APPLIED TOPOLOGY LECTURE NOTES

CHAD GIUSTI

UPDATED: OCTOBER 20, 2017

Topology of Data

When we move to the realm of science and engineering one of the fundamental changes from that of mathematics is that we only get to measure things, rather than assuming we know what we're studying – to put it another way, we move from topological spaces to a finite sample of points from those spaces. However, with a bit of cleverness, we can still use our topological tools to recover information about the underlying spaces in a sensible way.

Mapper

We want to understand the organizational structure underlying the point cloud $r(S)$. Clustering algorithms tend to collapse a lot of the detail in $r(S)$ while they're looking for component labels – this detail may tell us a great deal of interesting information about A and ℓ. Thus, we want to develop methods that retain that information.

One powerful tool is the *Mapper* algorithm. Mapper disassembles the point cloud into smaller pieces, and uses clustering in those individual pieces as building blocks for a low-dimensional visualization of the geometry structure of the point cloud.

Definition. Let $r(S) \subset X$ be a point cloud and fix a function $f : X \to \mathbb{R}$. Let $\mathcal{U} = \{U_i = (a_i, b_i)\}_{i=1}^m$ be a cover of $\mathbb{R}$ by overlapping open intervals $U_i = (a_i, b_i)$ with $a_1 = -\infty, b_m = \infty$ and $U_i \cap U_j \neq \varnothing$ if and only if $|i - j| \leq 1$. Write $r(S)_i = f^{-1}(U_i) \cap r(S)$. Select a clustering algorithm, and for each $r(S)_i$, perform the algorithm to obtain clusters $C_i^1, \ldots C_i^{p_i}$. The *mapper graph for f and $\mathcal{U}$* is the graph with vertices given by the collection of clusters for each U_i, $V = \{C_i^p\}$, and an edge $C_i^p C_j^q$ when $C_i^p \cap C_j^q \neq \varnothing$.

More generally, we can take $f : X \to Y$ for any topological space Y and use an open cover of Y to build a combinatorial structure. In this case, we build a *mapper complex* with simplices for each non-empty intersection of clusters.

Mapper is implemented commercially by its inventors through a company called Ayasdi, in a package called Ayasdi Core. The software offers a range of powerful visualization tools for the mapper

graph, as well as statistical tools for understanding the structure of the data set via manipulation of cluster

Simplicial complexes from data

Another option, the most common one in topological data analysis, is to use the points in $r(S)$ and the proximity function μ to build a simplicial complex directly. There are several possibilities.

Definition. Let $\epsilon > 0$ and P be a point cloud. The Čech complex for P at scale ϵ is the simplicial complex $\check{C}_\epsilon(P)$ with vertices P and simplices $\sigma \subseteq P$ whenever $\bigcap_{x\in\sigma} B_\epsilon(x) \neq \varnothing$.

The Čech complex is precisely the nerve of the canonical open cover by epsilon balls $B_\epsilon(x)$ of the space $\bigcup_{x\in P} B_\epsilon(x)$. Therefore, by the nerve theorem the simplicial complex built in this way recovers the homotopy type and, thus, has the proper homology for this union of local neighborhoods. If we were lucky enough to choose ϵ so that the structure of this union reflects the structure of interest, this is fantastic for us.

Unfortunately, the Čech complex is hard to compute – it involves looking at intersections of sets, which is an expensive task computationally. If we are fortunate enough to have μ symmetric, we can approximate the Čech complex using only the matrix $M(r(S))$.

Definition. Let μ a symmetric proximity function on X, $\epsilon > 0$ and P be a point cloud. The Vietoris-Rips complex for P at scale ϵ is the simplicial complex $VR_\epsilon(P)$ with vertices P and simplices $\sigma \subseteq P$ whenever $M(r(S))_{i,j} < \epsilon$ for all $i \neq j \in \sigma$.

The Vietoris-Rips complex is the clique complex of the graph $\Gamma_\epsilon(P)$ with vertices P and edges whenever the proximity of two points is less than ϵ. This saves the effort of checking for intersections in the ambient space, reducing the problem to finding cliques in a graph – which is still a hard computational problem, but more approachable in general.

In the special case when μ is a metric[1] (so X is a metric space), we have the following guarantee that there is a relationship between the Vietoris-Rips and Čech complexes.

[1] A *metric* is a function $\mu : X \times X \to \mathbb{R}_{\geq 0}$ for which (i) $\mu(x,y) = \mu(y,x)$, (ii) $\mu(x,y) = 0$ only if $x = y$ and (iii) $\mu(x,y) \leq \mu(x,z) + \mu(y,x)$ for every z.

Lemma 1. *Let $\epsilon > 0$ and P a point cloud in a metric space X, then*

$$\check{C}_\epsilon(P) \subseteq VR_\epsilon(P) \subseteq \check{C}_{2\epsilon}(P).$$

Proof. Exercise. □

Thus, if we want to approximate one of these complexes by another, we can "bound" the structure on either end by shifting ϵ.

