precise body of data, it would be a miracle that it would be false then we can reasonably infer the approximate truth of the hypothesis”. Though this argument has been often raised in philosophy of science, common wisdom in machine learning may be used to refute it [3]. It is not the degree of fitting of an hypothesis to historical data which makes it true. There is nothing extraordinary (in fact no miracle at all) in predicting a large body of data if the hypothesis is complex enough and built ad hoc. A well know example is an overfitting hypothesis, i.e. a too complex hypothesis which interpolates past data yet without any generalization power. The quality of an hypothesis derives from the quality of the learning procedure used to build it and from the correctness of its implicit assumption, not from the fact of predicting correctly a large (or very large) number of outcomes. This is made explicit in statistical learning by the notions of bias and variance of an estimator [2] which we will use later to establish our argument.

Popper’s [9] reaction to Hume’s problem was to simply deny induction. According to him, humans or scientists do not make inductions, they make conjectures (or hypothesis) and test them (or their logical consequences obtained by deduction). If the test is successful the conjecture is corroborated but never verified or proven. Confirmation is a myth. No theory or belief about the world can be proven: it can be only submitted to test for falsification and, in the best case, be confirmed by the evidence for the time being. Though the argument of Popper got an immense credit among scientists, it is so strict to close any prospect of automatic induction. If induction does not exist and the hypothesis generation is exclusively a human creative process, any automatic and human independent inductive procedure (like the ones underlying all the successful applications in data science) should be nonsense or at least ineffective. As data scientists who are assisting to an incredible success of inductive practices in any sort of predictive task, we reject the denial of induction and we intend to use arguments from statistical machine learning to justify its use in science.

Let us assume that induction is a goal-directed activity, whose goal is to generalise from observational data. As Howson stresses [4] there are two ways to justify its usage: an internal and an external way. The internal way consists in showing that the structure of the procedures itself inevitably leads to the goal, in other words that the goal (here the hypothesis) necessarily follows from the input (the data). As stressed by Hume and confirmed by well-known results in machine learning (notably the ”no free lunch” theorems [14]), induction does not necessary entail truth. There is no gold procedure that given the data returns necessarily the optimal or the best generalization.

The second way is to show that the inductive procedure achieves the goal most of the time, but the success or not of the entreprise depends on factors external to it. There, though there is no necessity in the achievement of the objective, the aim is to show that the inductive procedure is the least incorrect one.

This paper takes this second approach by getting inspiration from the Hans Reichenbach's *vindication* of induction [10], where it is argued that if predictive success is possible, only inductive methods will be able to pursue it. *Consequently we have everything to gain and nothing to lose by employing inductive methods* [11]. In this paper we take a similar pragmatic approach to justify induction as the most rational practice in front of experimental observations. In detail, we extend and formalize the Reichenbach argument by adopting notation and results from the estimation theory in order to assess and measure the quality of an estimator, or more in a general, of any inductive algorithm. Note that in our approach we replace the long-run approach of Reichenbach, which was criticized in terms of limiting values, by a short run or finite sample approach which quantifies the generalization accuracy of any inductive practice using a finite number of observations. What emerges is that, though no guarantee of correct prediction is associated to inductive practices, induction is the most rational choice (in the sense of lowest generalization error) if we have to choose between an inductive or a non inductive approach.

Note that our decision theory arguments differs form the "rule-circularity" argument since we make no assumption that past successes of induction necessarily extrapolate to future ones. Each application domain is different and there is no guarantee that what worked in other contexts (or times) will be useful in our, too. In our argument induction is not perfect, but it is the lesser evil and it has to be preferred whatever is the degree of regularity of the natural phenomenon: in other words if a regularity exists, induction is less error prone than non induction while, in absence of regularity, induction is as weak as non induction.

