Machine learning, artificial neural networks and social research

Giovanni Di Franco & Michele Santurro

Quality & Quantity
International Journal of Methodology

ISSN 0033-5177

Qual Quant
DOI 10.1007/s11135-020-01037-y

Springer

Quality & Quantity
https://doi.org/10.1007/s11135-020-01037-y

Machine learning, artificial neural networks and social research

Giovanni Di Franco[1] · **Michele Santurro**[1]

Accepted: 29 August 2020

Abstract
Machine learning (ML), and particularly algorithms based on artificial neural networks (ANNs), constitute a field of research lying at the intersection of different disciplines such as mathematics, statistics, computer science and neuroscience. This approach is characterized by the use of algorithms to extract knowledge from large and heterogeneous data sets. In addition to offering a brief introduction to ANN algorithms-based ML, in this paper we will focus our attention on its possible applications in the social sciences and, in particular, on its potential in the data analysis procedures. In this regard, we will provide three examples of applications on sociological data to assess the impact of ML in the study of relationships between variables. Finally, we will compare the potential of ML with traditional data analysis models.

Keywords Machine learning · Deep learning Artificial neural network · Supervised learning · Linear models · Nonlinear models

1 Introduction

ML is an automatic learning process that takes place through the processing of usually very large data sets. The procedures of the past, defined with the "symbolic artificial intelligence" label, operated on algorithms constituted by a logical set of instructions by which a given output (usually called target) was encoded for all possible inputs. Contrarily, the new ML systems "learn" directly from data and estimate mathematical functions that discover representations of some input, or learn to link one or more inputs to one or more outputs to be able to formulate predictions on new data (Jordan and Mitchell 2015).

The paper is the result of the collaboration of the two authors. The drawing up of the text is attributed as follows: Sects. 3 and 4 to Giovanni Di Franco; Sects. 1, 2 and 5 to Michele Santurro.

✉ Giovanni Di Franco
giovanni.difranco@uniroma1.it

Michele Santurro
michele.santurro@uniroma1.it

[1] Department of Social and Economic Sciences, Sapienza University of Rome, Rome, Italy

In recent years in various human sciences: economics (Varian 2014; Blumenstock et al. 2015; Athey and Imbens 2017; Mullainathan and Spiess 2017), political science (Baldassarri and Goldberg 2014; Bonikowski and DiMaggio 2016), sociology (Barocas and Selbst 2016; Evans and Aceves 2016; Baldassarri and Abascal 2017), communication science (Hopkins and King 2010; Grimmer and Stewart 2013; Bail 2014), etc., ML has started to be applied both in academic research and in areas related to the management of services provided by the public administration (Athey 2017; Berk et al. 2018) or by private companies.

Overall, many different approaches and tools are included under the ML label (Kleinberg et al. 2015). Here we will only consider ANNs that use supervised ML algorithms. In the supervised ML the algorithm observes an output for each input. This output gives the algorithm a target to predict and acts as a "teacher". On the contrary, unsupervised ML algorithms only observe the input and their task is to independently compute a function without a predetermined target (Hastie et al. 2009; Molina and Garip 2019). The goal of this paper is to apply ANNs to sociological data by comparing the results obtained with the results of traditional statistical techniques, to evaluate their strengths and weaknesses.

2 Short illustration of artificial intelligence and machine learning based on artificial neural networks

Artificial intelligence (AI) is a branch of computer science that encompasses a huge variety of computational operations, ranging from classical algorithmic production to ML and deep learning (DL) techniques (Russell and Norvig 2010; Kitchin 2014b). The substantial difference between these approaches is that while traditional AI problem solving methods are based on if–then rules, ML and DL seek to iteratively evolve an understanding of data sets without the need to explicitly code any rules. This allow the computing system on which they are implemented to automatically learn and make predictions starting from a set of input data, adjusting their parameters by optimizing a performance criterion defined on the data and reducing the error rate at each stage of the learning process (Alpaydin 2016; Goodfellow et al. 2016).

In other words, in ML the aim is to construct a software program that adapt and learn independently, that is, without having a pre-programmed system that establishes how it should behave. Algorithms can learn from their mistakes thanks to training data used as examples. Accordingly, how much a model learns depends on the quality and amount of example data to which it has been exposed (Nilsson 2010; Dong 2017).

The considerable availability of information, due to the deluge of big data gathered from all kinds of specialized sensors and digital devices, and the rapid growth in parallel and distributed computing systems, made possible by the advance of faster CPUs, the advent of general purpose GPUs, the use of faster network connectivity and better software infrastructure for distributed computing, have given a boost to this sector (National Research Council 2013; Schmidhuber 2015; Goodfellow et al. 2016). AI applications are constantly evolving, reaching high levels of complexity and fascinating results in many different tasks: language translation, speech recognition, visual processing, spam filtering, and so on.

It is intuitive how companies capable of collecting and storing data correctly are candidates to be at the top of the AI sector. Many of the applications of DL are highly profitable

(Goodfellow et al. 2016; Zuboff 2019).[1] Indeed, despite the emphasis around the state of the art, most big tech companies still use traditional ML models instead of more advanced DL, and depend on a traditional infrastructure of tools poorly suited to ML (Dong 2017).

Early findings of DL date back at least to the 1960s, when it was intended to be a computational model of biological learning, that is, a model of how learning happens or could happen in the brain. As a result, one of the names that DL has gone by is ANNs. (Schmidhuber 2015; Goodfellow et al. 2016). The two terms are often used as synonyms. To be precise, DL is a subfield of ANNs, that uses multi-layered neural networks to process information. The idea behind deep neural networks is that, starting from the raw input, each hidden layer—so named because its values are not given in the data—combines the values in its preceding layer and learns more complicated functions of the input. It is difficult for a computer to understand the meaning of raw input data. DL resolves this difficulty breaking the desired task into a series of nested concepts, each described by a different layer of the model (LeCun et al. 2015; Alpaydin 2016; Goodfellow et al. 2016).

