Brief Introduction to Topological Data Analysis : Perspective of Statistics Student

Taegyu Kang

Department of Applied Statistics, Yonsei University

November 13, 2019

Overview

1. Introduction
2. Topological Space and Its Representation
3. Topological Data Analysis
4. Persistent Homology
5. Mapper
6. Further Discussion on Topological Data Analysis

Motivation : Simple Linear Regression

With simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

We assumed that the true shape of data (x, y) is a 1-dimensional line in $\mathbb{R}^2$. In other words, each point p in data can be represented by a coordinate system $p \mapsto (x(p), \beta_0 + \beta_1 x(p))$.

Motivation : Principal Component Analysis

Let $X = \{x_1, \cdots, x_n\}$ be the given data and each $x_j \in \mathbb{R}^d$. Principal component analysis assumes that the true shape of data is a p-dimensional affine subspace of $\mathbb{R}^d$. Thus the true shape of data can be expressed by a basis of that affine subspace.

Motivation : Nonlinear Regression

Nonlinear regression model

$$y = m(x) + \epsilon, \quad \epsilon \sim (0, \sigma^2)$$

assumes that the true shape of data can be represented by a coordinate system $p \mapsto (x(p), m(x(p)))$.

Motivation : Manifold Learning

Manifold learning (or nonlinear dimension reduction) assumes that the true shape of the given data $X = \{x_1, \cdots, x_n\}$ is a p-dimensional submanifold in $\mathbb{R}^d$. In other words, it can be represented by an atlas $\{(U_i, \phi_i)\}_{i \in \mathcal{I}}$.

Motivations

All the previous examples are methods to estimate the true shape or pattern of the data. Some methods have simple structure, so they are easy to be interpreted. But such simple models entail very strict restrictions in re-constructing true shape of data. Other methods entail weaker restrictions but it is still not enough to cover the case of data having much more complex shape.

Primitive Structure of Data

Functional relationship or manifold structure may be strict in some cases. To alleviate such restriction, we have to consider a primitive structure of data.

In addition, such primitive structure at least include a notion of "relationship" such as closedness between two elements.

Topological Space

Topology is the most primitive mathematical structure where we can consider the notion of closedness.

(Definition) Topological Space

Let X be a nonempty set. A topology $\mathcal{T}$ on X is a subset of $\mathcal{P}(X)$ satisfying

1. $\emptyset, X \in \mathcal{T}$
2. for any family of sets $\{U_i\}_{i\in\mathcal{I}} \subset \mathcal{T}$, $\cup_{i\in\mathcal{I}} U_i \in \mathcal{T}$.
3. for any finite family of sets $\{U_1, \cdots, U_n\} \subset \mathcal{T}$, $\cap_{i=1}^{n} U_i \in \mathcal{T}$.

a pair $(X, \mathcal{T})$ is called a **topological space**.

Equivalence in the Category of Topological Spaces

Even though the outward representation of two topological space may be different, they can be same in the sense of topology.

(Definition) Homeomorphism

Let X, Y be two topological spaces, $f : X \to Y$ be a map between them. If f is continuous and bijective, and f^{-1} is also continuous, we call f a **homeomorphism** between X and Y. In this case, we say that X and Y are **homeomorphic**.

This is the analogous notion of **isomorphic** in the category of algebraic structures.

Topological Space as an Estimating Target

Likewise other traditional methods in statistics, topological data analysis aims to estimate the underlying true structure of the given data. However, topological data analysis differs from other methods in that it does not entail any strict structural restrictions.

More precisely, topological data analysis aims to estimate the underlying topology of given data.

Algebraic Topology : Theory of Representing Topological Space

In statistics, we often estimate the representation of our estimation target instead of estimating it directly.

(e.g.)

- parametric regression : target = mean response function, representation = parameter
- linear dimension reduction : target = dimension reduction subspace, representation = basis of dimension reduction subspace

Thus we need an object representing topological spaces.

Algebraic Topology : Theory of Representing Topological Space

Algebraic topology is a branch of mathematics specialised in constructing algebraic structure which can represent topological spaces.

In other words, the ultimate goal of algebraic topology is to construct a functor $\mathcal{F}$ from the category of topological spaces to the category of some algebraic structures satisfying $\mathcal{F}(X) = \mathcal{F}(Y)$ iff $X \cong Y$.

With data, we estimate $\mathcal{F}(X)$ instead of estimating X directly.

Representation of Algebraic Structure : Structure Theorem

There are various types of theorems depicts the characterisation of algebraic structure. One of popularly used theorem is the following. Topological data analysis also depends on such type of theorem.

