Machine learning and artificial neural networks

SAPIENZA
UNIVERSITÀ DI ROMA

Giovanni Di Franco
Michele Santurro

4-7 May 2020, Athens, Greece

Introduction (1/2)

- Machine learning (ML) is an automatic learning process that takes place through the processing of usually very large data sets

- In recent years in various human sciences: economics, political science, sociology, communication science, 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 or by private companies

Introduction (2/2)

- Overall, many different approaches and tools are included under the ML label, distinguished by architecture, learning rules, signal transfer function, etc.

- This presentation has two aims:
  a) to present the ML with the artificial neural network (ANN) algorithms in a simple and intuitive way;
  b) to apply them to sociological data by comparing the results obtained with the results of traditional statistical techniques, to evaluate their advantages and possible disadvantages

Literature Review (1/2)

- Historically ANNs have been proposed to emulate some functions of the human brain and nervous system, within an approach called connectionism

- The advantages offered by this approach are very interesting for the social sciences:
  - to simulate a phenomenon on a computer it is necessary to make explicit and formalize all the knowledge that is available;
  - it becomes possible to manipulate a phenomenon in ways that would not be allowed with other investigation techniques

Literature Review (2/2)

- The use of the calculator makes it possible to successfully study phenomena characterized by high dynamism, high parallelism and strong complexity, governed by rules of change that can be described by nonlinear equations

- Unlike the computational models used in expert systems, a network learns and generalizes through the experience it acquires rather than through a program that determines its behaviour

Methodology (1/4)

- Schematically an ANN consists of:
  - a large number of simple processing units (artificial neurons);
  - a large number of links between the units (artificial synapses);
  - a parallel and distributed control scheme;
  - a learning algorithm

Methodology (2/4)

- A feedforward ANN is made up of a number of units connected by links which are, in this type of network, unidirectional

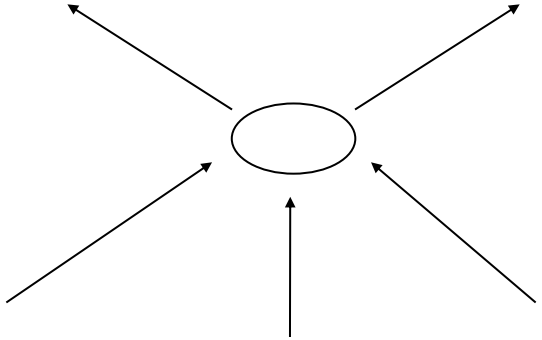

Fig. 1 – Unit with three incoming and two outgoing links

Methodology (3/4)

- The activation state of a unit is equal to a mathematical combination of all the activations and inhibitions that reach that unit through its incoming links

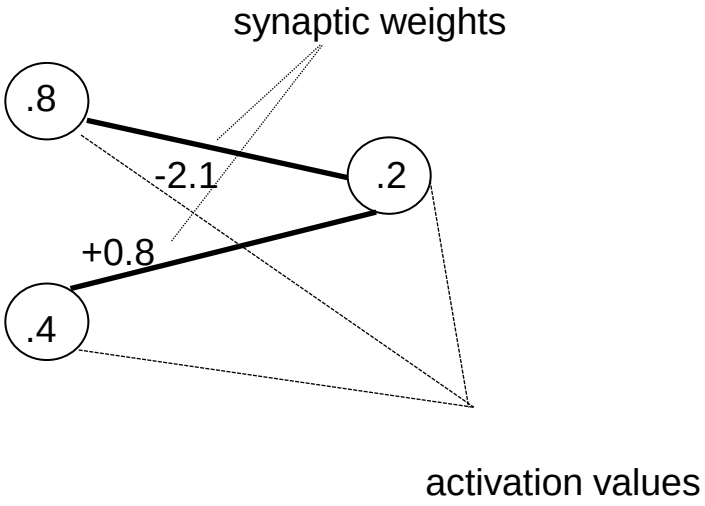

Fig. 2 – Connection weights

Methodology (4/4)

- The common feature of many ANNs is that they initially have randomly chosen connection weights. However, exposed to repeated experiences, the network progressively changes its weights so that they will produce the desired performance in the end

- Different learning techniques have been defined for the ANNs. What we will consider is a supervised learning, called error back-propagation, in which the algorithm is run for all the values of the training set until the network has reached a stable state, i.e. a minimum of the error function

Results and Discussion (1/10)

- We present an example of application which consists of a comparison between a multiple linear regression model and an ANN

- The data are taken from a matrix containing some information on the electoral political polls published in Italy by the mass media from 1 January 2017 to 29 February 2020. The information relating to these surveys was taken from the site www.sondaggipoliticoelettorali.it of the Presidency of the Council of Ministers, Department for Information and Publishing

Results and Discussion (2/10)

- The dependent variable is:
  - the percentage of voters who declared their intention to abstain or who declared their indecision regarding the election choice (‘no-vot’)

- The independent variables are:
  - the days of the survey (‘days’);
  - the sample size (‘n-sample’);
  - the completeness index of the information relating to the survey (‘ind-1’);
  - the ratio between the interview attempts and the interviews carried out (‘ind-2’)

