Persistent

Daniel Collin

January 30, 2021

Contents

1 Introduction

Using other measures than standard statistical analysis for datasets is an avenue that is explored in homology. From homology we can get a topological invariant of the data, interesting and can be further used as features in machine learning etc. Carlsson et al. describes it as exploring the shape of the data.

Persistent homology follows a basic principle that we can approximate a simplicial complex on a point cloud. This of course requires that the field is a metric space, because etc..

This thesis will serve as both an introduction the workings of persistent homology as well as (an?) example of persistent homology applied to a real dataset. Of course, since it is still quite a young field it is not entirely possible to answer to us what actual value persistent homology has as a tool within data science.

Etc etc.

2 Homology

Before go into what *persistent* homology it is well worth our time to clearly state what we mean by homology. (Why? Can this be skipped by experienced readers or are our definitions non-standard? Do we mostly follow hatcher?). In a general sense, homology is a particular of invariant of topological spaces. This has categorical reasons and others. Importantly we need to define simplicial complexes. There are other ways of defining this, notably singular homology, but for the computational aspect of persistent homology we do not have to dwell on this. For completion, we refer the reader to Hatcher for a more traditional treatment of homology.

2.1 Simplices

First we start with the simplex. An n-simplex is the smallest possible convex set in $\mathbb{R}^m$ containing $n + 1$ points $v_0, \ldots, v_n$ such that the vectors $v_1 - v_0, \ldots, v_n - v_0$ are linearly independent. The points $v_0, \ldots, v_n$ are known as the *vertices* of the simplex. The *standard* n-simplex is defined as the n-simplex with vertices being the unit vectors along coordinate axes

$$\Delta^n := \{(t_0, \ldots, t_n) \in R^{n+1} \mid \sum_i t_i = 1, t_i \geq 0 \quad \forall i\}$$

Definition. A face of a simplex is the convex hull of a subset of its vertices. (orientations?)

2.2 Simplicial complex

Definition. A simplicial complex K is a finite collection of simplices such that

1. $\sigma \in K$ and $\tau \subset \sigma$ implies that $\tau \in K$

2. $\sigma_1, \sigma_2 \in K$ implies that $\sigma_1 \cap \sigma_2$ is either empty or a face of both.

This is the geometric definition of a simplicial complex. However, since we are working with topological spaces it is advantageous to think of an abstract simplicial complex without concerning ourselves with the geometric connotations:

Definition (book). An abstract simplicial complex A is a finite collection of sets such that $\alpha \in A$ and $\beta \subseteq \alpha$ implies that $\beta \in A$

This abstract definition coincides with the geometric definition by calling the elements of A its simplices. The simplices of A are no longer geometric objects in Euclidean space, but simply combinatorial objects consisting of vertex sets.

It is easy to see how one can go from a geometric simplicial complex to an abstract simplicial complex simply by forgetting everything but the vertices themselves. However, most of the time our interest lies in the opposite direction: how do we go from an abstract simplicial complex to a geometric one? This is done by the geoemtric realization of A.

Theorem. Every abstract simplicial complex of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$.

Proof. See 12312424.

From here on we will simply refer to abstract simplicial complexes as a simplicial complex unless stated otherwise.

2.3 Simplicial homology

For a simplicial complex K of dimension n we define a free abelian group C_k on the oriented $k-$simplices of K. The elements of C_k are called k-chains and are formal sums of the type $\sum \alpha_i \sigma_i$ where α_i are coefficients in some ring R and σ_i are k-dimensional simplices. Furthermore, we have a collection of homomorphisms, known as boundary maps, which together with the groups form a chain complex. The kth boundary map

$$\partial_k : C_k \to C_{k-1}$$

takes a k-simplex to its boundary

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

where $\hat{v}_i$ signifies that this vertex has been omitted. This is a homomorphism so

$$\partial_k \sum \alpha_i \sigma_i = \sum \alpha_i \partial_k \sigma_i$$

Now a simplicial chain complex is a collection of chain groups together with their corresponding boundary maps as a sequence:

$$\ldots \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} C_{k-2} \xrightarrow{\partial_{k-2}} \ldots$$

Note that the boundary maps compose to become the zero map. From this definition we know that from every simplicial complex K we can associate a simplicial chain complex (this is a functor). We then define the kth homology group of K as the quotient group

$$H_k(K) = Ker(\partial_k)/Im(\partial_{k+1})$$

3 Persistent

In the world of data we rarely have access to perfect information. In the view of homology, perfect information would be knowing exactly what our topological space is and so we could calculate the homology group of that space. We could endow our the space which our data lives in with a topology, but if we give it the discrete topology the homology of that space will be trivial. What we are interested is what if there is an underlying topological space with a non trivial topology? Consider for example points sampled from a torus. If we know our space is a torus we know what the homology is of this space, but what if we do not? This is where persistent homology comes in, a way of gaining information about the homological structure of the data space.

The basic idea is quite simple. We let our data points be vertices in a simplicial complex and grow balls of radius ϵ around each vertex. If any two balls intersect we say that they are part of the same simplex.

Figure of this here.

However, a problem occurs here. How large do we want ϵ to be? If we grow ϵ too large we end up with all balls intersecting each other and with no homological information and if we let the balls grow too little then with a bunch of unconnected vertices and no interesting structure. Persistent homology addresses this by simply considering *all* of them and encoding the lifetime of homological features occuring in something called a barcode diagram.

Figure of barcode.

3.1 Views

We said that we put balls around each point, but this is just inuition. There are multiple ways of doing this which have different advantages and disadvantages.

(Explain the different complexes.) Cech complex more expensive to compute.

3.1.1 Cech complex

Definition. For a given selection of points $\{x_\alpha\}$ in some Euclidean space $\mathbb{R}^n$ the Cech complex C_ϵ is given by the abstract simplicial complex whose k-simplices are given by $k + 1$ points in the collection of points whose closed balls of radius $\epsilon/2$ have a point in common.

3.1.2 Vietoris-Rips Complex

Definition. For a given selection of points $\{x_\alpha\}$ in some Euclidean space $\mathbb{R}^n$ the Vietoris-Rips complex R_ϵ is the abstract simplicial complex whose k-simplices are given by $k+1$ points which are pairwise at most ϵ apart.

3.1.3 Witness complex

3.1.4 Cubical complex

Theorem 1. *hehe*

3.2 Persistence Diagrams

3.3 Barcodes

3.4 Metrics

3.5 Computation of

Some aspects of the computational part of this. How is it done in practice?

4 Data