Introduction

“Topological data analysis is the use, in computation, of techniques from the mathematical field of topology to learn about data. In particular, we find a way of transforming our data into a topological shape, from which we can apply topological techniques to compare, contrast, or learn about this data. The key insight from topology that we use here is that data fundamentally has shape, and that shape matters. Furthermore, topology has some key concepts that lend power to analysing data from this: Coordinate Invariance, Deformation Invariance, Compressed Representation.”

“Coordinate invariance means that the properties studied by topology are not affected by the coordinate system in which we view the shape, so long as we maintain the distance metric.”

“Deformation invariance means that the properties studied by topology are not affected by ‘small’ deformations – i.e. if we do not tear or reglue a shape, any stretching or squashing of it will not affect any properites.”

“Compressed representations means that instead of looking at data or an object directly, we approximate it with a finite representation, e.g. a hexagon for a circle; we have now 6 vertices and 6 edges rather than an infinite surface to study. Although some information, like curvature in this example, is lost, we maintain fundamental information such as the loop”.

“We are going to have to define a fair bit of maths here to understand how to apply these techniques; in order to perform this topological data analysis, we need a representation that conforms to a topological space, so we will start by defining this.”

“If we have a set of points X, a subset of these points is called an open set. A topology is then a set of open sets T, (and a subset of 2^X), where: the intersection of any two elements of T is itself in T; the union of any elements of T is in T; the empty set and the entire set X are elements of T. This leads us to defining the tuple (X, T) which is a topological space.”

“So we know what a topological space is. But how does this relate to our data? To obtain such a space from data, we often want to define the data as a simplicial complex. To understand what this is, we need to understand the ideas of affine independence, simplices and finally simplicial complexes.”

“We build up from affine spaces: affine spaces generalise vector spaces, effectively by having no origin. If A is an affine space and X is a subset of A, then the smallest affine subspace containing X is 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. A simplex is the arbitrary-dimensional generalisation of the notion of a triangle or a tetrahedron. To be more formal, a k-simplex is the convex hull of k+1 affinely independent points. The dimension of a simplex is k (where k+1 is the number of affinely independent points whose convex hull form the simplex. A face of a simplex sigma is any non-empty subset of the points that generate sigma, whose convex hull itself is a simplex. These notions can now lead us to being able to define simplicial complexes.”

“A simplicial complex is formed by ‘gluing together’ simplices. Formally: a simplicial complex Sigma is a finite set of simplices that satisfies the following gluing conditions: Any face of a simplex in Sigma is itself in Sigma. The intersection of any two simplices sigma1 sigma2 in Sigma is a face of both sigma1 and sigma2. The dimension of a simplicial complex Sigma, dim(Sigma) is the largest dimension of its simplices.”

“Once we have a simplicial complex, we can associate a topological space to it – this is called its geometric realisation. A Filtration is a nested family of simplicial complexes”.

“In this project I aim to investigate how such topological methods can be applied to image data. For example, if we take a certain image and process it so that we obtain some clusters of information, and subsequently convolve, obfuscate, or otherwise mix up this data, can topology help us to recover information about the original image? Comparing the original representation to the more ‘jumbled’ one, can topological data analysis techniques show us that these contain the same latent information – or if we alter an image slightly, can we distinguish these? We want to find out what properties of images, when processed in the normal way, are topologically invariant, and what topological structure is retained when we confuse the original data in some way.”

“There are pre-existing tools for TDA, such as (TTK, the topology toolkit), (Python Mapper) (JavaPlex) and more, which are all open- source, that we can use to perform TDA. Commercially, Ayasdi is the company leading the industry in TDA. Most existing methods are based on the following pipeline, outlined in Chazal and Michel’s paper of October 2017:

1. Assume the input is a finite set of points, with some sort of distance metric between them.
2. We build a continuous shape on top of the data, allowing us to highlight its topology (this is often a simplicial complex or a filtration)
3. Extract topological information from these structures
4. Obtain new families of features and descriptors.”