Working Group on Statistical Learning Seminar

Title: Textual data summarization using the Self-Organized Co-Clustering model

Speaker: Margot Selosse (Université de Lyon, Laboratoire ERIC)

Date: Mon 25th January 2021 at 12:00PM

Location: Online

Abstract: Recently, different studies have demonstrated the use of co-clustering, a data mining technique which simultaneously produces row-clusters of observations and column-clusters of features. Our work introduces a novel co-clustering model to easily summarize textual data in a document-term format. In addition to highlighting homogeneous co-clusters as other existing algorithms do, we also distinguish noisy co-clusters from significant co-clusters, which is particularly useful for sparse document-term matrices. Furthermore, our model proposes a structure among the significant co-clusters, thus providing improved interpretability to users. The approach proposed contends with state-of-the-art methods for document and term clustering and offers user-friendly results. The model relies on the Poisson distribution and on a constrained version of the Latent Block Model, which is a probabilistic approach for co-clustering. A Stochastic Expectation-Maximization algorithm is proposed to run the model's inference as well as a model selection criterion to choose the number of co-clusters. Both simulated and real data sets illustrate the efficiency of this model by its ability to easily identify relevant co-clusters.

https://ucd-ie.zoom.us/j/68316324831