Topological Data Analysis

Introduction

Adrien Jamelot

Instituto Superior Técnico

February 22, 2021

Introduction to Data Analysis - The classical Machine Learning Pipeline (1/3)

A bit of Linear Algebra + A bit of Statistics = Machine Learning
Phenomena are described through a set of **features** $F = \{X_1, X_2, \ldots, X_f\}$ and an **outcome** Y
Our objective is to learn how features relate to the outcome.
We need data!

Patient Number	F_1 : *Age*	F_2 = *Size*	F_3 = *Weight*	Y=Maximum velocity
1	10	130	35	15
2	90	160	55	6
...	...	...	...	...
n	30	175	65	38

Table: Caption

Introduction to Data Analysis - The classical Machine Learning Pipeline (2/3)

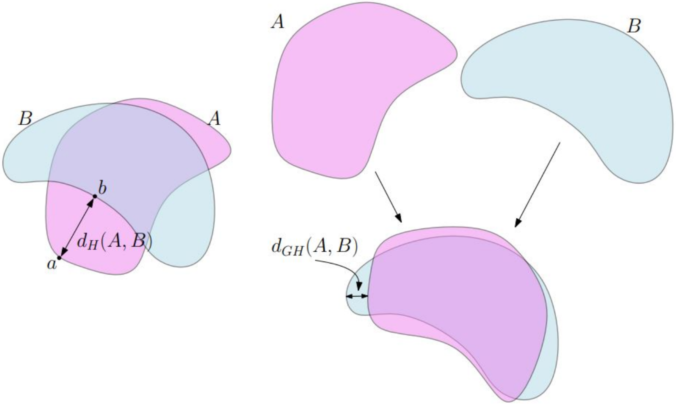
data:$X \in Mat_{n,f}(\mathrm{R})$ Try to extract relevant features (PCA, correlation tests, ...)
Apply and tune a machine learning algorithm (classification: kNN, random forests, SVM/ regression: Linear/Ridge/Lasso $\rightarrow$ Least Squares)

Introduction to Data Analysis - Example: The MNIST dataset (3/3)

TDA - Hausdorff Distance

$\mathrm{d}_H(A, B) = max\left(sup\left\{d(b, A), b \in B\right\}, sup\left\{d(a, B), a \in A\right\}\right)$
A bit better?

$$d_{GH}(M_1, M_2) = inf\{r \geq 0 : \exists(M, \rho), C_1, C_2 \text{ compacts isometric to } M_1, M_2 \text{ s.t } d_H(C_1, C_2) \leq r\}$$



TDA - Simplicial Complexes

Simplicial complexes $\simeq$ Convex Hull of a set of points.
$\sigma = [x_0, \ldots, x_k]$: k-simplicial complex

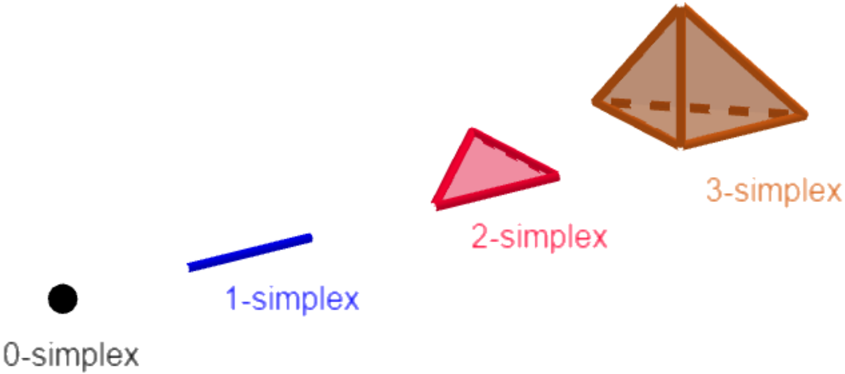


Figure: Some low-dimensional simplicial complexes

TDA - From data to complexes

Vietori-Rips complex —— Cech Complex

$$Rips_\alpha(\mathrm{X}) \subseteq Cech_\alpha(\mathrm{X}) \subseteq Rips_{2\alpha}(\mathrm{X})$$