There is no consensus about how much depth a model requires to qualify as deep. Discussions with DL experts have not yet yielded a conclusive response to this question. However, DL can be safely understood as the set of models that involve a greater amount of composition of either learned functions or learned concepts than traditional ML does (Schmidhuber 2015; Goodfellow et al. 2016).

DL is not a breakthrough in the scientific sense, rather it is a relevant breakthrough in efficient coding that makes a difference in several contexts. In practical applications, DL is able to achieve higher accuracy on more complex tasks as compared with traditional ANNs, although it requires more computational resources. Furthermore, DL needs less manual interference to craft the right features or the suitable transformations of data. It performs exceptionally precise operations on data that come from different modalities, such as images, texts and videos (Schmidhuber 2015; Alpaydin 2016; Goodfellow et al. 2016).

In summary, ML offers numerous mathematical tools to deal with a wide variety of problems. The main tool, very popular nowadays, are the ANNs, which are trained to solve a particular task. Neurons are organized into groups called layers and connected to each other precisely to form a network. As mentioned, when the number of layers is high, the neural network is defined as deep. The DL's approach attempts to mathematically model the way in which the human brain processes information in vision and hearing: the stimuli of eyes and ears, passing through the human brain, are initially broken down into simple concepts and gradually reconstructed in increasingly complex and abstract representations (Russell and Norvig 2010; Alpaydin 2016; Goodfellow et al. 2016).

Similarly, in a deep network a face is broken down in the form of an array of pixel values. The first layer can easily identify edges of different orientations. Subsequent layers combine these to form corners and extended contours. Layers that follow can detect entire parts of specific objects, by finding specific collections of contours and corners. Finally, these in turn are combined with some more layers of processing, allowing us to represent the faces we want to learn (Nilsson 2010; LeCun et al. 2015; Alpaydin 2016; Goodfellow et al. 2016).

[1] We must not forget that drivers and aspirations of corporations are quite different from the aims of academic researchers: we distinguish the former as motivated by financial concerns (e.g. prediction and control for product improvement and to identify new markets and opportunities), whereas the latter is focused on—at least should—the search for understanding and explanation of phenomena and processes (Crawford 2013; Kitchin 2014a; Lagoze 2014; Törnberg and Törnberg 2018).

So, the choice between ML or DL algorithms depends on the problem to be analyzed. If the problem is relatively simple, it is preferable to use ML based on ANNs with few layers of hidden units; if the problem is complex or requires the achievement of very specific and rigorous objectives, it is considered more useful to resort to DL.

3 Methodology

The starting point of our experiments is to evaluate whether, in the typical data analysis operations of social sciences, the techniques of ML based on ANNs can constitute an alternative, or at least a possible integration, with respect to the traditional data analysis tools which basically consist of linear and logistic regression models.

As is known, in general, multivariate data analysis models perform an empirical control of one or more hypotheses derived from a theory and the results consist in the comparison between the so-called expected, or theoretical, data and the empirical data. If the outcome of this comparison is attributable to random effects, it is said that the model fits, or is compatible, with the data; otherwise the model must be revised or, if this is impossible, rejected (Di Franco 2017). It is therefore the so-called confirmatory-explanatory approach.

Starting from the work of data analysis pioneers such as Fisher (1925, 1935), Galton (1869, 1886), Spearman (1904, 1927) and many others, for many decades data analysis in the social sciences has been characterized by this approach which fundamentally seeks to identify, from associations between a set of empirically detected variables, causal links between the same variables. In this context the model (i.e. the theory) is prevalent and the data are used to evaluate the goodness of fit of the model, expressed by the values of a coefficient of statistical significance (p-value).

Alternative approaches to data analysis, based on induction, exploration-description, simulation, etc. which have also been proposed over time (among others by Benzécri 1969, 1992; Benzécri et al. 1973a, b; Tukey 1977; Gifi 1981, 1990) have received less interest among social sciences researchers. The characteristic of these alternative approaches is the inversion of the relationship between data and theory: data are more important than the model. This means that starting from the data it is necessary to find the model that best represent them; while in the causal approach the starting point is a model and the data are used to test it.

Thanks to the recent developments in different disciplines such as applied mathematics, statistics, information technology, approaches based on data prevalence have become established, or are emerging, in many disciplinary areas of natural and biometric sciences. Over time these approaches have taken on different names such as data mining, statistical learning, machine learning (ML), deep learning (DL) and others.

In addition to the innovations to which we referred, starting from the development in information and communication technologies and web platforms, the current historical period is strongly characterized by the so-called big data and their management through mathematical algorithms that are able to independently process them to extract information useful for various purposes. As a result, many ML techniques exist today. A common feature of these techniques is that they are exploratory and rely on computer assisted analysis.

One large subdivision of these techniques uses a single outcome and tries to make an optimal prediction of this outcome from multiple predictor variables (supervised learning techniques). The second subdivision does not require any outcome and merely classifies

inputs into subgroups based on similarities among a set of variables (unsupervised learning techniques).

For the purpose of our experiments we will use the ML which adopts the ANNs which have units arranged on three layers (input, hidden and output) and unidirectional connections between each unit of one layer and all the other units of the next layer.

Being essentially a distributed processor built in analogy with the human central nervous system, an ANN is generally composed of elementary computational units called neurons, conceivable as nodes of a network with certain processing capacities and interconnected.[2] Artificial neurons are able to receive a combination of signals from the outside or from other neurons, and then transform them through a particular function called activation function, thus storing data in the network parameters and in particular in the weights associated with every connection.

Then there is the return of an output: a result generally dependent on the purpose for which the ANN was built (classification, recognition, approximation, etc.).

The relationship between incoming and outgoing data is generally determined:

- From the type of elementary units used: complexity of the internal structure, class of activation function used;
- From the formal structure of the network: number, orientation and direction of the nodes, which can be represented according to the tool of graph theory;
- From the values of the internal parameters associated with the neurons and the related interconnections: to be determined using appropriate learning algorithms.

