Notes on Topological Data Analysis

Zhang Liu

February 10, 2021

Abstract

This document is to serve as a set of notes to fill the gaps in my understanding of Topological Data Analysis relevant to the project. The reference book for this set of notes is *Algebraic Topology* [2], *Topology and Data* [1], and *Topology for Computing* [3].

1 Motivation

First of all, there are four major advantages for using topological methods to deal with point clouds in data analysis.

1. Topology provides qualitative information which is required for data analysis.
2. Metrics are not theoretically justified. Compared to straightforward geometric methods, Topology is less sensitive to the actual choice of metrics.
3. Studying geometric objects using Topology does not depend on the coordinates.
4. Functoriality. This is the most important advantage.

   Definition 1. For any topological space X, abelian group A, and integer $k \geqslant 0$, there is assigned a group $H_k(X, A)$.

   For any A and k, and any continuous map $f : X \to Y$, there is an induced homomorphism $H_k(f, A) : H_k(X, A) \to H_k(Y, A)$. Then *functoriality* refers to the following conditions:

   - $H_k(f \circ g, A) : H_k(f, A) \circ H_k(g, A)$
   - $H_k(Id_X; A) = Id_{H_k(X,A)}$.

   Functoriality addresses the ambiguities in statistical clustering methods - in particular the arbitrariness of various threshhold choices. We now illustrate how exactly functoriality could be used in questions related to clustering.

   Let X be the full data set and X_1, X_2 are the subsamples from the data set. If the set of clusterings $C(X_1), C(X_2), C(X_1 \cup X_2)$ correspond well (this notion will be defined formally in later section), then we can conclude that the subsample clusterings correspond to clusterings in the full data set X.

2 Homotopy

Definition 2. *Homotopy* is a family of maps $f_t : X \to Y$ where $t \in I$ such that $F : X \times I \to Y$ defined by $F(x,t) \mapsto f_t(x)$ is continuous.

Definition 3. Two maps f_0, f_1 are *homotopic* if $\exists$ a homotopy f_t between f_0 and f_1.

A special case of homotopy is the deformation retraction.

Definition 4. A *deformation retraction* of X onto a subspcae A is a homotopy from the identity map of X to a retraction of X onto A, $r : X \to X$ such that $r(X) = A$ and $r|_A = \mathbb{1}$ (or equivalently, retraction is the map $r : X \to X, r^2 = r$).

Retraction is the topological analog of projection. To visualize this analogy, we give an example of how some deformation retractions arise from the mapping cylinder.

Definition 5. For a map $f : X \to Y$, the *mapping cylinder* M_f is the quotient space of the disjoint union $(X \times I) \sqcup Y$.

Definition 6. A map $f : X \to Y$ is a homotopy equivalence if there is a map $g : Y \to X$ such that

- $f \circ g$ is homotopic to the identity map on Y, and
- $g \circ f$ is homotopic to f.

Two spaces X, Y are *homotopy equivalent* if there exiss a homotopy equivalence $f : X \to Y$.

Definition 7. If f and g are homotopic, then $H_k(f, A) = H_k(g, A)$. Then it follows that if X and Y are homotopy equivalent, then $H_k(X, A) \cong H_k(Y, A)$.

Definition 8. For any field F, $H_k(X, F)$ will be a vector space over F. Then if F is finite dimensional, its dimension is referred to as the k-th Betti number with coefficients in F, denoted as $\beta_k(X, F)$.

The k-th Betti number corresponds to an informal notion of the number of independent k-dimensional surfaces. If two spaces are homotopy equivalent, then all their Betti numbers are equal.

Note that the Betti numbers can vary with the choice of the coefficients in F.

3 Simplicial Complexes

Definition 9. An *abstract simplicial complex* is a pair (V, Σ), where V is a finite set, and Σ is a family of non-empty subsets of V such that

$$\sigma \in \Sigma, \tau \subseteq \sigma \implies \tau \in \Sigma.$$

Associated to a simplicial complex is a topological space $|(V, \Sigma)|$. $|(V, \Sigma)|$ may be defined using a bijection $\phi : V \to \{1, 2, \ldots, N\}$ as the subspace of $\mathbb{R}^N$ given by the union

$$\cup_{\sigma \in \Sigma} c(\sigma),$$

where $c(\sigma)$ is the convex hull of the set $\{e_{\phi(s)}\}_{s \in \sigma}$, where e_i denotes the ith standard basis vector.