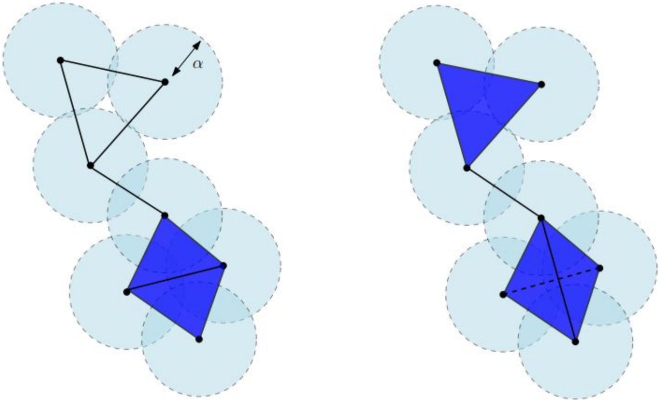


Figure: Left: $Cech_\alpha(X)$, Right: $Rips_{2\alpha}(X)$

TDA - The nerve of a cover

Definition (Nerve of a cover)

$\mathcal{U} = (U_i)_{i \in I}$ cover of X
$C(\mathcal{U}) = \{\sigma = [U_{i_0}, \ldots, U_{i_k}] : \bigcap_{j=0}^{k} U_{ij} \neq \emptyset\}$

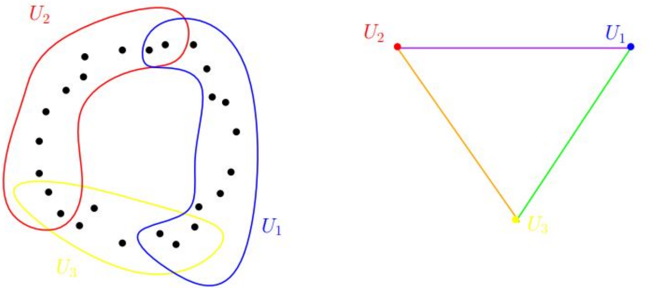


Figure: Nerve of U_1, U_2, U_3

TDA - Homotopy (1/4)

Definition (Homotopic maps)

$f_0, f_1 : X \to Y$ continuous are homotopic if
$\exists H \in C(\mathrm{X} \times [0,1], Y)$ s.t $\forall x \in \mathrm{X}, H(x,0) = f_0(x)$ and $H(x,1) = f_1(x)$

TDA - Homotopy (2/4)

Definition (Homotopy equivalent spaces)

X,Y are homotopy equivalent if $\exists f, g : X \to Y$ s.t $f \circ g$ and $g \circ f$ are homotopic respectively to id_Y and id_X

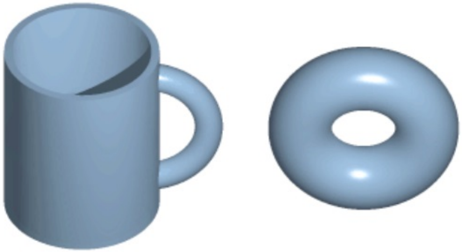


Figure: The surface of a mug and a torus are homotopy equivalent

TDA - Homotopy (3/4)

Definition (Homotopy equivalent spaces)

X,Y are homotopy equivalent if $\exists f : X \to Y, g : Y \to X$ s.t $f \circ g$ and $g \circ f$ are homotopic respectively to id_Y and id_X

Property

X,Y homeomorphic $\Rightarrow$ X,Y homotopy equivalent

TDA - Homotopy (4/4)

Definition (Contractible space)

X is contractible if it is homotopy equivalent to a point

Examples

- Balls
- Convex sets

TDA - Nerve theorem

Definition (Nerve of a cover)

$\mathcal{U} = (U_i)_{i \in I}$ cover of X
$C(\mathcal{U}) = \{\sigma = [U_{i_0}, \dots, U_{i_k}] : \bigcap_{j=0}^{k} U_{ij} \neq \emptyset\}$

