DSS Statistics Seminar

November 5, 2021, 12:00

https://uniroma1.zoom.us/j/86881977368?pwd=SWRFcVFjMDZTa0lXZk05TE1zNm5adz09

Passcode: 432940

Optimal coding of categorical data in Machine Learning

Agostino Di Ciaccio

Department of Statistical Science
Sapienza University of Rome

If we have to analyze large data sets, with hundreds of features, we will generally have many quantitative variables and many qualitative variables, which could have many modalities. In classical statistics these data are very difficult to analyze, but even in machine learning an optimal approach has not been proposed.
The purpose of this presentation is to suggest a method to analyze categorical variables with many categories, in machine learning methods.
Several approaches have been proposed in the literature, in this presentation we will focus on the problem of coding categorical data in order to apply neural networks.
The traditional methods that are used to encode categorical variables can be divided into three categories: methods that do not use the target variable or other variables; methods that use only the target variable; One Hot Encoding based methods that use a dummy variable for each category.
These methods have numerous drawbacks. Starting from a definition of optimal quantification, we will see that through a low-dimensional multiple quantification we can obtain a very effective coding that allows us to build more efficient Neural Networks with a low number of parameters. Some examples will show the usefulness of this method.

SAPIENZA
UNIVERSITÀ DI ROMA