We often use abstract simplicial complexes to approximate topological spaces. For simplicial complexes the homology can be computed using only the linear algebra of finitely generated $\mathbb{Z}$-modules. In particular, for simplicial complexes, homology is algorithmically computable (unlike the standard methods for computing the Smith normal form).

Definition 10. Let X be a topological space, and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be any covering of X.

The *nerve* of $\mathcal{U}$, denoted by $N\mathcal{U}$, will be the abstract simplicial complex with vertex set A, and where a family $\{\alpha_0, \ldots, \alpha_k\}$ spans a k-simplex if and only if $U_{\alpha_0} \cap U_{\alpha_1} \cap \cdots \cap U_{\alpha_k} \neq \emptyset$.

One reason that this construction is very useful is the following "Nerve Theorem." This theorem gives the criteria for $N(\mathcal{U})$ to be homotopy equivalent to the underlying topological space X.

Theorem 11. *Suppose that X and U are as above, and suppose that the covering consists of open sets and is numerable. Suppose further that for all $\emptyset \subseteq A$, we have that $\bigcap_{s \in S} U_s$ is either contractible or empty. Then $N(\mathcal{U})$ is homotopy equivalent to X.*

Definition 12. For any subset $V \subseteq X$ for which $X = \bigcup_{v \in V} B_\varepsilon(v)$, one can construct the nerve of the covering $\{B_\varepsilon(v)\}_{v \in V}$. This construction is referred to as the "Čech complex" attached to V and is denoted as $\check{\mathrm{C}}(V, \varepsilon)$.

Theorem 13. *Let M be a compact Riemannian manifold. Then there is a positive number e so that $\check{\mathrm{C}}(M, \varepsilon)$ is homotopy equivalent to M whenever $\varepsilon \leqslant e$. Moreover, for every $\varepsilon \leqslant e$, there is a finite subset $V \subseteq M$ so that the subcomplex of $\check{\mathrm{C}}(V, \varepsilon) \subseteq \check{\mathrm{C}}(M, \varepsilon)$ is also homotopy equivalent to M.*

However, this construction is computationally expensive. A solution is to construct a simplicial complex which can be recovered solely from the edge information, which motivates the following construction known as the "Vietoris-Rips complex."

Definition 14. Let X be a metric space with metric d. Then the *Vietoris-Rips complex* for X, attached to the parameter ε, denoted by $VR(X, \varepsilon)$, will be the simplicial complex whose vertex set is X, and where $\{x_0, \ldots, x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leqslant \varepsilon$ for all $0 \leqslant i, j \leqslant k$.

Proposition 15. *Comparing the Čech complex and the VR compelx:*

$$\check{\mathrm{C}}(X, \varepsilon) \subseteq VR(X, 2\varepsilon) \subseteq \check{\mathrm{C}}(X, 2\varepsilon).$$

However, even the VR complex is computationally expensive. A solution, again, is to the Voronoi decomposition which studies the subspaces of Euclidean space.

Theorem 16. *Let X be any metric space, and let $\mathcal{L} \subseteq X$ be a subset (called the set of landmark points). Given $\lambda \in \mathcal{L}$, we define the Voronoi cell associated to λ, V_λ, by*

$$V_\lambda = \{x \in X | d(x, \lambda) \leqslant d(x, \lambda')\} \forall \lambda' \in \mathcal{L}.$$

Definition 17. Similar to how we define the Čech complex above, we define the Delaunay complex attached to $\mathcal{L}$ to be the nerve of this covering.

However, for finite metric spaces, the Delaunay complex generically produces degenerate (i.e. discrete) complexes with no 1-simplices. To solve this, we modify the definition to accommodate pairs of points which are "almost" equidistant from a pair of landmark points. We thus have the definition below:

Definition 18. Let X be any metric space, and suppose we are given a finite set $\mathcal{L}$ of points in X (called the landmark set), and a parameter $\varepsilon > 0$. For every point $x \in X$, we let m_x denote the distance from this point to the set $\mathcal{L}$, i.e., the minimum distance from x to any point in the landmark set.