Results and Discussion (3/10)

- We first present the results of linear multiple regression
- The four independent variables reproduce a little less than a third (31.1%) of the variance of the dependent variable

Tab. 1 – Multiple regression model summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.563[a]	.317	.311	8,4224

Results and Discussion (4/10)

- The contribution of the four independent variables is significant in explaining the variance of the dependent one

Tab. 2 – 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

Results and Discussion (5/10)

- The analysis of the residual statistics shows the good adaptation of the model to the data

Tab. 3 – Multiple regression residual statistics

	Min.	Max.	Mean	Std. Deviation	N
Predicted Value	30.684	56.227	41.261	57.111	506
Residual	-261.793	353.160	.0000	83.890	506
Std. Predicted Value	-1.852	2.621	.000	1.000	506
Std. Residual	-3.108	4.193	.000	.996	506

Results and Discussion (6/10)

- Let’s now evaluate the results obtained with the ANN. For ANN applications, we used SPSS Multilayer Perceptron procedure

- The cases submitted to the network are the same 506 used in the regression. In this case, 70% of the cases (359) were used in the training set and the remaining 30% (147) for the testing set

Results and Discussion (7/10)

- In the training set the relative error was equal to .225. It grows slightly in the testing set reaching the value of .327

Tab. 4 – 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**

Results and Discussion (8/10)

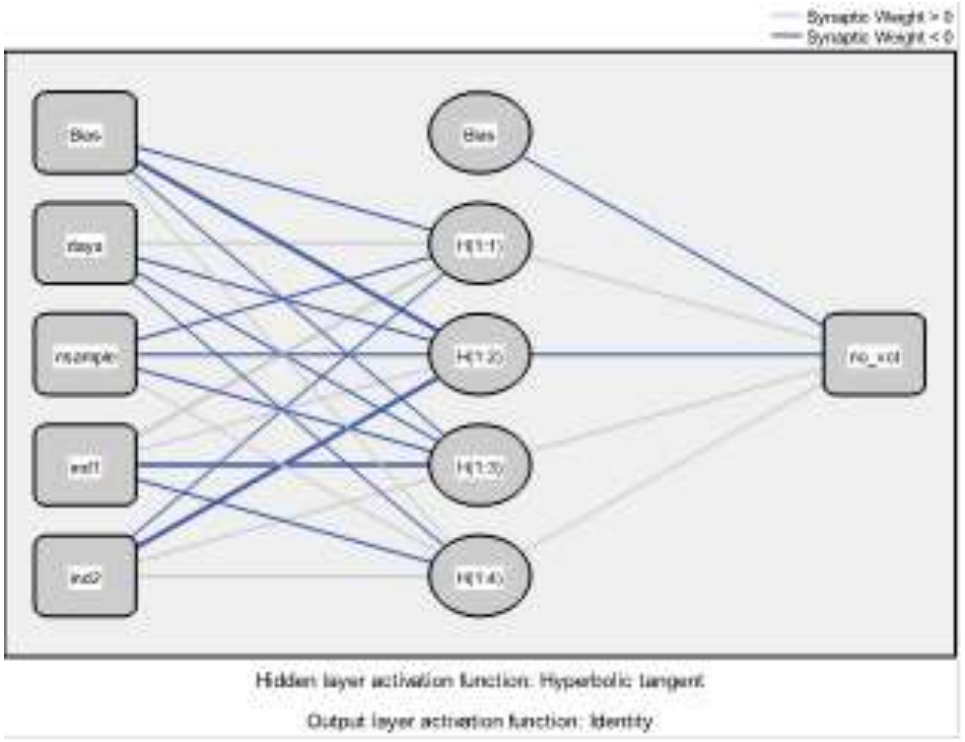

Fig. 3 – The architecture of the ANN

Results and Discussion (9/10)

- The comparison between the results of the multiple regression and the ANN leaves no doubt about the better predictive performance of the network

Tab. 5 – Correlations between predicted value of regression and ANN and the value 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 0.01 level (2-tailed)

Results and Discussion (10/10)

- 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

Conclusions (1/3)

- Network learning techniques are applications of known statistical methods (stochastic approximation) to a new class of nonlinear regression models. The determination of the network weights can be interpreted as a nonlinear regression applied to an ANN function

- The use of ANNs is therefore effective as a criterion for identifying hidden nonlinear relationships

Conclusions (2/3)

- Since a prediction problem can be referred to an approximation and extrapolation problem, it is also possible to use the networks to approximate the regularities present in the variations over time of the variable to be predicted

- ANNs are suitable for the processing of data that are incomplete or affected by noise or detection errors. By virtue of their ability to adapt to data, they are very robust, i.e., they have a high resistance to failures and malfunctions

Conclusions (3/3)

- The critical points of the ANNs are:
  - long and scarcely incremental learning;
  - also for the ANNs, as in any other case, it is necessary to have a data set that is rich and representative of the problem under study;
  - there are no strict criteria for designing the most suitable network for a given problem, but it is necessary to proceed by trial and error;
  - even when ANNs succeed in the assigned task, they do not allow to explain the relationships between variables