The question we ask ourselves is whether ANN can be usefully applied in social research, besides as a complex of nonlinear data processing algorithms, also as a tool to simulate social phenomena (Capecchi 1996).

It is difficult to assimilate social phenomena to neurophysiological ones; for this reason, the analogies of the nodes of an ANN with neurons, of its connections with synapses, etc., that are possible in the study of the brain, are not possible in these other cases. However, it is a question of assessing whether the abstractness of the structures and processes postulated in ANNs, understood as models of complex nonlinear dynamic systems, does allow their application also to the study of social phenomena. In this case it is necessary to determine the interpretation to be given to concepts such as node, connection, excitation/inhibition, connection weight, learning rule, equilibrium and so on.

On the other hand, the use of ANNs allows the possibility of partially overcoming some limitations of the analyses conducted with traditional statistical techniques. For example, the use of ANNs does not require any hypothesis on the distributions of the system variables and their reciprocal associations. For this reason, the treatment of cardinal, ordinal and/or categorical variables is possible (Di Franco 2017). By such approach the actual analysis of the system is left to the network, which alone creates its own criteria to reproduce its behaviour and consequently enables itself to formulate predictions on the system itself. In Fabbri and Orsini's (1993) judgement, this is both a strength and a weakness of ANNs: it is a strength because in this way the researcher

[2] Neurons are typically arranged along horizontal lines called layers, they communicate with neurons in the lower and upper layers by transforming the signals from layer to layer non-linearly. Their weights are iteratively modified thanks to certain ML algorithms – one of the best known and most useful is that of the stochastic gradient descent.

is not conditioned by a priori hypotheses in the choice of the units of the network; the weakness consists in the fact that the network is not able to do anything else but reproduce in a phenomenological manner the behaviour of the analysed system, without contributing to the knowledge of the internal relationships between the single parts of the system. This problem, however, can be partially overcome as some devices, that allow us to interrogate the network about what it was able to reproduce, have been fine-tuned. (Di Franco 1998).

If the simulation approach of ANNs to social phenomena proved to be possible and useful (Capecchi et al. 2010), this would allow significant progress in the social disciplines because it would also contribute to the foundation of a consistent basis of simulation concepts, models and techniques. If social phenomena can be thought of as complex dynamic systems then it is necessary to accept the possibility of simulating them on a computer with more meaningful results than those obtainable with traditional data analysis tools.

We now describe the methodology used in the examples whose results we present in the next paragraph. The data used in the three examples are taken from a matrix containing some information on the electoral polls published in Italy by the mass media from 1 January 2017 to 29 February 2020. The information relating to these electoral polls was downloaded from the institutional website of the Presidency of the Council of Ministers: www.sondaggipoliticoelettorali.it.

In the period indicated above we collected 825 polls focused on voting intentions for the next political elections. As mentioned, the results of these polls have been disseminated by the mass media and are governed by rules that require the drafting of an information note that presents methodological information useful for assessing the correctness of the polls carried out by the various agencies (Di Franco 2018).

The Italian regulation on the publication and dissemination of electoral polls in the mass media lists the information that must compulsorily be inserted in the document that is published on the institutional website. These are the fifteen information items:

1. Title of the poll;
2. Subject who carried out the poll;
3. Client;
4. Buyer;
5. Date or period in which the poll was carried out;
6. Name of the mass media in which the poll is published or disseminated;
7. Date of publication or diffusion;
8. Topics covered by the poll;
9. Reference population;
10. Territorial extension of the poll;
11. Sampling method;
12. Representativeness of the sample including indication of sampling error;
13. Method of collecting information;
14. Sample size, number and percentage of non-respondents and replacement made;
15. Full text of all questions and percentage of people who answered each.

From our analysis it emerged that in many documents there are important gaps with respect to what is required by current legislation, especially in relation to purely methodological information.

To assess the quality of the documents as a whole, we have developed a completeness index of the poll information, adding the presence of the following six elements on which we have identified the most critical issues:

1. The proportions between the breakdown of interviews conducted with mixed interview methods;
2. The confidence interval for the estimates;
3. The number of subjects contacted;
4. The number of refusals and replacements for the interviews carried out;
5. The percentage of the undecideds;
6. The percentage of voters or abstainers.

We have coded each of the six elements with the value one when it is present and with the value zero when it is absent. We then normalized the values of the index by dividing the sum by the number of elements in order to obtain values in the range between zero (which involves the absence of all six elements) and one (when all six elements are present).

The minimum value found on the completeness index (label ind-1) is .17 (which means only one element out of six); the maximum 1 is recorded only in 9 polls out of 825. The average score is .56, the standard deviation is .18. Most of the polls analyzed (58.5%) obtain a value on ind-1 equal or less than .5.

By analyzing the mean values of ind-1 for the interview techniques we can see in which cases the most critical issues are recorded.

Let's first analyze the polls conducted using a single data collection technique. When the interviews are carried out using a panel of respondents the average on ind-1 is equal to .69; the polls carried out with the CATI (computer assisted telephone interviewing) technique present an average on ind-1 equal to .63; the polls carried out with the CAWI (computer assisted web interviewing) technique present an average on ind-1 equal to .33.

When the polls are carried out using mixed techniques of data collection the average results on ind-1 are: .65 with CATI-CAWI; .60 with CATI-CAMI (computer assisted mobile interviewing); .49 with CATI-CAMI-CAWI.

We have computed the eta coefficient to quantify the strength of the association between ind-1 and the technique of conducting interviews. The value obtained (.641) allows us to establish the existence of a significant association between the two variables examined.

By examining the information notes of each poll, we considered all the other information provided with respect to the current legislation, focusing our attention on the elements that concern very important aspects for the evaluation of the results of the poll such as the days in which it was carried out, the sample size, the sampling error, the confidence interval, the number and percentage of those not available, non-respondents and replacements made, the full text of all questions and the percentage of interviewees who they answer each of them.