Another aspect of our approach is that it applies whatever is the adopted inductive procedure (e.g. regression, classification, prediction). This allows us to extend the conventional discourse about induction to other domains than simply induction by enumeration. Finally this paper aims to corroborate the idea that the interaction between machine learning and philosophy of science can play a beneficial role in improving the grasp of the induction process (see [6], [12] and other papers in the same special issue).

2 The machine learning argument

We define first what we intend by inductive process and more specifically how we can assess in a quantitative manner its accuracy. In what follows we will represent variability by having recourse to a stochastic notation. We will use the bold notation to denote random variables and the normal font to denote their value. So by $\mathbf{x}$ we denote the random variable while by x we refer to a single realization of the r.v. $\mathbf{x}$. We will have resort to the terminology of the estimation theory [7] where an estimator is defined as any algorithm taking as input a finite dataset D_{tr} made of N_{tr} samples and returning as output an estimation (or prediction) $\hat{\theta}_N$ of an unknown quantity θ. Note that this definition is extremely general since it encompasses many statistical modeling procedures as well as plenty of supervised machine learning and data mining algorithms.

Suppose we are interested in some phenomena characterized by a set of variables for which we have a set of historical records. In particular our interest concerns some properties of these variables, like a parameter of the distribution, correlation or more in general their dependency. We use the target quantity θ to design the object of our estimation. For instance in regression θ could denote the expected value of $\mathbf{y}$ for an observed value x while, in case of binary classification, θ could design the probability that $\mathbf{y} = 1$ for a given x.

Since induction looks for regularity, but there is no logical necessity for regularity in nature, we should assess the benefits of induction (with respect to no induction) by taking into account that either regular or irregular phenomena can occur. We assume that the observed data are samples of a generative model characterized by a parametric probability distribution with parameter $\boldsymbol{\theta}$ where the parameter is random, and no assumption is made about the nature of the distribution. For the sake of simplicity, we will assume that $\boldsymbol{\theta}$ is scalar and with variance $\mathcal{V}$, though the results may be generalized to the multivariate case.

We distinguish between two settings: *regularity* (or uniformity of nature) and *irregularity.* A regular phenomenon is described by a distribution where the parameter θ_{tr} is unknown but constant. An irregular phenomenon is described by a distribution where the parameter is random. In this paper the degree of randomness of $\boldsymbol{\theta}$ is used to denote the degree of irregularity of the phenomenon.

In a regular setting two subsequent observations (or datasets) are i.i.d realizations of the same probability distribution, i.e. characterized by the same parameter θ_{tr}) In an irregular setting two subsequent observations are realizations of two different distributions, for example a training one with parameter θ_{tr} and a test one with parameter θ_{ts}.

Hence after, θ will denote both the parameter of the data distribution and the target of the prediction: for instance θ could be the conditional mean (in a regression task) or the a posteriori probability (in a classification task).

Machine learning decomposes the induction process into three major steps: the collection of a finite dataset D_{tr} made of N samples, the generation of a family $\mathcal{E}$ of estimators (or hypothesis) and the model selection[1]. We call estimator [7] any algorithm[2] taking as input a training set D_{tr} and returning as output an estimation (or prediction) $\hat{\theta}_{tr}$ of the target quantity θ_{tr}.

An estimator is then the main ingredient of any inductive process, since it makes explicit the mapping $\hat{\theta}_{tr} = \hat{\theta}(D_{tr})$ between data and estimation. Since data are variable (or more formally the dataset $\mathbf{D}_{tr}$ is random), the output of the estimation process is the random variable $\hat{\boldsymbol{\theta}}_{tr}$. It follows that we cannot talk about the accuracy of an inductive process without taking into account this variability. This is the reason why, if the goal is to predict θ we cannot assess the quality of an estimator (or more in general of any inductive procedure) in

[1] for simplicity we will not consider here the case of combining estimators

[2] here we will consider only deterministic algorithms

terms of a single prediction but more properly in terms of statistical average of the prediction error.