Theorem

$\mathcal{U} = (U_i)_{\in I}$ cover of X topological space by open sets such that $\forall J \subset I, \bigcap_{j \in J} U_j$ is either empty or contractible $\Rightarrow$ X is homotopy equivalent to $C(\mathcal{U})$ the nerve of $\mathcal{U}$

Consequence of the Nerve theorem

X set of points, $\mathit{Cech}_\alpha(\mathrm{X})$ is homotopy equivalent to $\bigcup_{x \in \mathrm{X}} B(x, \alpha)$

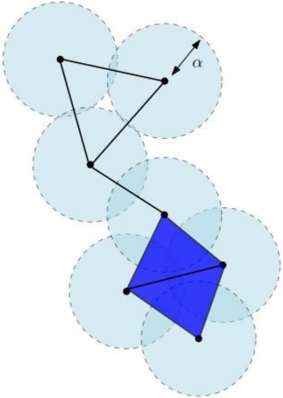


Figure: The simplicial complex preserves the structure of the union of balls

Generalization: notion of filtration

X set of points, $Cech_\alpha(\mathrm{X})$ is homotopy equivalent to $\bigcup_{x\in\mathrm{X}} B(x,\alpha)$

Figure: A binary image

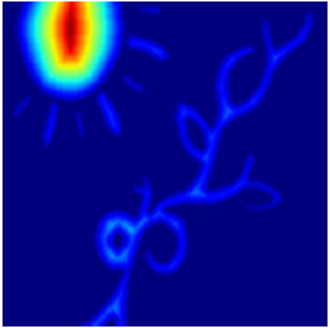


Figure: Its filtration by erosion

Homology - Chains

Definition (k-chain)

$C_k(K) = span\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ where $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ is the set of k-simplices of K

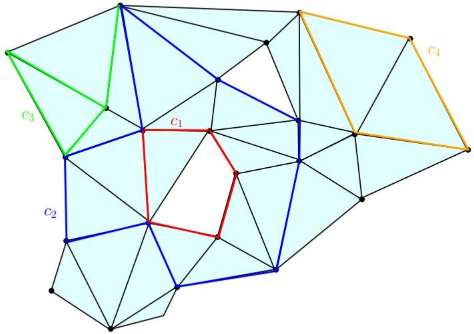


Figure: 0 to 2 dimensional chains

Homology - Boundaries

Definition (Boundary of a k-chain)

$\sigma = [v_0, \ldots, v_k]$ k-simplex

$\partial_k(\sigma) = \sum_{i=0}^{k}(-1)^i[v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k]$

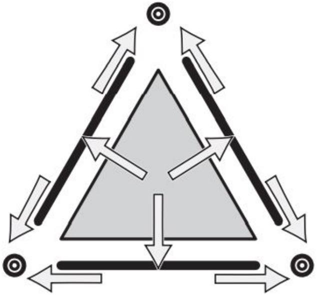


Figure: Boundaries of the triangle, the edge, and the vertex

Homology - Cycles

Definition (k-cycles)

k-chain having boundary 0

Some notations

- Space of k-chains: $C_k(K)$
- Space of k-boundaries: $B_k(K) = Im(\partial_{k+1})$
- Space of k-cycles: $Z_k(K) = Ker(\partial_k)$

Properties

- $\forall k \geq 1, \partial_{k-1} \circ \partial_k = 0$
- $B_k(K) \subseteq Z_k(K) \subseteq C_k(K)$

Homology - k-th simplicial homology group of K

Definition (Homologous cycles)

Two cycles are homologous if they differ by a boundary

Definition (k-th simplicial homology group of K)

Quotient vector space $H_k(K) = Z_k(K)/B_k(K)$. Its elements are the equivalence classes of homologous classes.