Finally, suppose we have a collection of *witnesses* $\{w_1, \dots w_m\} \subseteq X$, and rather than measuring distance between the points in P, we have measurements of distance to these witnesses. Witnesses model sensors with limited range in a space.

Definition. Let P be a point cloud in X and $\{w_1, \dots w_m\} \subset X$ a collection of witnesses, $\epsilon > 0$. The *witness complex* $W_\epsilon(P)$ at scale ϵ is the simplicial complex with vertices P and simplices $\sigma \in W_\epsilon(P)$ if $\mu(x_i, w_j) < \epsilon$.

The witness complex is a special case of a general construction of a simplicial complex for rectangular matrices called a *Dowker complex*.

Definition. Let M be an $n \times m$ matrix, $\epsilon > 0$. The *Dowker complex* of M at scale ϵ, $\text{Dow}_\epsilon(M)$ is the simplicial complex with vertices $[n]$ and a face $\sigma_j, j = 1 \dots m$ for each column given by $i \in \sigma_j \Leftrightarrow M_{ij} < \epsilon$.

The process for constructing a Dowker complex may result in repeated addition of faces to the complex; this is not a problem because faces form a set. The witness complex is just the Dowker complex at scale ϵ for the matrix with rows corresponding to points in the cloud, columns corresponding to witnesses, and entries corresponding to their proximity.

The choice to make rows into vertices and columns into faces seems arbitrary, and it is. We could consider instead the alternative complex with roles reversed, which would just be $\text{Dow}(M^T)$. These complexes are called Dowker complexes because of his remarkable observation about how this choice doesn't matter.

Theorem 2 (Dowker's theorem)**.** *Let M be a matrix, $\epsilon > 0$. Then $H_*(Dow_\epsilon(M)) \cong H_*(Dow_\epsilon(M^T))$.*

Persistent homology

Of course, all of these constructions depend on a choice of scale parameter ϵ, and we definitely don't want to have to make such a choice about data we don't yet understand. Fortunately, we can simply decide not to make a choice.

Write S_ϵ for any choice of $\check{C}_\epsilon(P), VR_\epsilon(P)$, or $\text{Dow}_\epsilon(M)$. Choose a new scale paramter $\epsilon' > \epsilon$, and observe that $S_\epsilon \subseteq S'_\epsilon$ in each case: the vertices remain the same, but we might introduce some new higher simplices. Thus, there is a canoncial inclusion map

$$\iota_{\epsilon,\epsilon'} : S_\epsilon \hookrightarrow S_{\epsilon'}$$

which is the identity on the vertices. If we follow this down to the level of homology, we get

$$(\iota_{\epsilon,\epsilon'})_* : H_*(S_\epsilon) \to H_*(S_{\epsilon'}).$$

Further, observe that since there are finitely many points in P or elements in M, ϵ-balls will form new intersections or new vertices will be added to simplices in $\mathrm{Dow}_\epsilon(M)$ only finitely many times. Thus, the structure of S_ϵ can only change finitely many times as ϵ increases. We record the parameters at which this structure changes as a list $-\infty = t(0) < t(1) < \cdots < t(k) < t(k+1) = \infty$, and observe that $S_\epsilon = S_{\epsilon'}$ if $\epsilon, \epsilon' \in (t(i), t(i+1)]$, and thus $(\iota_{\epsilon,\epsilon'})_*$ is an isomorphism, so the only interesting information is contained in the maps $(\iota_{t(i),t(i+1)})_* : H_*(S_{t(i)}) \to H_*(S_{t(i+1)})$.

Definition. Let $S_{t(0)}, S_{t(1)}, \ldots, S_{t(k+1)}$ be simplicial complexes so that $\iota_{t(i),t(i+1)} : S_{t(i)} \hookrightarrow S_{t(i+1)}$ is an injective homomorphism. Such a sequence S of simplicial complexes is called a *filtration*. For $j > i$, the $(t(i), t(j))$*-persistent homology in degree* p of the filtration is

$$H_p^{t(i)\to t(j)}(S) = \mathrm{im}((\iota_{t(i),t(j)})_p).$$

If $[x] \in H_p(S_{t(i)}), [x] \neq [0]$ and $[x] \notin H_p^{t(i-1)\to t(i)}(S)$, we say $[x]$ is *born* at $t(i)$ and write $b_{[x]} = t(i)$. If $[x] \in H_p(S_{t(i)}), [x] \neq [0]$ is born at $t(i)$, $[\iota_{t(i),t(j)}(x)] \neq [0] \in H_p^{t(i)\to t(j)}(S)$ but $[\iota_{t(i),t(j+1)}(x)] = [0] \in H_p^{t(i)\to t(j+1)}(S)$ then we say $[x]$ *dies* at $t(j+1)$ and write $d_{[x]} = t(j)$. The *lifetime* of $[x]$ is $\ell_{[x]} = d_{[x]} - b_{[x]}$.