Following the descriptive analysis, we found a second serious gap in the methodological notes concerning the information on the number of contacts and that of refusals : 128 out of 825 published surveys (equal to 15.5%) do not report this information. To take this important information into account, we have designed a second index (ind-2) which consists in the relationship between the number of people contacted and the number of interviews carried out. Thanks to this index, we can evaluate for each poll how many subjects it was necessary to contact to carry out a valid interview. The mean value of ind-2 is equal to 5.313 (the standard deviation is 3.768) which indicates that to carry

Table 1 Descriptive statistics of the main variables available in the data matrix

	N	Min	Max	Mean	Standard deviation
Days	825	1	25	2.62	1.966
N-sample	825	500	16,000	1243.62	779.537
Error	825	1.30	4.40	3.033	.5647
N. contacts	697	1788	21,567	5090.62	2706.72
N. of refusals	697	41	20,321	3944.53	2827.57
ind-1	825	.17	1.00	.5626	.18238
ind-2	697	1.02	18.06	5.313	3.768
No-vot	591	3.0	66.0	39.008	12.983

out a valid interview it was necessary to contact just over five subjects. In other words: on average, for each interview carried out more than four refusals were registered.

By analyzing the mean values of ind-2 for the interview techniques we can see in which cases the most critical issues are recorded.

Let's first analyze the polls conducted using a single data collection technique. When the interviews are carried out using a panel of respondents the average on ind-2 is equal to 1.178; the polls carried out with the CATI present an average on ind-2 equal to 5.23; almost all the polls carried out with the CAWI technique do not provide this information. It seems clear that the problem particularly concerns the use of the CAWI data collection technique and this suggests that in fact the institutes that use this technique do not carry out a probability sampling but draw on a form of convenience selection if not a self-selection of the subjects who frequent the web.

When the polls are carried out using mixed techniques of data collection the average results on ind-2 are: 9.494 with CATI-CAWI; 5.119 with CATI-CAMI; 5.690 with CATI-CAMI-CAWI.

We have computed the eta coefficient to quantify the strength of the association between ind-2 and the technique of conducting interviews. The value obtained (.647) allows us to establish the existence of a significant association between the two variables examined.

By examining the differences between the ind-2 values, it is possible to find a significant effect of the data collection technique on the ratio between the number of contacts and the number of interviews carried out. Undoubtedly, the polls that resort to web and mobile interviews have significantly higher ratios than those conducted only with CATI and those that resort to the combination of CATI and CAWI. Polls conducted on a panel are an exception because the sample is composed of subjects who agree to be interviewed repeatedly over time, therefore they have very low values on ind-2.

In Table 1 we show the descriptive statistics relating to the following variables: duration of the survey in number of days (label days), sample size (nsample), sampling error (error), number of subjects contacted (n. contacts), number of subjects who refused the interview (n. of refusals), value recorded on ind-1 (ind-1), value recorded on ind-2 (ind-2), percentage of respondents who declared their intention not to vote or who declared themselves undecided (no-vot).

On average, the polls analyzed were carried out in just over two days (2.6 days; 1966 the standard deviation; 1 the minimum value; 25 the maximum value).

The sample sizes vary in a range from 500 to 16,000 cases; the average is 1243.62, the standard deviation is 779.537.

Table 2 Multiple regression model summary

R	R square	Adjusted R square	Std. error of the Estimate
.563	.317	.311	84,224

Predictors: (Constant), days, n-sample, ind-1, ind-2. Dependent Variable: no-vot

Table 3 Multiple regression coefficients

	Unstandardized coefficients		Standardized coefficients		
	B	Std. error	Beta	t	Sig
(Constant)	21.168	2.609		8.114	.000
Days	−.663	.297	−.105	−2.229	.026
n-sample	.007	.001	.284	4.986	.000
ind-1	25.129	2.872	.348	8.749	.000
ind-2	−.705	.156	−.226	−4.521	.000

Dependent Variable: no-vot

Linked to the size of the sample is the level of sampling error that in the analyzed polls varies between 1.3 and 4.4%. The average error is 3%.

Finally, with regard to the request to provide information on the number and percentage of subjects who do not answer the poll questions in our analysis—since we have considered only the question relating to voting intentions, whose formulation is: "if you voted today [or, if you had voted yesterday] for the Chamber of Deputies, which party would you vote for [or, would you have voted]?"—we have taken into consideration the presence of the percentages of the undecideds and those who intend to abstain from voting.

In 28.36% (234 cases) of the polls, neither the percentage of undecideds nor that of abstainers was reported.

4 Results and discussion

The first example consists of a comparison between a multiple linear regression model and an ANN Multilayer Perceptron.

We first present the results of multiple linear regression. The dependent variable is the percentage of voters who declared their intention to abstain or who declared their indecision regarding the election choice (label 'no-vot'). The independent variables are the following four: the duration of the poll in days (label 'days'); the sample size (label 'n-sample'); the completeness index of the information relating to the poll (label 'ind-1′); the ratio between the interview attempts and the interviews carried out ('ind-2′).

Table 2 presents the fitting results of the multiple regression model. Considering the adjusted R square, we find that the four independent variables reproduce a little less than a third (31.1%) of the variance of the dependent variable. Table 3 shows the regression coefficients and Table 4 the residual statistics.

