Research Proposal

Title: Applications of Zigzag Homology to the Study of Clustering Behavior in Social Media

Author: Martin Skilleter

Supervisor: Dr Katharine Turner

Background

Topological data analysis takes tools from algebraic topology and uses them to study large data sets. Its effectiveness has already been demonstrated in areas as wide-ranging as bioinformatics, sensor detection, market analysis and more. A major technique used is persistent homology. Persistent homology describes the changes in homology as a space evolves. Given a metric on a data set, we can form for each $\varepsilon > 0$ a simplex by connecting all points in the data set that are within ε of each other. Within this simplex, we can then count the number of "holes" by computing the dimension of the n^{th} homology group. This dimension is called the n^{th} Betti number.

Repeating this process for arbitrary ε creates a function from the positive real numbers to the nonnegative integers; given an $\varepsilon > 0$, compute the n^{th} Betti number. If the dimension stabilises at a particular value as ε varies, the space is said to have a persistent homological structure.

The theory of zigzag homology expands persistent homology. In persistent homology, a functor $H_p(-, \mathbb{F})$ is applied to a topological space to generate a sequence of homology groups

$$H_1(X) \to X_2(X) \to ... \to H_n(X).$$

Distinct homology groups do not largely influence each other, but we can relate them through the use of zigzag homology. In the diagram above, this allows arrows to point in both directions so we have naturally induced maps between higher and lower homology groups. If each dimension were to represent a variable in a data set, this would then show how holes in data sets are affected by multiple variables interacting. This information can be used to refine sampling techniques, so that more complete data is available in the future.

Aim

We will use existing tools in topological data analysis, including zigzag homology, persistent homology, and other associated methods, to describe and cluster data relating to social media usage. We will also experiment with different metrics on the sample data sets to determine their interaction with the techniques of persistent homology.

Part of the purpose of this project will be to learn the underlying theory of persistent and zigzag homology and understand how this can be applied practically. As the tools of zigzag homology are not widely tested on large data sets, we will also be developing and refining these tools to work efficiently.

Method

We will begin by understanding the mathematical theory of persistent and zigzag homology as in [2]. This paper does not consider the techniques of topological data analysis with the intention of practical application, and so we will explore the methods developed and apply them to understand social media user behavior. As the data sets we will be utilising are non-Euclidean, we will also need to consider ways of embedding the data in a suitable metric space before we can perform the clustering.

The primary data set we will use consists of Reddit posts. This project will thus involve finding suitable metrics we can place on such social media usage. There are several standard metrics which are used in topological data analysis and we will perform computations with multiple such metrics to determine which give the most useful homological data.

If possible, we will use other publicly available data sets to determine whether there are underlying features which may be more receptive to the tools of zigzag homology i.e. whether there are features which affect if the persistent homological structure of a data set offers useful insights.

Software and Hardware Requirements

The intention is that any code required for the project will be written by myself, likely using any of Python, Julia, Dionysus (a package in Python) or Eirene (a pack in Julia). We expect all computations to be made on a computer with 16-32 Gb of RAM. As we are exploring the applications of these methods to different data sets, these requirements may change.

Applications

Persistent homology, upon which zigzag homology is built, is a common computational tool used in data analysis. The intention of this project is to explore the applicability of zigzag homology to social media. If it reveals useful insights, this could have implications for the use of homological analysis in advertising.

Persistent homology is used to identify holes in data sets so that sampling techniques can be refined to minimise the number of these holes in future. It can also reveal when analysts have misinterpreted the meaning of a data sample by measuring against an expectation. By embedding data sets in a metrizable vector space, we can take the difference between the expected and actual data and consider the resultant homology. If there are non-trivial holes then there is a gap between what analysts expect and the gathered data set.

References

[1] Frédéric Chazal and Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *arXiv e-prints*, page arXiv:1710.04019, Oct 2017.

[2] Andrew Tausz and Gunnar Carlsson. Applications of zigzag persistence to topological data analysis. 08 2011.