1 Overview

1.1 Motivation

Before we jump into any of this, we ask the basic question: why should we care?

Suppose you're given a big pile of data, and told to find interesting things about it. For being such a vague-sounding task, it's actually pretty common – tons of research is centered around looking for structure in big collections of images, or social networks, or large sets of signals, etc., where we might not know exactly what we're looking for, but we have good reason to believe there *are* a lot of interesting things we should be able to find. This is the purview of *unsupervised learning*, a subset of machine learning that's had a lot of influence from computational algebraic topology (CAT). Unsupervised learning is all about finding structure we might not be able to notice; CAT does this with some additional structure around what might make something that is interesting to look for and that is likely not to be noise.

Topology is the study of the properties of mathematical shapes where we're not allowed to "tear" apart or "glue" together shapes; that is, where we can continuously deform the shape. In other words, it's less concerned with *where* things are than *how* they're all connected together. In data that has one to three dimensions, we don't really have to be concerned with any of this. We can plot our data and just look at the shape, and that's way easier to do than reasoning about the interconnections. But in higher dimensions, data are often sparse and it's not obvious what we're looking for. By focusing on connections, we can build a more systematic idea of what we think is "really important" and why.

1.2 Overview

The basic structure in CAT is the *simplicial complex*, which is essentially "a graph with multidimensional vertices", where you can cluster together any number of points instead of just two. (In particular, this means graph $\subset$ simplicial complex, so for anything that follows you can just read "graph" where I've written "simplicial complex" and it'll work.) Formally, you could say a simplicial complex connects combinatorics and geometry, but to my mind that just invites more questions. Here's how I like to think of it.

In learning multivariable calculus and linear algebra, we took ideas we learned in one or two dimensions with visual aids, and learned how to use them as a scaffold to think about notions of volume, surfaces, subspaces, principal directions, and so on, in $\mathbb{R}^n$, where we would otherwise have no hope of understanding what's going on. We'll do the same thing here – simplicial complexes are things it makes sense to draw out (if you have few enough dimensions that you can), so what do their shapes look like? Can we make sense of them as subsets of $\mathbb{R}^n$, like we did with the kernel and range of a matrix?

These shapes can get more complicated than they do in linear algebra; for some very scientific evidence, I always have a much harder time drawing a tetrahedron – the simplest "solid" simplicial complex (in graph language, a complete graph) 3D – than a cube. This is where the "topology" part of CAT comes in. We introduce the idea of *homotopy equivalence* being able to smoothly (infinitely-differentiably) map between two simplicial complexes, because if you can do that it'd seem to signal "it doesn't matter how these are laid out in space, these are actually kind of the same thing!" in a way that might be useful to us. We'll talk about *homology*, an easier way to compute homotopy equivalence by finding fixed quantities like Euler's $V + E - F = 2$.

After that there's a couple of chapters I don't understand yet, but they help with building tools for the next big idea: *persistence*. If you look at a feature in data at several different "resolutions" or "length scales", how does it change? Is it even still there if you look at it really closely and then from really far away? Can multiple "lenses" tell us something about what's signal and what's noise? From what I've seen, this is the concept it's easiest to apply to visualizing high-dimensional data in a way that gets at the Thing of Interest way more easily than we'd otherwise be able to find. I also don't understand the last two chapters, entitled "sheaves" and "gradients", but I'll get there!

I chose CAT as the first topic I wanted to teach myself because I was able to draw all these connections to questions I'd practically care about if I were trying to analyze data. Hopefully, over the course of doing this, I'll develop a whole new toolkit for my scientific life!