Abstracts: Topological Data Analysis and Clustering

Friday, April 22, 2022

Dan Shiebler (Oxford)

Title: Kan Extensions for Data Science

Abstract: A common problem in data science is "use this function defined over this small set to generate predictions over that larger set." Extrapolation, interpolation, statistical inference and forecasting all reduce to this problem. The Kan extension is a powerful tool that generalizes this notion. We will first explore how the Kan extension can be used to derive a simple classification algorithm that we can run on data. Next, we use the Kan extension to derive a procedure for learning clustering algorithms from labels and explore the performance of this procedure on data.

Time: 09:00-10:00 EDT

Luis Scoccola (Northeastern)

Title: Density-based clustering, multiparameter persistence, and relative homological algebra

Abstract: Density-based approaches to clustering typically frame the clustering problem as the problem of estimating the most prominent modes of the distribution the data is assumed to have been sampled from. Some successful clustering schemes use persistence to define the notion of prominence. I will describe joint work with Alex Rolle in which we use multiparameter persistence to provide a unified view of these clustering schemes, and to define modifications of the algorithms that enjoy better stability properties and for which one can use techniques from persistence homology to do parameter selection. I will put this work into the broader context of multiparameter persistence and discuss recent connections to relative homological algebra.

Time: 10:15-11:15 EDT

Katharine Adamyk (Western)

Title: Compressions of Hierarchical Clusterings

Abstract: Many topological methods for clustering data depend on a choice of one or more real parameters. Heuristically, a (covariant) hierarchical clustering is such a method where clusters appear, grow and/or merge together as any individual parameter increases. The amount of information required to describe a hierarchical clustering can be reduced by recording only points of significant change in the clusters. Layer points and branch points are both examples of such compressions.

In this talk, I will discuss recent work on the stability of layer points with respect to subsampling, followed by a discussion of ongoing work on similar compressions of hierarchical clusterings. While these theoretical results concern hierarchical clusterings in general, there will also be a focus on examples of topologically constructed clusterings.

Time: 11:30-12:30 EDT

Leland McInnes (Tutte Institute)

Title: Clustering High Dimensional Data; challenges and opportunities

Abstract: Clustering data is a challenging problem in and of itself. However, dealing with high dimensional data, with its quirks and the curse of dimensionality, is especially challenging. This talk will first look at topological and density based clustering, and then look at why these approaches struggle when working with high dimensional data. We will then look at techniques used to mitigate some of these challenges, and how and why they might be effective. Finally we will look ahead to what a technique for successful high dimensional data clustering might look like, and potential future research directions.

Time: 13:30-14:30 EDT

Problem Session

Time: 14:45-15:45 EDT