The accuracy of $\hat{\boldsymbol{\theta}}_{tr} \in \mathcal{E}$ may be expressed in terms of the Mean Squared Error

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{tr}) = E_{\hat{\boldsymbol{\theta}}_{tr}}[(\theta_{tr} - \hat{\boldsymbol{\theta}}_{tr})^2]$$

which can be notoriously decomposed in the sum

$$(E_{\hat{\boldsymbol{\theta}}_{tr}}[\hat{\boldsymbol{\theta}}_{tr}] - \theta_{tr})^2 + \mathrm{Var}\left[\hat{\boldsymbol{\theta}}_{tr}\right] = B_{tr}^2 + V_{tr}$$

where B_{tr} and V_{tr} denote the bias and the variance, respectively. While the bias is a sort of systematic error, typically due to the family of estimators taken into consideration, the variance measures the sensitivity of the estimator to the training dataset and decreases with the increase of the number N of observations.

What is remarkable here is that those quantities formalize the reliability and the accuracy of any inductive process. On one hand they show that it does not make sense to assume a perfect induction since any induction depends on data and, being data variable, induction is variable too. On the other hand, though perfect induction is illusory, it is possible to have degree of reliability according to the property of the target quantity, of the observed data and the estimator. In what follows we will use these quantities to show that no rational alternative to induction exists if our aim is to perform prediction or more in general extract information from data.

Let

$$\hat{\boldsymbol{\theta}}^* = \arg\min_{\hat{\theta}_{tr} \in \mathcal{E}} \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{tr}) \tag{1}$$

be the estimator in the family $\mathcal{E}$ with the lowest mean-squared-error. The aim of the selection step is to assess the MSE of each estimator in the family $\mathcal{E}$ and return the one with the lowest value, i.e.

$$\hat{\boldsymbol{\theta}}^*_{tr} = \arg\min_{\hat{\theta}_{tr} \in \mathcal{E}} \widehat{\mathrm{MSE}}(\hat{\boldsymbol{\theta}}_{tr}) \tag{2}$$

where $\widehat{\mathrm{MSE}}$ is the estimation of MSE returned by validation procedures (e.g. cross-validation, leave-one-out or bootstrap).

In order to compare inductive with non inductive practice we need to specify what we mean by alternative to induction. In qualitative terms, as illustrated by Salmon [11], *we might make wild guesses, consult a crystal gazer or believe what is found in Chinese fortune cookies.* In our estimation framework, a non inductive practice corresponds to a special kind of estimator which on purpose uses no data (i.e. $N = 0$), i.e. is data independent.

Let us now quantify the expected error of inductive and no-inductive procedures in the two contexts: regularity and no regularity.

Let us first consider the regular setting (denoted by the superscript $^{(r)}$) where the training and test datasets are i.i.d. samples of the $\boldsymbol{\theta} = \theta_{tr}$ distribution. The error of the non inductive process is

$$\mathrm{MSE}_0^{(r)} = E_{\hat{\boldsymbol{\theta}}_0}[(\theta - \hat{\boldsymbol{\theta}}_0)^2] = B_0^2 + V_0$$

where B_0 and V_0 denote the bias and the variance of the non inductive estimator $\hat{\boldsymbol{\theta}}_0$. According to (2) the error of the inductive process is $\mathrm{MSE}_I^{(r)} = \mathrm{MSE}(\hat{\boldsymbol{\theta}}^*_{tr})$.

Now, if we consider that a non inductive procedure is just a specific instance of estimator which is using no data, we can include $\hat{\boldsymbol{\theta}}_0$ in the family $\mathcal{E}$ by default[3]. It follows that the probability that MSE_I is bigger than MSE_0 amounts to the probability of wrong selection of $\hat{\boldsymbol{\theta}}^*$ due to the error in estimating the MSE terms . Since it is well known that this probability can be made arbitrarily small by increasing the number N of samples, we can conclude that the an inductive process cannot be outperformed by a non inductive one in the regular setting.

