Working Group on Statistical Learning Seminar

Title: Modal clustering for categorical data

Speaker: Noemi Corsini (University of Padua)

Date: Thu 19th October 2023 at 3:00PM

Location: E0.32 (beside Pi restaurant)

Abstract: Despite the ill-posedness of the clustering task, there exists a broad consensus regarding the definition of clusters in the continuous setting. Here, the idea of similarity between subjects finds, to a greater or lesser extent, well-grounded counterparts in the notions of density and distance. Conversely, when dealing with categorical data, the lack of a total order among categories makes somewhat controversial even the notion of distance, leading to potential arbitrariness of the target to reach and undermining the soundness of the inherent methods.

A novel notion of cluster is introduced. It complies with natural intuition and relies on the twofold concept of high frequency and association between variables. Groups are defined as highly populated aggregations of cross-categories of the observed variables leading to a large contribution of mutual information.

The former concept complies with the notion of cluster described by the modal formulation of the clustering problem, which we take advantage of to borrow some operational tools. The proposed procedure jointly extends, if not formally, at least conceptually, the ideas of connected sets, gradient ascent, and density, which are typical of the

nonparametric clustering settings.