Then we define the*strong witness complex* attached to this data to be the complex $W^s(X, \mathcal{L}, \varepsilon)$ whose vertex set is $\mathcal{L}$, and where a collection $\{l_0, \ldots, l_k\}$ spans a k-simplex if and only if there is a point $x \in X$ (the witness) so that $d(x, l_i) \leqslant m_x + \varepsilon$ for all i.

We can also consider the version of this complex in which the 1-simplices are identical to those of $W(X, \mathcal{L}, \varepsilon)$, but where the family $\{l_0, \ldots, l_k\}$ spans a k-simplex if and only if all the pairs (l_i, l_j) are 1-simplices. We will denote this by W^s_{VR}.

A modified version of the strong witness complex is also useful:

Definition 19. We construct the*weak witness complex*, $W^w(X, \mathcal{L}, \varepsilon)$, attached to the given data by declaring that a family $\Lambda = \{l_0, \ldots, l_k\}$ spans a k-simplex if and only if Λ and all its faces admit ε weak witnesses.

Similar to the definition for strong witness complex, we can also consider the version of the weak witness complex in which the 1-simplices are identical to those of $W(X, \mathcal{L}, \varepsilon)$, but where the family $\{l_0, \ldots, l_k\}$ spans a k-simplex if and only if all the pairs (l_i, l_j) are 1-simplices. We will denote this by W^w_{VR}.

4 Category Theory Pre-requisites

> It frames a possible template for any mathematical theory: the theory should have nouns and verbs, i.e., objects, and morphisms, and there should be an explicit notion of composition related to the morphisms; the theory should, in brief, be packaged by a category.
>
> Barry Mazur, "When is one thing equal to some other thing?"

Definition 20. A *category* consists of

- a collection of objects $X, Y, Z, \ldots$
- a collection of morphisms $f, g, h, \ldots$

so that:

- Each morphism has specified domain and codomain objects; the notation $f : X \to Y$ signifies that f is a morphism with domain X and codomain Y.
- Each object has a designated identity morphism $\mathbb{1}_X : X \to X$.
- For any pair of morphisms f, g with the codomain of f equal to the domain of g, there exists a specified composite morphism gf whose domain is equal to the domain of f and whose codomain is equal to the codomain of g, i.e.,:

$$f : X \to Y, g : Y \to Z \rightsquigarrow gf : X \to Z.$$

A natural question to ask is: what is a morphism between categories? This leads to the definition of a functor:

Definition 21. A *functor* $F : C \to D$, between categories C and D, consists of the following data:

- An object $F_c \in D$, for each object $c \in C$.
- A morphism $Ff : Fc \to Fc' \in D$, for each morphism $f : c \to c' \in C$, so that the domain and codomain of Ff are, respectively, equal to F applied to the domain or codomain of f.

The assignments are required to satisfy the following two functoriality axioms:

- For any composable pair $f, ginC, Fg \cdot Ff = F(g \cdot f)$.
- For each object c in C, $F(\mathbb{1}_c) = \mathbb{1}_{Fc}$.

Put concisely, a functor consists of

1. a mapping on objects and
2. a mapping on morphisms that preserves all of the structure of a category, namely domains and codomains, composition, and identities.

As already mentioned in Section 1, functoriality plays a key role in topological data analysis. []

Definition 22. Let A be a ring. A *left module* M over A consists of an abelian group (also denoted M) and a law of composition $A \times M \to M$ (denoted $(a, x) \mapsto ax$) such that

$$a(bx) = (ab)x \text{ for } a, b \in A, x \in M, \tag{1}$$
$$1x = x \text{ for } x \in M, \tag{2}$$
$$(a + b)x = ax + bx \text{ for } a, b \in A, x \in M, \tag{3}$$
$$a(x + y) = ax + ay\text{for } a \in A, x, y \in M. \tag{4}$$

(4) asserts that $\rho(a) : M \to M$defined by $\rho(a)(x) = ax$ is an endomorphism of the underlying abelian group of the module, while the first three statements assert that $\rho : A \to End(M)$ is a ring homomorphism. Conversely, given such a homomorphism, we may define a module structure on M by setting $ax = \rho(a)(x)$.