The analysis of the beta weights confirms that the contribution of the four independent variables is significant in explaining the variance of the dependent one. Of the four

Table 4 Multiple regression residual statistics

	Min	Max	Mean	Std. deviation	N
Predicted value	30.684	56.227	41.261	5.7111	506
Residual	−26.1793	35.3160	.0000	8.3890	506
Std. Predicted value	−1.852	2.621	.000	1.000	506
Std. residual	−3.108	4.193	.000	.996	506

Dependent Variable: no-vot

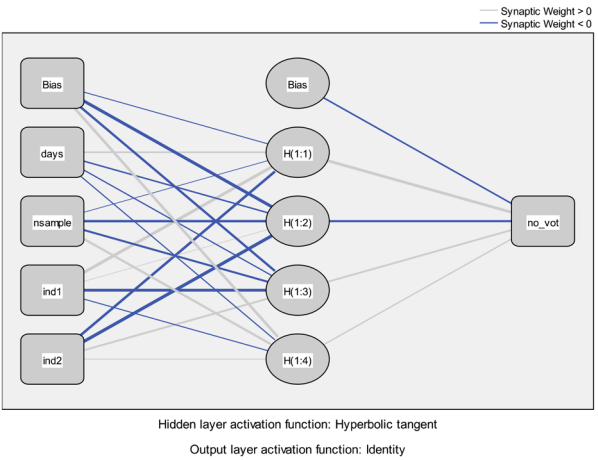

Fig. 1 The architecture of the ANN

Table 5 ANN model summary

Training	Sum of squares error	40.218
	Relative error	.225
	Stopping rule used	1 consecutive step(s) with no decrease in error
	Training time	0:00:00.194
Testing	Sum of squares error	19.528
	Relative error	.327

independent variables ind-1 (.348) and n-sample (.284) have a positive beta weight; ind-2 (−.226) and days (−.105) have a negative beta weight.

In other words, the percentage of non-voters (dependent variable) is directly proportional to the completeness of the poll information and the poll sample size and is inversely proportional to the ratio between attempts and interviews carried out and the increase in the duration in days of the poll.

The analysis of the residual statistics also shows a good fit of the model to the data (Table 4).

Table 6 Correlations between predicted values of regression and ANN and values of dependent variable

No-vot	1		
Unstandardized predicted value: regression	.563**	1	–
Predicted value for no_vot: ann	.866**	.391**	1

**Correlation is significant at the .01 level (2-tailed)

Table 7 Binary logistic regression coefficients

	B	S.E	Wald	df	Sig	Exp(B)
days	1.582	.318	24.719	1	.000	4.863
nsample	.004	.002	8.808	1	.003	1.005
err	5.077	1.134	20.052	1	.000	160.220
ind-1	− 14.427	2.055	49.276	1	.000	.000
no_vot	− .109	.026	17.996	1	.000	.897
ind-2	.301	.128	5.511	1	.019	1.351
Constant	− 9.075	5.048	3.232	1	.072	.000

Let's now evaluate the results obtained with the ANN comparing them with those obtained with the multiple linear regression (Fig. 1).[3]

The cases submitted to the network are obviously the same 506 used in the regression. In this case, however, 70% of cases (359) were used in the training set and the remaining 30% (147) in the testing set. Table 5 presents the model summary. In the training set the relative error was equal to .225. In the testing set it grows slightly reaching the value of .327. Recall that in the testing set the network predicts the value of the dependent variable using the weights that it computed on the cases observed during the training. So basically, we assess the ability of the network to generalize what it has learned in the training.

We do not report the parameter estimates (i.e. the weights calculated for each node of the network) as their examination does not clarify the impact of each independent variable in the estimate of the dependent one.

The comparison between the results of the multiple regression and the ANN leaves no doubt about the better predictive performance of the network (Table 6). The correlation between the values predicted by the multiple regression and the actual values of the dependent variable is equal to .563; the correlation between the values predicted by the ANN and the actual values of the dependent variable is thirty points higher, rising to .866.

Evidently in the relationship between the independent variables and the dependent one, the network managed to capture nonlinear trends which allow for a better estimate of the values.

In the second example we compare a binary logistic regression model with a network. The data refer to the same matrix used in the previous example. The dependent variable, in this case, is a dichotomy and represents the data collection method used in the poll (label 'met'). The first category represents polls that use only one technique (CATI or CAWI

[3] For the ANN applications we used the Multilayer Perceptron procedure available in SPSS for Windows.

Table 8 Binary logistic regression model summary

−2 Log likelihood	Cox & Snell R square	Nagelkerke R square
197.373	.527	.772

Table 9 Binary logistic regression classification table

Observed	Predicted		(%) Correct
	Single	Multi	
Single	120	9	93.0
Multi	14	349	96.1
Overall (%)	–	–	95.3

[a]The cut value is .500

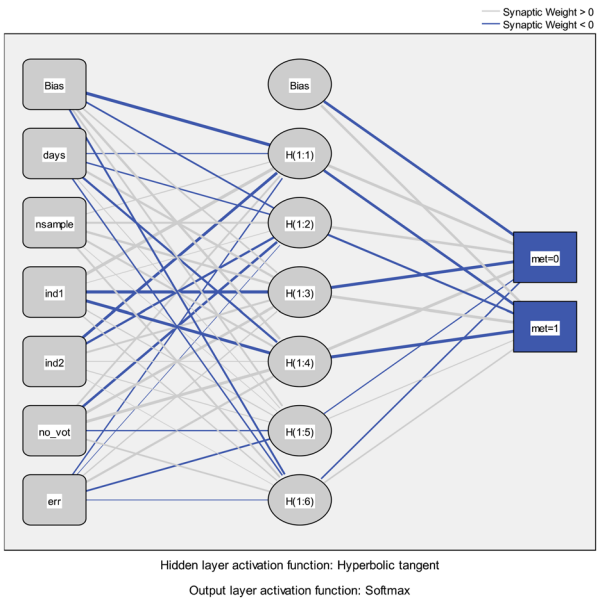

Fig. 2 The architecture of the ANN

or panel); the second category represents polls made with more than one technique (e.g. CATI and CAWI; CATI, CAMI and CAWI, etc.). The independent variables are the same four used in the first example plus the sampling error (label 'err') and the percentage of non-voters (label 'no-vot').

Table 7 reports the coefficients of the logistic regression. Also in this case the coefficients of all six independent variables are significant.

Table 8 presents the fitting results of the binary logistic regression model which express the statistical significance of the model in the estimate of the dependent variable.

Examining the values of the pseudo R-square (respectively: .527 Cox & Snell and .772 Nagelkerke), they show the good fit of the model to reproduce the dependent variable.

