Topological Data Analysis

Akhil Lohia

June 8, 2017

Topics to be covered

- Manifold Learning
- Intro to Topology
- Persistent Homology: The tool that takes a heavy duty dataset and gives a topological summary
- Functorality: Very abstract mathematical concept - for clustering algorithms
- Hodge Theory (for statistical ranking)
- Mapper Algorithm

Manifold Learning (aka NLDR)

Basic problem: Curse of dimensionality

$M \subset \mathbb{R}^d$. Here M is a manifold. $dim M < d$.

Geometry of manifolds

- Metric (Riemmanian)
- Geodesic (Shortest parth between 2 points)

Isomap

$X \subset \mathbb{R}^d$ our dataset. We want to believe that:

P:[0,t] $\rightarrow M$ paths. Geodesic path is the shortest path.

P:[0,t] $\rightarrow M$ D(p) = $\int_0^t p(t)r(t)dt$. Distance along path

X is actually a manifold with a metric and geodesics in disguise. If we know the distance between $x_1, x_2 \in X$ along a geodesic, then we want to *move* them into $\mathbb{R}^{d_0}(d_0 < d)$ in such a way that we preserve that distance.

C(M) - Continuous functions f:M $\rightarrow \mathbb{R}$ S(M) - Simplicial complex. Generalization of a graph.

1. Construct a graph (V,E) from X. V = X. E = {k-NN on X or ϵ-balls.}
2. Find the shortest graph distance between any 2 data points x_1, x_2 (Dijkstra's Algorithm)
3. "Scale" the data into a smaller $\mathbb{R}^d$ using *multidimensional scaling.*

Topology

- Qualitative
- Connected? : Topology tries to summarize things.
- Summaries
- Topology doesn't care about your coordinate system. It's coordinate-free
- Metric-free

Homotopy

Homology

M attach to it a sequence algebraic structures. Algebraic structures are called homology groups. Each of which contains info about M.

H^0 - Number of connected components, ie, Clusters. H^1 - Number of holes . . . H^n - n^{th} dimension connectivity info.

Persistent Homology

1. Data → simplical complex

metric spaces, ϵ-balls review, exercises

2. Simplical complex → chain complex

ker, Im, quotient space review

3. Homology groups

exercises

4. Persistence

Section 1

Let X be a dataset. What is a simplicial complex?

V = {1,2,3,4}
E = {(1 2),(2 3),(3 4),(4 2)}

In a graph, edges have 2 vertices. In a simplicial complex, a *k-simplex* has k vertices

eg: V = {1,2,3,4} = 0-simplicies
{(1 2),(2 3),(3 4),(4 2)} = 1-simplicies
{(2 3 4)} = 2-simplicies

If a complex has a k-simplex A, then $P(A)$ must be a subset of the complex.

Čech Complex

k-simplicies are defined by the k+1 points whose $\epsilon/2$-balls intersect

eg: 0-simplex {0,1,2}, 1-simplex {(1,2)} Two balls (number 1 and 2) of $\epsilon/2$ radius intersecting and another one (number 3) independent.

Rips Complex

k-simplicies are defined by the chain of k+1 data points within ϵ-distance of each other (often some embedding into $\mathbb{R}^d$)

Lemma: $R_\epsilon \leftrightarrow \epsilon_{\epsilon\sqrt{2}} \leftrightarrow R_{\epsilon\sqrt{2}}$

Manifold Learning

Applications

- Numerous
- Feature Engineering

Implementations

- `scikit-learn`

Simplical complexes

Order matters!! (1 2 3) $\neq$ (1 3 2)

Persistent Homology

1. Data $\rightarrow$ Simplicial Complex
2. Complex $\rightarrow$ Chain
3. Chain $\rightarrow$ Homology
4. Persistence
5. Implementation and Applications

Chain Complex

Let X be a dataset. S denotes the Cech ϵ-complex of X S_k is the set of k-simplicies. eg: S_0 = vertices = data points
S_1 = edges

Definition: A k-chain is a function $f : S_k \rightarrow \mathbb{Q}$, M looking C(M) = $f : M \rightarrow \mathbb{R}$ C_k to denote all k-chains.