In the irregular setting the training set and the test set are generated by two different distributions with parameters θ_{tr} and θ_{ts}. In particular we assume that θ_{ts} is a realization of $\boldsymbol{\theta}_{ts}$ whose mean is θ_{tr} and whose variance is $\mathcal{V}$. The Mean Squared Error of the inductive process is now obtained by averaging over all possible training sets and test parameters:

$$\begin{aligned}\mathrm{MSE}_I^{(irr)} &= E_{\hat{\boldsymbol{\theta}}^*_{tr},\boldsymbol{\theta}_{ts}}[(\boldsymbol{\theta}_{ts} - \hat{\boldsymbol{\theta}}^*_{tr})^2] = \\ &= E_{\hat{\boldsymbol{\theta}}^*_{tr},\boldsymbol{\theta}_{ts}}[(\boldsymbol{\theta}_{ts} - \theta_{tr} + \theta_{tr} - \hat{\boldsymbol{\theta}}^*_{tr})^2] = \\ = E_{\boldsymbol{\theta}_{ts}}[(\boldsymbol{\theta}_{ts} - \theta_{tr})^2] &- 2E_{\hat{\boldsymbol{\theta}}^*_{tr},\boldsymbol{\theta}_{ts}}[(\boldsymbol{\theta}_{ts} - \theta_{tr})(\hat{\boldsymbol{\theta}}^*_{tr} - \theta_{tr})] + E_{\hat{\boldsymbol{\theta}}^*_{tr}}[(\theta_{tr} - \hat{\boldsymbol{\theta}}^*_{tr})^2] = \\ &= \mathcal{V} - 2\mathit{\Gamma} + E_{\hat{\boldsymbol{\theta}}^*_{tr}}[(\boldsymbol{\theta}_{tr} - \hat{\boldsymbol{\theta}}^*_{tr})^2] = \mathcal{V} - 2\mathit{\Gamma} + \mathrm{MSE}_I^{(r)} \qquad (3)\end{aligned}$$

In the equation above $\mathcal{V}$ quantifies the variability (or irregularity) of the phenomenon and

$$\mathit{\Gamma} = E_{\hat{\boldsymbol{\theta}}^*_{tr},\boldsymbol{\theta}_{ts}}[(\boldsymbol{\theta}_{ts} - \theta_{tr})(\hat{\boldsymbol{\theta}}^*_{tr} - \theta_{tr})] \qquad (4)$$

denotes the covariance between the estimator and $\boldsymbol{\theta}_{ts}$. This term can be different from zero only if the learning procedure incorporates some knowledge (also called inductive bias) about the $\boldsymbol{\theta}_{ts}$ distribution. If no knowledge about $\boldsymbol{\theta}_{ts}$ is available, $\mathit{\Gamma} = 0$.

Note that the equation (3) decomposes the testing error in an irregular setting into three terms: a term depending on the variability of phenomenon, a term representing the impact of inductive bias and a term denoting the MSE estimated on the basis of the training set only.

Analogously, for the non inductive case, we have

$$\mathrm{MSE}_0^{(irr)} = \mathcal{V} + \mathrm{MSE}_0^{(r)} \qquad (5)$$

Since model selection ensures $\mathrm{MSE}_I^{(r)} \leq \mathrm{MSE}_0^{(r)}$ for every θ_{tr}, from (3) and (5) we obtain that the error of the inductive process is not larger than the non inductive one. So in the irregular setting too, the inductive approach cannot be outperformed by the non inductive one. Though in the irregular case the resulting error is definitely much larger than in the regular case (notably if

[3] it is indeed a common practice to add random predictors in machine learning pipelines and to use them as null hypothesis against which the generalization power of more complex candidate algorithms is benchmarked

$\mathcal{V}$ is large), it still holds that an inductive approach is on average more accurate that a non inductive one. This can be explained by the fact that an irregular setting is analogous to an estimation task where a single observation (in this case θ_{tr}, or better its proxy $\hat{\theta}^*_{tr}$) is available about the target distribution (i.e. $\boldsymbol{\theta}$). Though a single sample is not much for returning an accurate estimation, it is nevertheless recommended to use it rather than discard it. A rational agent is therefore encouraged to choose, whatever is its assumption about the reality, an inductive technique to take into account the observed data.

