ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute of Theoretical Computer Science
Patrick Schnider

Scribe notes by Simon Weber. Please contact me for corrections.
Lecture date: June 1, 2023
Last update: Thursday 1st June, 2023, 14:00

Applications

The Space of Image Patches

We consider a dataset of about 4000 greyscale images taken around Groningen (Netherlands). Each such image is described by giving each pixel a number between 0 (white) and 1 (black). From each such image, we sample 5000 patches of three by three pixels. To get interesting patches, we only look at the top 20% with the highest contrast. Thus, we end up with about 4 million three by three patches, each describable as a vector in $\mathbb{R}^9$. We further want to normalize the contrast (such that the darkest and brightest pixel are 1 and 0, respectively) and the overall norm of the vectors. Thus, we end up with points on S^7. How does this point cloud look? Is there some interesting structure in this point cloud?

Since there are a lot of points, the researchers conducting this experiment only sampled from the densest parts of the data. They computed the persistent homology of the Witness complex in dimensions 0, 1, and 2, and found that $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 0$, suggesting that the data set looks like a circle. PCA also shows a circular arrangement of the data points. This circle can be interpreted as the possible angles of a detected "edge" in the images.

What happens if we additionally also sample some points from parts of the data of intermediate density? Suddenly, instead of one 1-cycle, we end up with five. If we again consider the PCA, we see a cross in the middle of the circle seen previously. However, this only shows four circles! Where is the fifth one? Most likely some cycle is orthogonal to the projection chosen by PCA. The interpretation picked from this situation is that the dataset consists of three great circles of some S^2 that are all orthogonal. We can interpret these three circles as the circle described above, a circle describing different translations of vertical edges, and one circle describing translations of horizontal edges.

However, if we finally also include points from the lowest density parts of the data, we start seeing some feature in the persistent 2-homology, and reduce β_1 to most likely 2 again. How can this space look? It must somehow include the three circles found before.

We can embed the circles as in the figure (on the slides), indicating that the space is a Klein bottle.

However, this is not a proof. The persistent homology computed agrees with the Klein bottle, but it would also agree with a torus. Here, we can compute persistent homology over a different field, with which we can distinguish between Klein bottles and tori.

Diabetes and Breast Cancer

The following application comes from the first paper describing the Mapper approach. The data comes from a study in 1979 on 145 participants, with six quantities measured per participant. After applying various classical projection methods, the original study came up with a picture containing a blob of healthy people, with two strains coming out of this blob, called type 1 and type 2 diabetes. This is something that Mapper should easily detect: As the filter function they used a density estimator for each point. In the Mapper you also see areas of large density, with 2 flares going out of it.

The Mapper approach has later also been applied to genomic data from a breast cancer study. This data is very high dimensional, each patient is a data point in $\mathbb{R}^{262}$. As a filter function they used the distance to some baseline healthy tissue. With this method, they found a type of tumor which has previously not been classified, while also confirming the strains of breast cancer known previously. Since then, topological data analysis has been used in many studies in medicine. TDA seems to excel in these high-dimensional datasets since many features seem to be the result of higher-order interactions of different coordinates. TDA can also deal with much smaller data sets than other approaches such as deep neural networks, that need tons of data points to train, validate, etc. In medicine, studies often have very few data points due to the large monetary cost, workload, and also ethical questions involved with data gathering.