Topological Data Analysis Project Background Chapter

Shanik Dassenaike

February 9, 2018

id15663@my.bristol.ac.uk

What is Topological Data Analysis?

Topological Data Analysis (TDA) is a rapidly-evolving technology for analysing data. It uses techniques from the mathematical field of topology to provide a greater insight into data. To perform TDA, we must consider our data as a shape of some sort. Indeed, this is exactly the key insight from topology that we take advantage of: data has shape, and moreover, this shape is important. TDA has the potential to show us structures and patterns that are hidden from traditional methods of data analysis.

Chazal and Michel's paper of October 2017 explains the approach from a data science perspective, describing "fundamental and practical aspects" of TDA. Although largely mathematical and theoretical in nature, it provides a basis from which to understand TDA, as well as an illustration of a practical implementation for data science, using the Python library Gudhi. Furthermore, the paper outlines a pipeline from which most TDA implementations build up. Lum et al., produced a paper in February 2013 describing applications, showing how TDA can extend traditional analysis techniques such as principal component analysis, and ultimately "find subgroups in data sets that traditional methodologies fail to find". This paper demonstrates that TDA is applicable to a wide range of datasets, applying their method to gene expression data, voting data, and basketball players' performance data. It is clear from these papers that there is a strong foundation and case to be made for TDA, and that it can indeed practically produce information or describe relationships that are not otherwise easily found with more traditional techniques.

Mathematical Foundations

The following section is intended to provide a mathematical basis for the work to follow. We assume the reader has knowledge of basic Linear Algebra and

Group Theory.

Affine Spaces

Affine Spaces are structure that generalise vector spaces, avoiding concepts of distance and measure of angles; the only properties that remain are related to parallelism and ratio of lengths of parallel line segments.
Suppose we had a vector space V over a field $\mathbb{F}$. Let A be a non-empty set. For any vector $\mathbf{v} \in V$ and element $p \in A$, we define addition $p + \mathbf{v} \in A$ with the following conditions:

- $(p + \mathbf{v}) + \mathbf{w} = p + (\mathbf{v} + \mathbf{w})$
- $\forall q \in A, \exists! \mathbf{w} \in V$ s.t. $q = p + \mathbf{w}$
- $p + \mathbf{0} = p$ (Note this is implied by the above)

Then A is an affine space, and $\mathbb{F}$ is called the coefficient field.

An intuitive understanding of affine spaces is to consider them as vector spaces in which we have 'forgotten' where the origin is. As such, no vector has a unique origin and thus cannot be associated with any particular point.
An example of an affine space is the plane in $\mathbb{R}^3$ defined by $< x, y, 1 >$, i.e. the xy-plane sitting at $z = 1$; this is clearly not a vector space as it does not contain the origin. However, we might still subtract two points in this plane and obtain a vector, just as in a vector space. On the other hand, in this space we cannot just take a point and find a vector with it, as there is no origin to define this vector from; similarly, we cannot add two points together to obtain a vector as they are not measured relative to an origin.
Many notions carry over from vector spaces to affine spaces in some way; suppose we have an affine space A and a subset $X \subseteq A$. Then:

- The smallest affine subspace containing X is called the affine span of X
- X is affinely independent if the affine span of any proper subset of X is a proper subset of the affine span of X.

Simplicial complexes

Discretised objects, such as mathematical graphs, or digital images, can be represented by simplicial complexes. These are collections of "well-glued" bricks called simplices.

Simplex

A simplex generalises the notion of a triangle or tetrahedron to arbitrary dimensions. Formally, suppose we had $k+1$ affinely independent points, $u_0, u_1, ..., u_k \in$

$\mathbb{R}^k$. The convex hull of these points is the k-simplex σ they determine; it is the set of points

$$C = \left\{ \theta_0 u_0 + ... + \theta_k u_k | \sum_{i=0}^{k} \theta_i = 1 \text{ and } \theta_i \geq 0 \ \forall i \right\}$$

We have that a 0-simplex is a point, a 1-simplex an edge, a 2-simplex a triangle, a 3-simplex is a tetrahedron and the notion generalises on. The dimension of σ, $dim(\sigma) = k$.
For a simplex σ, any non-empty subset of the points generating σ whose convex hull itself is a simplex is called a face of σ.

Simplicial Complex

We can "glue together" simplices to form a simplicial complex.
A simplicial complex Σ is a finite set of simplices that satisfies the following "gluing" conditions:

- Any face of a simplex in Σ is itself in Σ
- The intersection of any two simplices $\sigma_1, \sigma_2 \in \Sigma$ is a face of both σ_1 and σ_2

The dimension of a simplicial complex Σ, $dim(\Sigma)$, is the largest dimension of its simplices.

What is a Topological Space?

- We have a set of points, X
- An open set is a subset of X.
- A topology is then a set of open sets $T \subset 2^X$, such that:
  1. If $S_1, S_2 \in T$, then $S_1 \cap S_2 \in T$
  2. If $\forall j \in J$, we have that $S_j \in T$, then $\bigcup_{j \in J} S_j \in T$
  3. $\varnothing, X \in T$
- We then achieve $\mathbb{X} = (X, T)$ is a topological space.
- Different topologies are possible
- A metric space is an open set defined by some metric

Using TDA for images

The aim of this paper is to demonstrate the use of TDA for image data. For example, we would like to know, after processing an image in the usual way, if this data is obfuscated or 'jumbled' in some way, will TDA techniques still be able to recover the original image? Will it be able to recognise it as the same image if given the source image? If TDA techniques can recover or distinguish the image data, is this in general - is there some way that we can 'mix up' the data that removes the topological structure that TDA is identifying, without destroying the data itself? On the other hand, how similar would images need to be for their topological structure to be the same, or very similar?

We will take advantage of three key topological concepts: Coordinate Invariance, Deformation Invariance and Compressed Representations.

1. **Coordinate Invariance:** Topology studies shapes in a coordinate-free manner; what that means is that the coordinate system in which we view a shape does not affect the properties of it that we study. The topological constructions depend only on the distance function intrinsic to the metric space within which the shape is specified.

2. **Deformation Invariance:** The properties studied in topology are also invariant under "small" deformations - despite any "stretching" or "squashing", as long as the shape is not "torn" or "reglued", any property topology studies is unchanging. For example, if we were to write down the capital letter "A" on an elastic surface, and then stretched it in some directions, it will still retain its closed triangle and two legs pointing out. Similarly, a capital "A" written in different fonts is still clearly a letter A (and in fact humans are particularly good at recognising deformation-invariant properties), as the fundamental parts of the letter haven't changed.

3. **Compressed Representations:** oftentimes an object that we want to study is very highly (potentially infinitely) complex in its detail and information contained. Topology allows us to approximate the object with a finite representation - a triangulation. This may mean identifying the object with a simplicial complex or a network, for example identifying a sphere with an icosahedron or a circle with a hexagon. In either case we go from infinite points on a surface to a small, finite number of points, edges and faces, so we lose some information (with these approximations curvature is an example) but we retain the important topological feature e.g. a loop

Finally, to undertake this analysis, we will follow the TDA pipeline as outlined in Chazal and Michel:

1. Assume the input is a finite set of points, with a notion of distance/similarity between them.

2. Build a ”continuous” shape on top of the data, to highlight the topology underpinning it. Frequently, this is a simplicial complex or a nested family of simiplicial complexes - a filtration - reflecting how the data is structured at different scales.
3. Extract topological or geometric information from the structures built on top of the data.
4. Achieve new families of features and descriptors of the data, from the extracted topological and geometric information.