Theorem: C_k is a finite dimensional vector space in rational numbers.

example:
$f \in C_2$, $f(1\ 2\ 3) = 1/2$
$f(\sigma) = 1/2$ if $\sigma = (1\ 2\ 3)$, 0 elsewhere

Proof: Let $\sigma \in 1_k$

$$f_\sigma(\delta) = \begin{cases} 1 & \delta = \sigma \\ 0 & \text{elsewhere} \end{cases}$$

$$g \in C_k g = \Sigma_{\sigma \in S_k} g(\sigma) f_\sigma g(\delta) = \Sigma_{\sigma \in S_k} g(\sigma) f_\sigma(\delta) = g(\delta) f_\sigma(\delta) \qquad \delta = \sigma = g(\delta)$$

$n = n(S_k)$, $S_k = \{\sigma_1, \ldots, \sigma_n\}$
$x_1, \ldots, x_n \in \mathbb{Q}$
$f = \Sigma_{i=1}^n x_i f_{\sigma_n} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{Q}^n$

$\mathbb{R}^n$ basis $e_1, \ldots, e_n$
$x_1, \ldots, x_n \in \mathbb{R}$
$V = \Sigma_{i=1}^n x_i e_i = \{x_1, x_2, \ldots, x_n\}$

$\partial_k : C_k \to C_{k-1}$
$\sigma = (v_0, v_1, \ldots, v_k)$
$\partial_k(f_\sigma)(v_0, v_1, \ldots, v_i, \ldots, v_k) = (-1)^i$

Let $[v_0, v_1, \ldots, v_k]$ denote $f_{(v_0,\ldots,v_n)}$
$\sigma = (1\ 2\ 3)$
$\partial_k(f_\sigma)(1\ 2) = 1$
$\partial_k(f_\sigma)(1\ 3) = -1$

$\partial_k[v_0, v_1, \ldots, v_k] = \Sigma_{r=0}^{k-1}(-1)^r[v_0, \ldots, v_r, \ldots, v_k]$

The *chain* in *chain complex*:

$\ldots \to C_k \xrightarrow{\partial_k} C_{k-1} \to \ldots \to C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$

Theorem $\partial_{k-1} \circ \partial_k = 0$
$\partial_{k-1} \circ \partial_k(x) = \partial_{k-1}(\partial_k(x)) = 0 \quad \forall x$

Aside

Let V, W be vector spaces. $T : V \to W$ linear.
$ker(T) := \{v \in V : Tv = 0\}$
$T = [1\ 2\ ,0\ 3]$ $\qquad [x,\ y]$ $\qquad T \cdot [x,\ y] = 0 = [x + 2y,\ 3y]$
ker(T) $= \{[x,\ y] : x = y = 0\}$ $\qquad$ ker(T) $\subset V$

Image(T) = Im(T) = $\{w \in W : \exists v \quad Tv = w\}$. Also called column space.

Quotient Space

$V, W \subset V, V/W, V = W \oplus W^\perp$.

Let $v \in V$
$v = v_0 + w_0, \quad w_0 \in W, \quad v_0 \in W^\perp$
V/W is all $v_0 s$

k-Homology

$H^k(S) = \frac{ker\partial_k}{Im\partial_{k+1}}$

Day 2

TDA

- Persistent Homology
- Funtors
- Hodge theory and ranking
- Mapper
- The End

1.