MSE	Regular setting $\boldsymbol{\theta} = \theta_{tr}$	Irregular setting $\boldsymbol{\theta}$
inductive process	$\min_{\mathcal{E}}[B_{tr}^2 + V_{tr}]$	$\mathcal{V} + [\min_{\mathcal{E}}[B_{tr}^2 + V_{tr}]]$
non inductive process	$B_0^2 + V_0$	$\mathcal{V} + [B_0^2 + V_0]$

Table 1. MSE for induction and non inductive processes in the case of regular and irregular nature.

The reasoning is summarized in Table 1 which is on purpose remindful of Table I in [11]. We extended the "nothing to lose" vindicationist argument presented by Reichenbach by using the notion of mean-squared error to quantify the short-run prediction accuracy. Given the impossibility of assured success in obtaining knowledge from data, it is nevertheless possible to show quantitatively that inductive policies are preferable to any competitor. Induction is preferable to soothsaying since it will work if anything will.

2.1 Example

In order to illustrate the result (3) let us consider a simple task where the goal is to learn the expectation of a distribution. Let us suppose that we are in an irregular setting and that $\boldsymbol{\theta}_{ts}$ is distributed according to a Normal distribution with mean θ_{tr} and variance $\mathcal{V} = 1$. Suppose that we can observe a dataset D_N of size $N = 50$ distributed according to a Normal distribution with mean θ_{tr} and variance 1. Let us compare an inductive approach which simply computes a simple average ($\hat{\theta}^*_{tr} = \frac{\sum_{i=1}^N z_i}{N}$) with a number of noninductive strategies $\hat{\theta}_0$ which differ in term of inductive bias (or a priori knowledge), since $\hat{\theta}_0$ is distributed according to a Normal distribution with mean zero and standard deviation $\sigma_0 \in [0.05, 2]$. Figure illustrates $\text{MSE}_I^{(irr)}$ (upper dotted horizontal line) and $\text{MSE}_0^{(irr)}$ for different values of σ_0^2. It appears that for all values of σ_0^2 the inductive approach outperforms (i.e. lower generalization error) the non inductive one. As far as the a priori becomes less informative, the accuracy of the non inductive approach deteriorates due to the increasing of $\text{MSE}_0^{(irr)}$ (continuous increasing line).

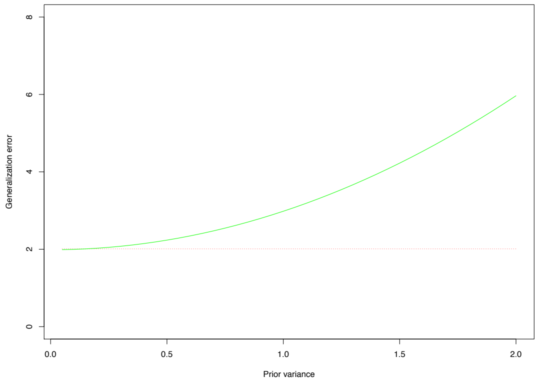

Fig. 1. Generalization error in the irregular setting for inductive and non inductive case: x-axis represents the amount of variance of the noninductive estimator. The horizontal line represents $\mathrm{MSE}_I^{(irr)}$ while the increasing line represent $\mathrm{MSE}_0^{(irr)}$.

3 Discussion

The results of the previous section leave open a set of issues which are worth to be discussed.