Table 9 presents the classification table which in a simple way allows us to evaluate the predictive power of the logistic regression model. Overall, the model registers 95.3% of

Table 10 ANN model summary

Training	Cross entropy error	9.397
	Percent incorrect predictions	.6%
	Stopping rule used	1 consecutive step(s) with no decrease in error
	Training time	0:00:00.178
Testing	Cross entropy error	10.040
	Percent incorrect predictions	1.9%

Table 11 ANN classification table

Sample	Observed	Predicted		
		Single	Multi	(%) Correct
Training	single	87	2	97.8
	multi	0	245	100.0
	Overall (%)	26.0	74.0	99.4
Testing	single	39	1	97.5
	multi	2	116	98.3
	Overall (%)	25.9	74.1	98.1

Table 12 Multinomial logistic regression model fitting information

Model	Model fitting criteria −2 Log likelihood	Likelihood ratio tests Chi-square	df	Sig
Intercept Only	1.456E3	–	–	–
Final	269.947	1.186E3	30	.000

correct classifications. Considering the two categories of the dependent variable, the first registers 93% of correct classifications and the second 96.1%.

We now consider the results obtained with the ANN by comparing them with those obtained with the binary logistic regression (Fig. 2).

The network architecture in this example consists of an input layer of six nodes (one for each independent variable), a hidden layer of six nodes and an output layer of two nodes (one for each of the two categories of the dependent variable).

The comparison between the results of the binary logistic regression model and the ANN leaves no doubt about the better predictive performance of the network (Table 10). The network correctly classifies 99.4% of the polls used in the training set and 98.1% of the polls used in the testing set.

Examining in detail the result of the network (Table 11), 97.8% of the first category and 100% of the second one of the dependent variable are correctly classified in the training set; in the test set 97.5% of the first category and 98.3% of the second category are correctly classified.

In the third example we compare a multinomial logistic regression model with an ANN. The dependent variable this time is a polytomous variable with five categories that

Table 13 Multinomial logistic regression model goodness of fit statistics

	Chi-Square	df	Sig
Pearson	850.578	2495	1.000
Deviance	269.947	2495	1.000

Table 14 Multinomial logistic regression model pseudo R-square statistics

Cox and Snell	.904
Nagelkerke	.958
McFadden	.815

Table 15 Multinomial logistic regression classification table

Observed	Predicted					
	cati	cati-cami	cati-cami-cawi	cati-cawi	Panel	(%) Corr.
cati	40	6	0	0	0	87.0
cati-cami	3	45	8	0	0	80.4
cati-cami-cawi	0	2	219	7	0	96.1
cati-cawi	0	1	12	68	0	84.0
Panel	0	0	0	0	94	100.0
Overall %	8.5	10.7	47.2	14.8	18.8	92.3

represents the technique of conducting the interviews used in the polls. The categories are: 1 = CATI, 2 = CATI-CAMI, 3 = CATI-CAMI-CAWI, 4 = CATI-CAWI, 5 = panel. The independent variables are the same as in the previous example.

Tables 12 and 13 present the goodness of fit statistics of the multinomial logistic regression model; Table 14 shows the values of the pseudo R^2, which in this case are very good, demonstrating the excellent fit of the model to the data.

As is known (Di Franco 2017), the model produces a different parameter of each independent variable for each category of the dependent one, except the last category which is set as the reference category (in our example it is the 'panel' category). We omit to report the parameter estimates.

In Table 15 we report the classification table of the multinomial logistic regression model. Through the reading of the results, we can see how overall the model reproduces correctly 92.3% of the cases. If we consider the single categories of the dependent variable, the best performance is obtained with the panel category (100% of correct classifications) and with the CATI-CAMI-CAWI category (96.1% of correct classifications). For the other three categories of the dependent variable (CATI, CATI-CAMI and CATI-CAWI) the percentage of correct classifications varies from 80.4 to 87%.

Let's now see the results of ANN, obviously applied to the same data and to the same variables used for the multinomial logistic regression model. Figure 3 show the architecture of the network which is composed of six input nodes, four hidden nodes on a single layer and five output nodes.

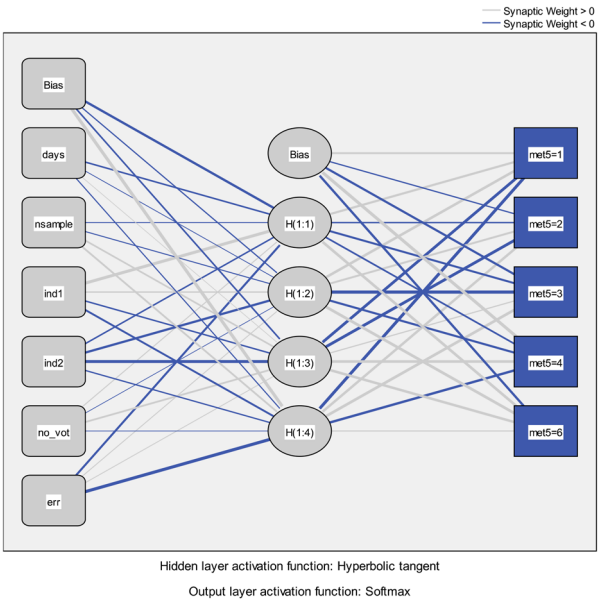

Fig. 3 The architecture of the ANN

Table 16 ANN model summary

Training	Cross entropy error	61.649
	Percent incorrect predictions	5.5%
	Stopping rule used	1 consecutive step(s) with no decrease in error
	Training time	0:00:00.296
Testing	Cross entropy error	34.673
	Percent incorrect predictions	6.8%

Table 17 ANN classification table

Sample	Observed	Predicted					
		cati	cati-cami	cati-cami-cawi	cati-cawi	Panel	(%) Corr.
Train.	Cati	31	2	0	1	0	91.2
	cati-cami	0	29	4	0	0	87.9
	cati-cami-cawi	0	0	155	5	0	96.9
	cati-cawi	0	0	7	45	0	86.5
	Panel	0	0	0	0	64	100.0
	Overall %	9.0	9.0	48.4	14.9	18.7	94.5
Testing	cati	11	1	0	0	0	91.7
	cati-cami	0	19	3	1	0	82.6
	cati-cami-cawi	0	0	64	4	0	94.1
	cati-cawi	0	1	1	27	0	93.1
	Panel	0	0	0	0	30	100,0
	Overall %	6.8	13.0	42.0	19.8	18.5	93,2

Table 16 shows the ANN model summary. As usual, the data were divided into training set (70%) and testing set (30%).

