Conditional density estimation

Background

Conditional density estimation (CDE) is a key problem in machine learning and data mining. Consider random vector X and Y, estimating the conditional density, denoted as f(Y|X), provides more information than simply estimating the expected conditional mean E(Y|X), which is exactly the goal of regression analysis.

Currently both tree-based methods [4,5] and kernel-based methods [1,2] for the task of CDE exist, but their practical performance is understudied, especially when the data is of mixed type (i.e., containing both categorical and continuous variables). As a result, there is a wide range of possible research questions that could be pursued in the context of a Master's thesis.

Possible research directions

One (or multiple) of the following research directions could be investigated:

- Compare impurity criteria used in tree-based methods and possibly propose improved impurity criteria.
- Compare different kernels and different bandwidth selection methods [3] in kernel-based conditional density estimation method.
- Propose a novel method by combining the above two approachs.

Requirements

- We are looking for a highly self-motivated Master student with good programming skills and experience with data science projects.
- Students who are aiming for an excellent thesis grade and who are looking for a challenging thesis project with a certain degree of research freedom are particularly welcome.
- Knowledge of at least non-parametric density estimation methods or the minimum description length (MDL) principle, and (a very basic understanding of) modern probability theory based on measure theory, including integration with respect to a measure and Radon-Nikodym derivative.

Supervisors

Lincen Yang, Matthijs van Leeuwen

References

[1] Hansen, Bruce E. "Nonparametric conditional density estimation." Unpublished manuscript (2004).

[2] Holmes, Michael P., Alexander G. Gray, and Charles Lee Isbell. "Fast nonparametric conditional density estimation." arXiv preprint arXiv:1206.5278 (2012).

[3] Bashtannyk, David M., and Rob J. Hyndman. "Bandwidth selection for kernel conditional density estimation." Computational Statistics & Data Analysis 36.3 (2001): 279-298.

[4] Cousins, Cyrus, and Matteo Riondato. "CaDET: interpretable parametric conditional density estimation with decision trees and forests." Machine Learning 108.8-9 (2019): 1613-1634.

[5] Pospisil, Taylor, and Ann B. Lee. "RFCDE: Random forests for conditional density estimation." arXiv preprint arXiv:1804.05753 (2018).