X dataset. $X \leftrightarrow \mathbb{R}^d$
X = 0-simplex S_0

Simplicial complexes

Two ways

Cech - C_ϵ
$S_k = k+1$ points whose $\epsilon/2$-balls intersect.
Rips - R_ϵ
$S_k = k+1$ points within ϵ-distance pairwise.
Lemma : $R_\epsilon \leftrightarrow C_{\epsilon\sqrt{2}} \leftrightarrow R_{\epsilon\sqrt{2}}$

2.

Definition : k-chain $f : S_k \to \mathbb{Q}$ $(\mathbb{F}_2)$
C_k = set of all k-chains.
$f_\sigma(\delta)$ as defined earlier are basis elements for $C_k \Rightarrow C_k$ is a vector space over $\mathbb{Q}$.
dim $C_k = n(S_k)$

Example: S is a set.
$S_0 = \{1, 2, 3, 4\}$
$S_1 = \{(1\ 2), (2\ 3), (3\ 4), (4\ 2)\}$
$C_0 =< f_1, f_2, f_3, f_4 >_{\mathbb{Q}}$ = all linear combinations of $f_1, f_2, f_3, f_4 = \mathbb{Q}^4$
$C_1 =< f_{(1\ 2)}, f_{(2\ 3)}, f_{(3\ 4)}, f_{(4\ 2)} >\cong \mathbb{Q}^4$
$C_2 = \phi$
$f_1 = [1]$, $f_{(2\ 4)} = [2\ 4]$. Square brackets means the basis elements.
$[2\ 4] = -[4\ 2]$ because $(2\ 4) = -(4\ 2)$. Also, $f_{(4 2)}(2 4) = -f_{(4 2)}(4 2) = -1

k-Homology

$H^k(S) = \frac{ker\partial_k}{Im\partial_{k+1}}$
$H^0 = \frac{ker\partial_0}{Im\partial_1} = \frac{C_0}{Im_{\partial_1}}$
$H^1 = \frac{ker\partial_{d_1}}{Im_{\partial_2}}$

$\beta_k := dim(H^k(S))$ - this is k-th *Betti* number.

Theorem
β_0 = Number of connected components.
β_1 = Number of “holes”.
β_2 = Number of “cavaties”.
.
.
.
β_n = Number of “k-dim holes”. We never go beyond β_2.

3.

Example of computing $H^\star(S)$. $H^\star$ refers to all H^0's.

S = graph with 3 connected vertices 1,2 and 3. $S_0 = \{1, 2, 3\}$.

$S_1 = \{(1\ 2), (2\ 3), (3\ 1)\} \qquad H^0(S) = \frac{ker\partial_0}{Im\partial_1} = \frac{\mathbb{Q}^3}{Im\partial_1}$

$0 \xrightarrow{\partial_2} C_1 \cong \mathbb{Q}^3 \xrightarrow{\partial_1} C_0 \cong \mathbb{Q}^3 \xrightarrow{\partial_0} 0$

$\partial_1[1\ 2] = [2] - [1]$
$\partial_1[2\ 3] = [3] - [2]$
$\partial_1[3\ 1] = [1] - [3]$
$[\partial_1] = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ The dimension of this is 2 because one linear dependent column.
$Im\partial_1 \cong \mathbb{Q}^2 \quad \Rightarrow H^0(S) \cong \frac{\mathbb{Q}^3}{\mathbb{Q}^2} \cong \mathbb{Q}$
$\beta_0 = 1$

Now try to get H^1:
$H^1 = \frac{ker\partial_1}{Im\partial_2}$
$\Rightarrow H^1 = ker\partial_1 \qquad \because Im\partial_2 = \phi$
$\ker \begin{bmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cong \mathbb{Q}$
$\beta_1 = dim H^1(S) = 1$
$H^2 = \frac{ker\partial_2}{Im\partial_3}$

Example of computing a homology group

$_\rightarrow\Delta$ - Vertices of triangle are 2,3 and 4. First one is 1. Find $H^0, \ldots, H^1$.

