ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute of Theoretical Computer Science
Patrick Schnider

Introduction to Topological Data Analysis **Scribe Notes 26** **FS23**

Scribe notes by Simon Weber. Please contact me for corrections.
Lecture date: June 2, 2023
Last update: Friday 2nd June, 2023, 12:55

Applications (continued)

Time series

Time series is data given as a sequence of points x_t in some metric space X, where t is a (discrete) variable. The goal in analyzing time series is often analyzing and finding periodic behaviour in the time series.

We can embed time series in some higher-dimensional space, by always considering a sequence of $l+1$ consecutive points, i.e., $\mathcal{I}_l := \{(x_{t_i}, x_{t_{i+1}}, \dots, x_{t_{i+l}})\} \subset X^{l+1}$. The idea is that periodicity in the time series translates to loops (1-dimensional holes) in $\mathcal{I}_l$.

As an example, consider $x_t = \sin(\frac{\pi}{4}t)$, and consider $l = 1$. Then, we get a loop in $\mathbb{R}^2$. If we however consider $x_t = t$, $\mathcal{I}_1$ is just a straight line.

This approach has been used for analyzing motion capture data, where cameras track the location in 3-space of certain points on a human body marked by physical markers. The data considered are six seconds long recordings of movement such as boxing. Every data point in this set is roughly 70-dimensional. Using PCA into 2 dimensions, some loops can be seen but we cannot really distinguish different loops. In persistent homology, there are six different loops that persist a long time. Looking back at the input data, these six different loops corresponded to six different boxing movements.

Politics

The data considered in this application is gathered from the sessions of the US house of representatives through the year 2010. For every vote, we set $x_i = 1$ if the member i says Yes to the vote, -1 for No, and 0 otherwise. This gives 664 datapoints in $\mathbb{R}^{447}$. Doing 2-dimensional PCA and coloring the data by splitting the members into the two parties, we can see that there are four main "corners" in which issues lie, ones that get bipartisan support, those that get bipartisan rejection, and those that are supported by republicans and rejected by democrats, and those where it is the opposite way. We can see that there are almost no votes that lie in between two of these corners, and

especially few that lie in the middle of all four corners. We can also see that the corner with bipartisan rejection is very sparsely populated, since such issues rarely make it far enough to be voted on.

If we do persistent homology of the same data, we see nothing. Why? Persistent homology is quite vulnerable to low-density noise. The very few points in the middle of the 2-axis diagram quickly fill in the perceived hole and kill it in persistent homology. If we again estimate the density around each data point and only consider the densest 99%, we start seeing a very clear hole.

Shape segmentation

Given a three-dimensional shape by points, edges, and triangles, we want to label different parts of the shape. For example, given a model of a human body, we want to segment it into categories such as "head", "torso", "upper arm", etc.

We can pick some point on the body, and start growing a ball around it, using the geodesic distance (length of the shortest path along the surface). On the resulting filtration, we can perform persistent homology. If we do this for a point on the palm of a hand, for example, we get a one-dimensional hole for every finger. If we do this for a point on a finger, the persistence diagram looks very different. We can then classify the persistence diagrams to segment the shape. But, how can we do this? We need to somehow insert persistence diagrams into classical ML pipelines.

TDA in ML

Many machine learning pipelines require input points to be in Euclidean space (and not just any metric space, which the persistence diagrams would already fulfill), or in the case of kernel methods, at least in some space that has an inner product.

There are many ways to turn persistence diagrams into elements of metric spaces. These methods are also called vectorizations. On https://persistent-homology.streamlit.app one can see examples and play around with various vectorization methods.

Persistence Statistics

Persistence statistics are measures we can analyze parts of a persistence diagram on, for example we can take the mean, standard deviation, median, interquartile range, full range, percentiles, etc., of birth times, death times, midpoints, or lifespans. This already gives a very large number of features, which hopefully captures enough information about the persistence diagram.

Persistence Landscapes

We can draw a horizontal and vertical line segment between each point in a persistence diagram to the diagonal. Flipping this arrangement of line segments such that the diagonal is horizontal, we can look at the k-th envelope, the function describing the height of the k-th highest line segment at each point on the diagonal. These envelopes are piecewise linear, and piecewise linear functions lie in an L^p-space, which is an inner product space and thus allows Kernel methods to be applied.

Betti Curves

The Betti curve is the function $\beta : \mathbb{R} \to \mathbb{N}_{\geq 0}$ which assigns each time t the current Betti number. This is again a piecewise linear function, capturing all the births and deaths, but throwing away the pairing between births and deaths.