Theorem (Structure Theorem of Finitely Generated Modules on a PID)

Let M be a finitely generated module over a principal ideal domain R, then M can be expressed as following

$$M \cong \bigoplus_{i=1}^{n} R/(d_i)$$

where $d_1 | \cdots | d_n$, and this representation is unique.

Limitation of Tools from Algebraic Topology

However, there is no such tool yet. Precisely, most of tools $\mathcal{F}$ developed in algebraic topology are well-defined. i.e., $X \cong Y \Rightarrow \mathcal{F}(X) = \mathcal{F}(Y)$, but not perfectly characterise topological spaces.

Moreover, these tools identify much wider class of topological spaces called **homotopic equivalence**.

Homotopy

(Definition) Homotopy

Let $f, g : X \to Y$ be two continuous functions. A map $H : [0,1] \times X \to Y$ is called a homotopy between f and g if it satisfies

1. $H(0,x) = f(x)$ for all $x \in X$
2. $H(1,x) = g(x)$ for all $x \in X$
3. H is a continuous map from $[0,1] \times X$ to Y.

In this case, we write $f \simeq g$.

Homotopy

(Definition) Homotopy Equivalence

Let X, Y be two topological spaces. X and Y are said to be **homotopy equivalent** if there exist continuous maps $f : X \to Y$, $g : Y \to X$ such that

1. $f \circ g \simeq 1_Y$
2. $g \circ f \simeq 1_X$

In this case, we write $X \simeq Y$.

Homotopy Equivalence and Algebraic Topology

Most of tools $\mathcal{F}$ in algebraic topology satisfy

$$X \simeq Y \Rightarrow \mathcal{F}(X) = \mathcal{F}(Y)$$

But we cannot say that $X \cong Y$ or $X \simeq Y$ even though $\mathcal{F}(X) = \mathcal{F}(Y)$. This limitation prevents us from re-constructing the topology of data accurately, so its utilization may be restricted.

Where to Use This Method

As I mentioned before, estimating algebraic representation is not suitable for use in re-construct the exact topological space.

However, it is a good way of classify the structure of given data or determine whether there are structural differences between several data sets or not.

Tools from Algebraic Topology

1. Homotopy Theory
   For given topological space X, construct a group $\mathcal{F}(X) = \pi_p(X)$ where $\pi_p(X)$ is called p-th homotopy group of X.

2. Homology Theory
   For given topological space X, construct $\mathcal{F}(X) = \big(H_p(X)\big)_p$ where $H_p(X)$ is called p-th homology group of X.

Homology used in Topological Data Analysis

There are various homology theories in algebraic topology (e.g : singular homology, homology of CW complexes, simplicial homology, etc.), but simplicial homology is the one can be derived via algorithmic calculation.

Thus most of method is topological data analysis rely on the theory of simplicial homology.

Idea of Persistency

Every data entails some noise in it. To estimate the true shape of data, it is important to distinguish the true shape and the shape made by noise.

The idea is persistency is simple. Suppose that our result of estimation depends on some parameter t (think of it as time), then we can consider the sequence of our result $(R_t)_t$. If some shape persist for a long time, we regard them as true shape. Some shape may disappear in very short time period, then we regard them as noise.

Construction of Persistent Homology

Let $X = \{x_1, \cdots, x_n\}$ be a given data set and suppose that X can be embedded into some metric space (M, d), then we can calculate the distance between each pair of elements in X.

Let $\epsilon > 0$ be fixed, and consider the ball $B(x_j; \epsilon) = \{y \in X \mid d(y, x_j) < \epsilon\}$ for each $j = 1, \cdots, n$.

Construct the corresponding set $\mathcal{S}(X)$ as following:
For any finite subset $J \subset \{1, \cdots, n\}$, if $\cap_{j \in J} B(x_j\ \epsilon) \neq \emptyset$, then $J \in \mathcal{S}(X)$.

Construction of Persistent Homology

Now sort all the elements in $\mathcal{S}(X)$ with respect to its degree. More precisely, construct the following subsets of $\mathcal{S}(X)$ as following:

Let $\mathcal{S}_0(X) =$ all elements in $\mathcal{S}(X)$ whose cardinality $= 1$ (0-dimension)
Let $\mathcal{S}_1(X) =$ all elements in $\mathcal{S}(X)$ whose cardinality $= 2$ (1-dimension)
and so on $\cdots$.

Construction of Persistent Homology

With each $\mathcal{S}_p(X)$, construct free module $C_p(X)$ over a PID R. Then each $C_p(X)$ is a finitely generated module over a PID.

The most popular choice of $R = \mathbb{Z}_2$ (It has advantages on calculation). Since $\mathbb{Z}_2$ is a field, the corresponding $C_p(X)$'s are vector spaces over $\mathbb{Z}_2$. But theoretically, the choice of a coefficient ring R is not important (see the universal coefficient theorem in homological algebra).