Definition (k-th Betti number)

$\beta_k(K) = dim(H_k(K))$

Homology - Chains

Definition (k-chain)

$C_k(K) = span\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ where $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ is the set of k-simplices of K

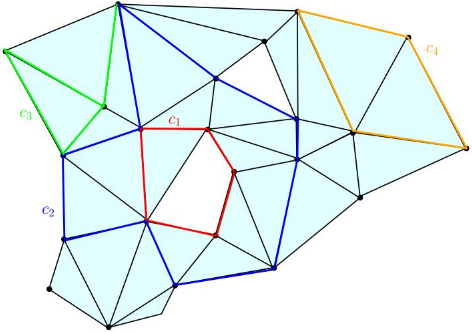


Figure: Finding chains, cycles and boundaries...

TDA - The torus example (1/3)

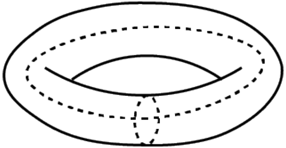


Figure: The torus and its two 1-cycles

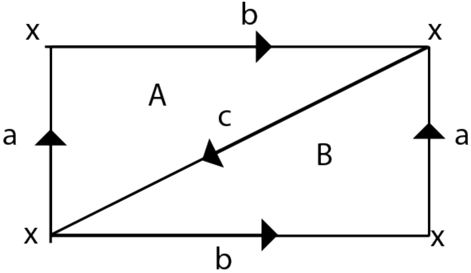


Figure: Simplicial representation of the torus

TDA - The torus example (2/3)

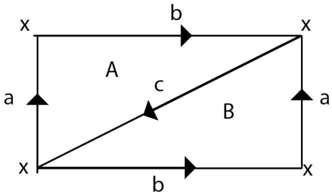


Figure: Simplicial representation of the torus

$$
\begin{aligned}
S &= \{A, B, a, b, c, x_{ul}, x_{ur}, x_{bl}, x_{br}\} \\
&= \{A, B, a, b, c, x\} \\
A &= [x_{ul}, x_{ur}, x_{bl}] \\
B &= [x_{ur}, x_{br}, x_{bl}] \\
\partial_2(A) &= [x_{ur}, x_{bl}] - [x_{ul}, x_{bl}] + [x_{ul}, x_{ur}] \\
&= c - (-a) + b = a + b + c = \partial_2(B)
\end{aligned}
$$

so

$$B_1 = span_{\mathbb{Z}}(a + b + c)$$

TDA - The torus example (3/3)

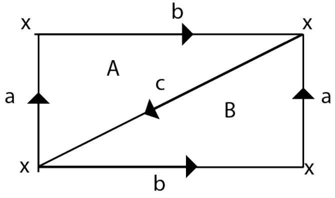


Figure: Simplicial representation of the torus

$$a = [x, x] = b = c$$

so

$$\partial a = x - x = \partial b = \partial c = 0$$

therefore,

$$Z_1 = span_{\mathbb{Z}}(a, b, c) = span_{\mathbb{Z}}(a + b + c, b, c)$$
$$B_1 = span_{\mathbb{Z}}(a + b + c)$$

Now $H_1 = Z_{1/B_1} = \mathbb{Z}^2$
The first Betti number of the torus is 2, in agreement with the figure!

Homology - Invariants

Definition (Euler characteristic)

S simplicial complex
$\chi(S) = \sum(-1)^i k_i$ where k_i is the number of simplexes of dimension i

Examples (Polyhedrons)

Tetrahedron: 4 vertices, 6 edges, 4 faces $\rightarrow \chi = 4 - 6 + 4 = 2$
Cube: 8 vertices, 12 edges, 6 faces $\rightarrow \chi = 8 - 12 + 6 = 2$