Finally, Table 17 reports the classification table obtained with ANN in the training set and in the testing set.

Also in the third example the ANN results are better than those of the multinomial logistic regression model. On the whole, if we consider the results of the training, ANN reaches 94.5% of correct classifications against 92.3% of the multinomial model. Even if we take into consideration the results of ANN testing, they are, albeit slightly, better (93.2%). As for the single categories of the dependent variable, ANN achieves the best performance with the panel category (100% of correct classifications for both training and testing set) and with the CATI-CAMI-CAWI category (96.9% for the training set and 94.1% for the testing set).

For the other three categories of the dependent variable (CATI, CATI-CAMI and CATI-CAWI) the percentage of correct classifications varies from 86.5 to 91.2% for the training set and from 82.6 to 93.1% for the testing set.

5 Conclusions

At the end of this *excursus* on feedforward ANNs we can summarize the most important aspects by highlighting their strengths and weaknesses.

As the phenomenon of generalization demonstrates, ANNs are capable of learning, namely, they allow solving problems by associating the sought solution with data. Indeed, network learning techniques are applications of known statistical methods (stochastic approximation) to a new class of nonlinear regression models. In this sense the determination of the network weights can be interpreted as a nonlinear regression applied to an ANN function. The advantage is to have an extremely flexible function, avoiding the subjective components of the specification error, as the parameters implicitly determine which is the latent function that a network approximates.

If the analytical form of the function underlying the problem under study is known, or can be assimilated to a known form, the problem of parameter estimation refers to the case of nonlinear least squares and the use of ANNs is not justified; it becomes so when one is not able to formulate reliable conjectures on such form. In this case, the use of networks is easier and more productive than other complex procedures with restrictive assumptions. The use of ANNs is therefore effective as a criterion for identifying hidden nonlinear relationships.

The ability to learn is related to that to forecast. ANNs offer good performances both in univariate forecasting, that is, when one wants to predict the behaviour of a variable of a system that evolves over time on the basis of its past trend, and in multivariate analysis, when trying to predict the trend of a variable observing the past behaviour of several variables of the evolving system. Many studies have highlighted how ANNs allow good approximations and extrapolations to be made. Since a forecast problem can be referred to an approximation and extrapolation problem, it is possible to use networks to approximate the regularities present in the variations over time of the variable to be predicted. ANNs flexibly adapt to complex situations that change over time, directly if learning is unsupervised; by re-training if learning is supervised. They are also suitable for processing data that are incomplete or affected by noise or biases. By virtue of this ability to adapt to data, ANNs are very robust, viz. they have a high resistance to failures and malfunctions. Another important feature is the computational

speed that derives from their parallelism and the very rapid input–output association, since the computations to be performed are weighted sums and threshold selections; therefore, they constitute a valid alternative to traditional techniques for performing complex computations.

Obviously, ANNs are not magical boxes. As we have made clear, ANNs can achieve better performance than linear methods if there are nonlinearities and interactions in input data. It should be kept in mind that just because we have data, it does not mean that there are underlying rules that can be learned. ANNs offer an approach to analysis that is data-intensive and exploratory. The focus of these methods is on computational efficiency, not modelling. Of course, the results will not necessarily be good unless the variables are. As the old adage of computer science goes: "garbage in, garbage out".

The critical points of ANNs are, first of all, the long and scarcely incremental learning; in addition to requiring a large number of epochs before significantly reducing the error, learning must be repeated when the situation represented by the patterns undergoes substantial changes, unless such learning is continuous or unsupervised.

Obviously also for ANNs, as in any other case, it is necessary to have a data set that is rich and representative (of the problem under study) so that the training set and the testing set are effectively controllable.

Other problems may arise from the low accuracy and their uncertain reliability of the results provided by ANNs: the past performances of a network do not guarantee those future. There is a risk that the generalization is not complete and that therefore most of the inputs do not recall correct outputs. Furthermore, there are no strict criteria to design the most suitable network for a given problem, but it is necessary to proceed by trial and error with, as mentioned, numerous degrees of freedom in the choice of each parameter. Moreover, each network has its own specificity. If the same experiment is repeated on another network, there will not be the same results, although in most cases they tend to converge. This is another interesting feature of ANNs; they are able to provide similar results in terms of performance with a variety of weight settings. Clearly what is important is not the value of a certain weight, but the overall set of all connection weights.

Finally, the criticism most frequently raised against the usefulness of ANNs is that, even when they succeed in the assigned task, they do not allow us to explain their operation on a cognitive level (in the case of the sociological research we could say on the level of the analysis of relationships between variables). We expect from a model not only that it will be able to predict or reproduce its referent, but also that it will be transparent, that is, it will make us understand how it works, what mechanisms, processes and principles are behind it. ANNs, according to this criticism, risk obtaining the first goal, but not the second one. A network that was able to learn a certain task and is also capable of extending its performance to new situations, showing in this way that it has incorporated the mechanisms and principles underlying that task, may nevertheless be not very transparent as to these mechanisms and principles, not making them emerge clearly and thus not allowing their full explanation regarding the phenomenon in question. Their strictly quantitative nature, the interweaving of the links, the connection weights, the effects of a local phenomenon of activation on the rest of the network, are all factors that make the behaviour of networks dark as tools for explaining the relationships between variables.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