Analogically, we can define the right module over A.

Note that the concept of module is a generalization of the concept of vector space. The condition that "a vector space is finite dimensional" generalizes to the condition that "a module is finitely generated". A basis of a module is a generating set that is linearly independent over the ring. Unfortunately, such sets rarely exist: only free modules have bases. Usually, we have to consider a (minimal) system of generators instead of a basis.

Definition 23. A ring R is called *graded* (or more precisely, $\mathbb{Z}$-graded) if there exists a family of subgroups $Rn_{n\in\mathbb{Z}}$ of R such that

1. $R = \oplus_n R_n$ (as abelian groups), and
2. $R_n \cdot R_m \subseteq R_{n+m} \forall n, m$.

Definition 24. Let R be a graded ring and M an R-module. We say that M is a *graded R-module* (or has an R-grading) if there $M_n n \in \mathbb{Z}$ of M such that

1. $M = \oplus_n M_n$ (as abelian groups), and
2. $R_n \cdot M_m \subseteq M_{n+m} \forall n, m$.

If $u \in M \setminus \{0\}$ and $u = u_{i_1} + \cdots + ui_k$ where $u_{i_j} \in R_{i_j} \setminus \{0\}$, then $u_{i_1}, \ldots, ui_k$ are called *the homogeneous components* of u.

5 Persistence

The main idea of *persistence* is that instead of selecting a fixed value of the threshhold ε, we would like to obtain a useful summary of the homological information for all the different values of ε at once.

Definition 25. Let $\underline{C}$ be any category, and $\mathcal{P}$ a partially ordered set. We regard $\mathcal{P}$ as a category $\underline{\mathcal{P}}$ in the usual way, i.e. with object set $\mathcal{P}$, and with a unique morphism from x to y whenever $x \leqslant y$. Then by a $\mathcal{P}$-persistence object in C we mean a functor $\phi : \mathcal{P} \to C$.

More concretely, it means a family $\{c_x\}x \in \mathcal{P}$of objects of C together with morphisms $\phi : xy : cx \to cy$ whenever$x \leqslant y$, such that $\phi_{yz} \circ \phi_{xy} = \phi_{xz}$ whenever $x \leqslant y \leqslant z$. Note that the $\mathcal{P}$-persistence objects in C form a category in their own right, where a morphism F from ϕ to Φ is a natural transformation. Again, in more concrete terms, a morphism from a family $\{c_x, \phi_{xy}\}$ to a family $\{d_x, \Phi_{xy}\}$ is a family of morphisms $\{f_x\}$, with $f_x : c_x \to d_x$, and where the diagrams commute:

$$\begin{tikzcd}
c_x \arrow[r, "\phi_{xy}"] \arrow[d, "f_n"'] & c_y \arrow[d, "f_y"] \\
d_x \arrow[r, "\psi_{xy}"] & d_y
\end{tikzcd}$$

Although we do not have a classification theorem for $\mathbb{R}$-persistence abelian groups, which would then provide a summary of the behavior of the homology of all the complexes $\check{\mathrm{C}}(X, \varepsilon)$, we do have a classification theorem for a subcategory of the category of $\mathbb{N}$-persistence F-vector spaces, where F is a field.

From the $\mathbb{R}$-persistence simplicial complexes, we just need any partial order preserving map $\mathbb{N} \to \mathbb{R}$ to obtain an $\mathbb{N}-$persistence simplicial complex. Then we can use the classification theorem.

There are at least two useful ways to construct such maps.

References

[1] Gunnar Carlsson, *Topology and data*, BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY **46** (2009), 255–308.

[2] Allen Hatcher, *Algebraic topology*, Cambridge University Press, 2002.

[3] Afra J. Zomorodian, *Topology for computing*, Cambridge University Press, 2005.