State of play and perspectives on machine learning at Istat

Marco Di Zio[1]

Abstract

This paper discusses the use of machine learning methods in Istat. It broadly illustrates the road taken by the Institute starting from the first works published on the use of neural networks in official statistics in '90s, namely in the field of editing and imputation, to the current situation in which the use of big data in Istat has brought a great acceleration in the study and use of machine learning methods. In fact, these techniques are very useful for dealing with problems with unstructured data and allow to exploit and take advantage of data of large volume, aspects that characterize big data. These features will be briefly illustrated by means of some cases addressed in the activities concerning Trusted Smart Statistics in Istat. In addition to this favorable context, studies have been conducted for the application of machine learning in situations characterized by the integration of administrative and survey data. To this end, it is necessary to consider how to take into account aspects that characterize a sample survey in ML models. One of the first issues addressed in the studies, concerns the introduction of sample weights into the models. An application on real data is carried out with this aim. Other issues on which future studies are intended to focus on concern the application of ML methods in the case of longitudinal data, a particularly relevant aspect when using administrative data, and in the case where units occur in clusters, for example in the case of household surveys. Finally, an element that is particularly relevant for a Statistical Institute concerns the evaluation of the quality of statistics produced by machine learning techniques with particular attention to their accuracy. Thorough reflection is needed on this point since the number provided by NSIs are generally aggregated measure as instance means, totals, quantiles of a population, and this is in fact not the usual framework of machine learning methods.

Keywords: Quality, Trusted Smart Statistics, Multisource data

[1] Istat, dizio@istat.it