Construction of Persistent Homology

Define a map $\partial_p : \mathcal{S}_p(X) \to C_{p-1}(X)$ called a boundary map which gives the oriented boundary of elements in $\mathcal{S}_p(X)$, then by the property of free object, this map can be extended into a well-defined module homomorphism on $C_p(X)$.

Thus we get a chain complex

$$\cdots \overset{\partial_{p+2}}{\to} C_{p+1} \overset{\partial_{p+1}}{\to} C_p \overset{\partial_p}{\to} C_{p-1} \overset{\partial_{p-1}}{\to} \cdots$$

Define $H_p^{\epsilon}(X) = Ker(\partial_p) \big/ Im(\partial_{p+1})$. We call it a p-th simplicial homology group of X with coefficient in R.

Construction of Persistent Homology

For each ϵ, we can construct $(H_p^\epsilon)_p$, and there is a natural inclusion map from $C_p^{\epsilon_1}(X) \to C_p^{\epsilon_2}(X)$ whenever $\epsilon_1 < \epsilon_2$. By this property, we can construct a module homomorphism $f_{\epsilon_1 \to \epsilon_2}^p : C_p^{\epsilon_1}(X) \to C_p^{\epsilon_2}(X)$, and a corresponding module homomorphism from $H_p^{\epsilon_1}(X) \to H_p^{\epsilon_2}(X)$.

Thus for each p, we obtain a chain

$$\cdots \to H_p^{\epsilon_1}(X) \to H_p^{\epsilon_2}(X) \to H_p^{\epsilon_3}(X) \to \cdots ,$$

and we classify elements in homology groups into two classes, persist for a long time or die in a short time.

Construction of Persistent Homology

This naive description of persistent homology is explained with detail in many references. See (Carlsson, 2009) and (Zomorodian and Carlsson, 2005)

Visualisation via Topological Data Analysis

There is an approach to visualise the topology of very high-dimensional and complex data simply. The method called **Mapper** (Singh et al., 2007)

Construction of Mapper

Mapper depends on the idea on coordination. For a chosen coordinate function f on data set X, consider the clustering with respect to the value of f.

Precisely, Let $\mathcal{P}$ be a overlapping partition of $Im(f)$. For each $p \in \mathcal{P}$, let $U_p = f^{-1}(p) \subset X$. Then we obtain a set of clusters $(U_p)_p$. Then using the metric space structure of X, construct sub-clusters estimating the connected component of each U_p. Then we can obtain the 1-dimensional visualisation of data X.

Discussion on Mapper

Mapper reduces the given data into 1-dimensional object, so intuitively, it reduces much information from the data.

Also, construction of Mapper visualisation depends on the choice of f, and it is also very sensitive to the choice of $\mathcal{P}$. The choice of f needs a lot of domain knowledge.

How to Distinguish Noise and True Shape

There are several discussions on the problem of distinguishing noise and true shape of data using topological summary.

(Fasy et al., 2014) derive the confidence sets to determine whether an observed shape is noise or not in a single persistence diagram.

(Mileyko et al., 2011) describes the properties of the space of persistence diagram under a certain conditions and metric structure, and probability distribution on that space.

Inference with Topological Summary

Another issue is performing statistical inference with topological summaries obtained via methods from topological data analysis.

(Bubenik, 2015) transforms persistence diagrams into objects called persistence landscape and concerns the Banach space structure of persistence landscape, and use the probability theory on Banach space to obtain asymptotic results about persistence diagram.

Further Topics in Topological Data Analysis

- Information contained in topological summary
  The amount of information in topological summaries should be quantified to make a better statistical inference using them.

- Decision theory framework
  To quantify the quality of topological summaries and make a decision with them, appropriate object function should be defined on the space of topological summaries. It is also needed for the harmony of machine learning and topological data analysis.

- Algorithm to obtain another homology theory or homotopy theory
  If it is possible, we can obtain much sharper topological summaries.

References

Bubenik, P. (2015). Statistical topological data ananlysis using persistence landscape. *Journal of Machine Learning Research 16*, 77–102.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society 46*, 225–308.

Fasy, B. T., F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh (2014). Confidence sets for persistence diagrams. *The Annals of Statistics 42*, 2301–2339.

Mileyko, Y., S. Mukherjee, and J. Harer (2011). Probability measures on the space of persistence diagrams. *Inverse Problems 27*.

Singh, G., F. Mémoli, and G. Carlsson (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. *Eurographics Symposium on Point-Based Graphics*.

Zomorodian, A. and G. Carlsson (2005). Computing persistent homology. *Discrete Computational Geometry 33*, 249–274.