DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF SOUTHERN DENMARK, ODENSE

COMPUTER SCIENCE COLLOQUIUM

Density-Based Methods for Data Analysis: Some Recent Developments and Future Perspectives

Ricardo Campello
School of Mathematical and Physical Sciences
University of Newcastle, Australia

Tuesday, 11 June, 2019 at 14:15

DIAS conference room

Abstract:

Non-parametric density estimates are a useful tool for tackling different problems in statistical learning and data mining, most noticeably in the unsupervised and semi-supervised learning scenarios. In this talk, I elaborate on HDBSCAN*, a density-based framework for hierarchical and partitioning clustering, outlier detection, and data visualisation. Since its introduction in 2015, HDBSCAN* has gained increasing attention from both researchers and practitioners in data mining, with computationally efficient third-party implementations already available in major open-source software distributions such as R/CRAN and Python/SciKit-learn, as well as successful real-world applications reported in different fields. I will discuss the core HDBSCAN* algorithm and its interpretation from a non-parametric modelling perspective as well as from the perspective of graph theory. I will also discuss post-processing routines to perform hierarchy simplification, cluster evaluation, optimal cluster selection, visualisation, and outlier detection. Finally, I briefly survey a number of unsupervised and semi-supervised extensions of the HDBSCAN* framework currently under development along with students and collaborators, as well as some topics for future research.

Host: Arthur Zimek