References

Alpaydin, E.: Machine Learning: The New AI. The MIT Press, Cambridge (2016)
Athey, S.: Beyond prediction: using big data for policy problems. Science **355**(6324), 483–485 (2017)
Athey, S., Imbens, G.W.: The State of Applied Econometrics: Causality and Policy Evaluation. J. Econ. Perspect. **31**(2), 3–32 (2017)
Bail, C.A.: The cultural environment: measuring culture with big data. Theory Soc. **43**(3–4), 465–482 (2014)
Baldassarri, D., Abascal, M.: Field Experiments Across the Social Sciences. Ann. Rev. Sociol. **43**(1), 41–73 (2017)
Baldassarri, D., Goldberg, A.: Neither ideologues nor agnostics: alternative voters' belief system in an age of partisan politics. Am. J. Sociol. **120**(1), 45–95 (2014)
Barocas, S., Selbst, A.: Big data's disparate impact. Calif. Law Rev. **104**(3), 671–732 (2016)
Benzécri, J.-P.: Statistical analysis as a tool to make patterns emerge from data. In: Watanabe, S. (ed.) Methodologies of Pattern Recognition, pp. 35–74. Academic Press, New York (1969)
Benzécri, J.-P.: & Collaborateurs: L'Analyse des Données: 1. La Taxinomie. Dunod, Paris (1973a)
Benzécri, J.-P. & Collaborateurs: L'Analyse des Données: 2. L'Analyse des Correspondances. Dunod, Paris (1973b)
Benzécri, J.-P.: Correspondence Analysis Handbook. Marcel Dekker, New York (1992)
Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res. (2018). https://doi.org/10.1177/0049124118782533
Blumenstock, J., Cadamuro, G., On, R.: Predicting poverty and wealth from mobile phone metadata. Science **350**(6264), 1073–1076 (2015)
Bonikowski, B., DiMaggio, P.: Varieties of American popular nationalism. Am. Sociol. Rev. **81**(5), 949–980 (2016)
Capecchi, V.: Tre Castelli, una Casa e la Città inquieta. In: Cipolla, C., De Lillo, A. (eds.) Il sociologo e le sirene: La sfida dei metodi qualitativi, pp. 37–99. FrancoAngeli, Milano (1996)
Capecchi, V., Buscema, M., Contucci, P., D'Amore, B. (eds.): Applications of Mathematics in Models, Artificial Neural Networks and Arts. Springer, Dordrecht (2010)
Crawford, K.: Think Again: Big Data: Why the rise of machines isn't all it's cracked up to be. Foreign Policy. https://foreignpolicy.com/2013/05/10/think-again-big-data (2013) Accessed 01 August 2020
Di Franco, G.: Reti neurali artificiali e analisi dei dati per la ricerca sociale: un nuovo paradigma? Sociol. Ric. Soc. **19**(56), 35–75 (1998)
Di Franco, G.: Tecniche e modelli di analisi multivariata. FrancoAngeli, Milano (2017)
Di Franco, G.: Usi e abusi dei sondaggi politico-elettorali in Italia: Una guida per giornalisti, politici e ricercatori. FrancoAngeli, Milano (2018)
Dong, C.: The evolution of machine learning. TechCrunch. https://tcrn.ch/2vIQWXY (2017). Accessed 01 August 2020
Evans, J.A., Aceves, P.: Machine translation: mining text for social theory. Ann. Rev. Sociol. **42**(1), 21–50 (2016)
Fabbri, G., Orsini, R.: Reti neurali per le scienze economiche: I modelli del connessionismo per l'analisi statistica e la simulazione dei comportamenti economici. Franco Muzzio Editore, Milano (1993)
Fisher, R.A.: The Design of Experiments. Oliver and Boyd, Edinburgh (1935)
Fisher, R.A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh (1925)
Galton, F.: Hereditary Genius: An Inquiry into its Laws and Consequences. MacMillan, London (1869)
Galton, F.: Regression towards mediocrity in hereditary stature. J. Anthropol. Inst. G. B. Irel. **15**, 246–263 (1886)
Gifi, A.: Nonlinear Multivariate Analysis. Department of Data Theory, University of Leiden (1981)
Gifi, A.: Nonlinear Multivariate Analysis. Wiley, Chichester (1990)
Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press, Cambridge (2016)

Grimmer, J., Stewart, B.M.: Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. **21**(3), 267–297 (2013)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining Inference, and Prediction. Springer, New York (2009)
Hopkins, D.J., King, G.: A method of automated nonparametric content analysis for social science. Am. J. Polit. Sci. **54**(1), 229–247 (2010)
Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
Kitchin, R.: Big Data, new epistemologies and paradigm shifts. Big Data & Society **1**(1), 1–12 (2014a)
Kitchin, R.: The Data Revolution: Big Data, Open Data Data Infrastructures and Their Consequences. SAGE Publications, London (2014)
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z.: Prediction policy problems. Am. Econ. Rev. **105**(5), 491–495 (2015)
Lagoze, C.: Big Data, data integrity, and the fracturing of the control zone. Big Data Soc. **1**(2), 1–11 (2014)
LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
Molina, M., Garip, F.: Machine Learning for Sociology. Ann. Rev. Sociol. **45**(1), 1–25 (2019)
Mullainathan, S., Spiess, J.: Machine learning: an applied econometric approach. J. Econ. Perspect. **31**(2), 87–106 (2017)
National Research Council: Frontiers in Massive Data Analysis. The National Academies Press, Washington, D.C. (2013)
Nilsson, N.J.: The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge University Press, Cambridge (2010)
Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River (2010)
Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw **61**, 85–117 (2015)
Spearman, C.: "general intelligence", objectively determined and measured. Am. J. Psychol. **15**(2), 201–292 (1904)
Spearman, C.: The Abilities of Man: Their Nature and Measurement. MacMillan, London (1927)
Törnberg, P., Törnberg, A.: The limits of computation: A philosophical critique of contemporary Big Data research. Big Data & Society **5**(2), 1–12 (2018)
Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading (1977)
Varian, H.R.: Big data: new tricks for econometrics. J. Econ. Perspect. **28**(2), 3–28 (2014)
Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York (2019)