- Practical relevance: though the results discussed in the above section appear of no immediate application, there is a very hot domain in machine learning which could take advantage of such reasoning. This is the domain of transfer learning [8] whose aim is to transfer the learning of a *source* model θ_{tr} to a *target* model θ_{ts}, e.g. by using a small amount of target labeled data to leverage the source data. The transfer setting may be addressed in the derivation (3) by making the hypothesis that the distribution of θ_{ts} is no more centered on θ_{tr} but on a value $\mathcal{T} + \theta_{tr}$ where $\mathcal{T}$ denotes the transfer from the source to the target. In this case the result (3) becomes

  $$\mathrm{MSE}_I^{(irr)} = \mathcal{V} + \mathcal{T}^2 - 2\mathcal{T}B_{tr} - 2\Gamma + \mathrm{MSE}_I^{(r)}$$

  The difficulty of generalization is made then harder by the presence of additional terms depending on the transfer $\mathcal{T}$. At the same time this setting confirms the superiority of an inductive approach. The availability of a (however small) number of samples about θ_{ts} can be used to estimate the $\mathcal{T}$ term and then reduce the error of a data driven approach with respect to a non inductive practice which would have no manner of accounting for the transfer.
- Bayesianism: the Bayesian formalism has been more and more used in the last decades by philosophers of science to have a formal and quantitative interpretation of the induction process. If one hand, Bayesian reasoning

shed light on some famous riddles of induction (see [5]), on the other hand Wolpert [13] showed that conventional Bayesian formalism fails to distinguish target functions from hypothesis functions, and is then incapable of addressing the generalization issue. This limitation is also present in the conventional estimation formalism which implicitly makes the assumption of a constant invariable target behind the observations (like in the conventional definition of Mean Squared Error). In order to overcome this limitations Wolpert introduces the Extended Bayesian Formalism (EBF). Our work is inspired to Wolpert results and can be seen as a sort of Extended Frequentist Formalism (EFF) aiming to stress the impact on the generalization error of uncoupling the distribution of the estimator and the one of the target.

- The distinction between regular and irregular settings: for the sake of the presentation we put regular and irregular settings in two distinct classes. However as it appears from Table I, there is a continuum between the two situations. The larger is the randomness of the target (i.e. the larger $\mathcal{V}$) the least accurate is the accuracy of the inductive estimation.
- How to build an estimator: in the previous sections we introduced and discussed the properties of an estimator, intended as a mapping between a dataset and an estimation. An open issue remains however: how to build such mapping? Are there better ways to build this mapping? All depends on the (unknown) relation between the target and the dataset. Typically there are two approaches in statistics: a parametric approach where it is assumed the knowledge of the parametric link between the target and the dataset distribution and a nonparametric approach where no parametric assumption is made. However, it is worth to remark that though nonparametric approaches are distribution free, they are dependent on a set of hyperparameters (e.g. the kernel bandwidth or number of neighbors in nearest neighbour) whose optimal value, unknown a priori, depends on the data distribution. In other words any rule for creating estimators introduce a bias (in parametric or non parametric form) which has a major impact on the final accuracy. This aspect reconciles our vision with the denial of induction made by Popper. It is indeed the case that there is no *tabula rasa* way of making induction and that each induction procedure has its own bias [1]. Choosing a family of estimators is in some sense an arbitrary act of creation which can be loosely justified a priori and that can only be assessed by data. Machine learning techniques however found a way to escape this indeterminacy by implementing an automatic version of the hypotetico-deductionist approach where a number of alternative (nevertheless predefined) families of estimators are generated, then assessed and eventually ranked. This is accomplished by the model selection step whose goal is (i) to assess alternatives on the basis of data driven procedures (e.g. cross-validation or bootstrap), (ii) prune weak hypothesis and (iii) preserve the least erroneous ones.
- Why is inductive machine learning successful? the results of the previous section aim to show that the inductive process is not necessarily correct but necessarily better than non inductive alternatives. Its degree of success depends on the degree of regularity of the nature. Each time we are in front

of regular phenomena, or better we formulate prediction problems which have some degree of regularity, machine learning can be effective. At the same time, there are plenty of examples where machine learning and prediction have very low reliability or whose accuracy is slightly better than random guess: this is the case of long term prediction of chaotic phenomena (e.g. weather forecasting), nonstationary and nonlinear financial/economical problems as well as prediction tasks whose number of features is much larger than the number of samples (notably bioinformatics).