$$\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$0 \to C_2 \cong \mathbb{Q} \xrightarrow{\partial_2} C_1 \cong \mathbb{Q}^4 \xrightarrow{\partial_1} C_0 \cong \mathbb{Q}^4 \xrightarrow{\partial_0} 0$

$C_2 =< (2\ 3\ 4) >_{\mathbb{Q}}. \qquad (2\ 3\ 4) \mapsto [1].$
$\partial_2[2\ 3\ 4] = [3\ 4] - [2\ 4] + [2\ 3]$
$Im\partial_2 \cong \mathbb{Q} \qquad ker\partial_2 = 0$
$[1\ 2] \mapsto \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

$[\partial_2]_m = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 1 \end{bmatrix}$
$C_1 =< [1\ 2], [2\ 3], [3\ 4], [4\ 2] >_{\mathbb{Q}}$
$\partial_1[1\ 2] = [2] - [1]$
$\partial_1[2\ 3] = [3] - [2]$
$\partial_1[3\ 4] = [4] - [3]$
$\partial_1[4\ 2] = [2] - [4]$
$Im\partial_1 = \mathbb{Q}^3 \qquad ker\partial_1 = \mathbb{Q}$

$H^0 = \frac{\mathbb{Q}^4}{Im\partial_1} = \mathbb{Q}$
$H^1 = \frac{ker\partial_1}{Im\partial_2} = 0$

$H^2 = \frac{ker\partial_2}{Im\partial_3} = 0$
$H^3 = \frac{ker\partial_3}{Im\partial_4} = 0$

4. Persistence

Lemma : $R_\epsilon \leftrightarrow C_{\epsilon\sqrt{2}} \leftrightarrow R_{\epsilon\sqrt{2}}$

Also, $H^\star(R_\epsilon) \xrightarrow{LinComb} H^\star(C_{\epsilon\sqrt{2}}) \leftrightarrow H^\star(R_{\epsilon\sqrt{2}})$ Incresing dimensions in this direction ->

X dataset.
$\epsilon_1, \ldots, \epsilon_n$, $\epsilon_i > 0 \quad \forall i \qquad \epsilon_i < \epsilon_j$ if $i < j$.
$S_i :=$ Simplicial complex $R_{\epsilon_i}(X)$.

$i < j$
$S_i \leftrightarrow S_j$ with X on top of the arrow.
$S_i \to H^\star(S_i) \xrightarrow{X^\star} H^\star(S_j) \leftarrow S_j$

$H^\star(X) =: X^\star$

definition: For $i < j$, the $(i - j)$ persistent homology is:
$H^\star(X)(H^\star(S_i)) = X^\star(H^\star(S_i))$
$H^0(S_i) \xrightarrow{X^\star} H^0(S_j)$
$ImX^\star = X^\star(H^\star(S_i))$

X carries every k-simplex. So then every k-chain, so then every $Im\partial_0, ker\partial_0$, etc.. So then $X^\star : H^\star(S_i) \to H^\star(S_j)$

4. Barcodes

We are interested in basis elements that survive $X^\star = H^\star(X)$.

5. Implementation/Application

C++ : P.H. (Persistent Homology) Alg. Toolbox (2017) - Also has a python binding - does stuff over field of 2 elements $\mathbb{F}_2$.
R: `phom`(2014), `TDA` (2017)

1. **Time Series Classification via TDA** (2017)
   Feature Engineering PH(chaotic timeseries) $\to$ CNN.
2. **PH analysis of protein structure, flexibility & folding** (2014)
   Predicting protein folding
3. **Topology of viral evolution** (2013)
   Virus mutations appear on barcodes
4. **Persistent Homology of Syntax** (2015)
   Language $\mapsto$ All binary features.
   Language $\to$ PCA $\to$ PH.
   They did this process on 2 sets of languages: Indo-European and Niger-Congo

IE: H^0 2 persistent basis elements. 1 if you add Hellenic.
NC: H^0 has 1, H^1 has lots.