- Undetermination of theory by evidence: the bias/variance formulation of the induction problem supports the empiricist thesis of "undetermination of theory by evidence", i.e. the data do not by themselves choose among competing scientific theories. The bias/variance decomposition shows that it is possible to attain similar prediction accuracy with very different estimators: for instance a low bias large variance estimator could be as accurate as a large bias but low variance one. It is then possible that different machine learning algorithms generate alternative estimators which are compatible with the actual evidence yet for which prediction accuracy provides no empirical ground for the choice. In practice, other (non predictive) criteria can help in disambiguating the situation: think for instance to criteria related to the computational or storage cost as well as criteria related to the interpretability of the model (e.g. decision trees vs. neural networks).
- Practical relevance of those results: though the results discussed in the above section appear of no practical use, there is a very hot domain in machine learning which could take advantage of this reasoning. It is the domain of transfer learning where the issue is indeed to transfer the learning of a source model θ_{tr} to a target model θ_{ts}, for instance thanks to few samples characterizing the target tasks. This aspect could be taken into account in the derivation (3) by making the hypothesis that the distribution of θ_{ts} is not centered on θ_{tr} but on a value $\mathcal{T} + \theta_{tr}$ where $\mathcal{T}$ denotes the transfer from the source to the target. This setting it appears

4 Conclusion

Machine learning aims to extract predictive knowledge from data and as such it is intimately linked to the problem of induction. Its growing usage in theoretical and applied sciences raises additional pragmatic evidence in favour of induction. But as Hume stressed, past successes of machine learning are no guarantee for the future. Hume's arguments have been for more than 250 years treated as arguments for skepticism about empirical science. However, Hume himself considered that inductive arguments were reasonable. His attitude was that he had not yet found the right justification for induction, not that there was no justification for it. Reichenbach agreed with Hume that it is impossible to prove that induction will always yield true conclusions (validation). However, a pragmatic attitude is possible by showing that induction is well suited to the achievement of our aim (vindication).

This paper extends the ”nothing to lose” vindicationist argument presented by Reichenbach, by using the notion of mean-squared error to quantify the short-run prediction accuracy and by claiming the rationality of induction in a decision making perspective. Given the impossibility of assured success in obtaining knowledge from data, it is nevertheless possible to show quantitatively that inductive policies are preferable to any competitor. Induction is preferable to soothsaying since it will work if anything will.

Humans are confronted from the very first day to an incessant stream of data. They have only two options: make use of them or discard them. Neglecting data in no way can be better than using data. Inductive learning from data has therefore no logical guarantee of success but it is nonetheless the only rational behavior we can afford. Paraphrasing G. Box, induction is wrong but sometimes it is useful.

References

1. H.N. Bensusan. *Automatic bias learning: an inquiry into the inductive basis of induction.* PhD thesis, University of Sussex, 1999.
2. S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.
3. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
4. C. Howson. *Hume's Problem: Induction and the Justification of Belief.* Clarendon Press, 2003.
5. C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach.* Open Court, 2006.
6. K.B. Korb. Introduction: Machine learning as philosophy of science. *Minds and Machines*, 14(4):433–440, 2004.
7. E.L. Lehmann and G. Casella. *Theory of point estimation.* Springer, 1998.
8. S.J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.
9. K.R. Popper. *Conjectures and Refutationsf.* Basic Books, 1962.
10. H. Reichenbach. *The Theory of Probability.* University of California Press, 1949.
11. W. C. Salmon. Hans Reichenbach's vindication of induction. *Erkenntnis*, 27(35):99–122, 1991.
12. Jon Williamson. A dynamic interaction between machine learning and the philosophy of science. *Minds and Machines*, 14(4):539–549, 2004.
13. D. H. Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6:47–94, 1992.
14. D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8:1341–1390, 1996.