TDA - Persistence Homology

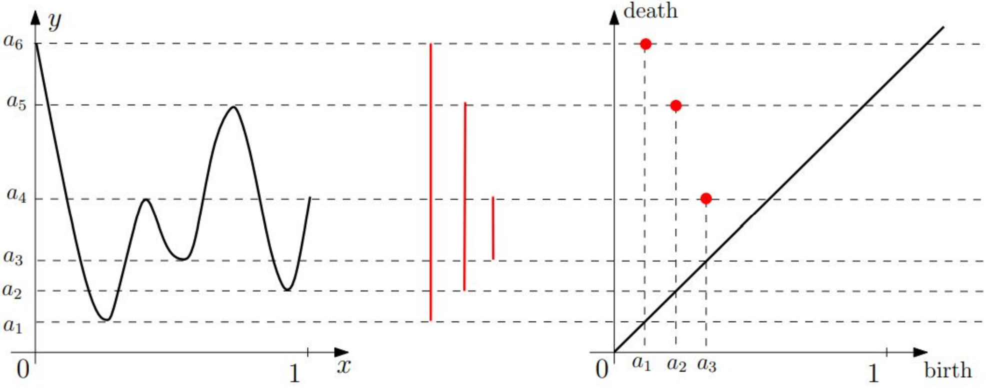


Figure: Peristent Homology for a continuous signal

TDA - Persistence Homology

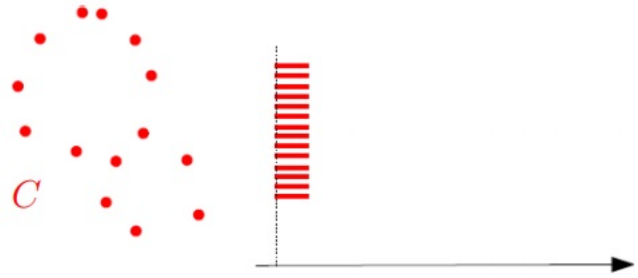


Figure: Step 1

TDA - Persistence Homology

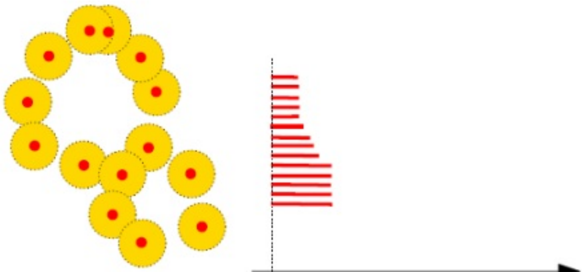


Figure: Step 2

TDA - Persistence Homology

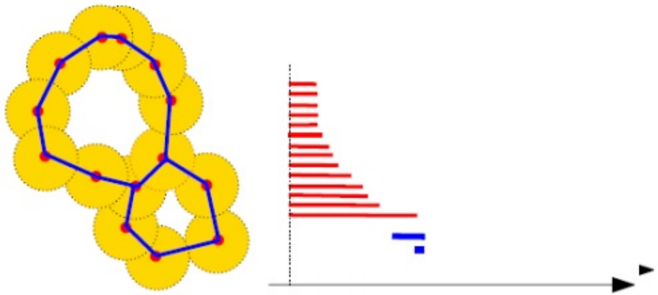


Figure: Step 3

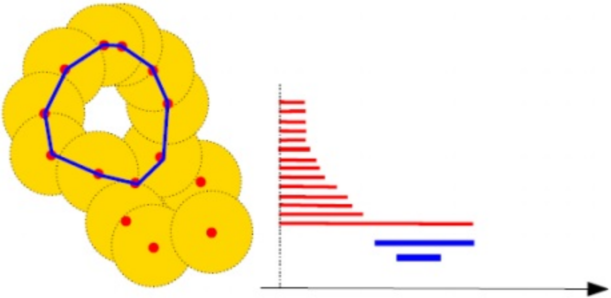


Figure: Step 4

References

Chazal, Frédéric and Bertrand Michel (Oct. 11, 2017). "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists". In:

Garin, Adélie and Guillaume Tauzin (Oct. 22, 2019). "A Topological "Reading" Lesson: Classification of MNIST using TDA". In: *arXiv:1910.08345 [cs, math, stat]*. URL: http://arxiv.org/abs/1910.08345.

Ghrist, Robert (Sept. 1, 2014). *Elementary Applied Topology*. 1st edition.

– (Nov. 15, 2018). "Homological algebra and data". In: *IAS/Park City Mathematics Series*. Vol. 25. American Mathematical Society, pp. 273–325. DOI: 10.1090/pcms/025/06. URL: http://www.ams.org/pcms/025.

Insights into Mathematics (Dec. 10, 2012). *Delta complexes, Betti numbers and torsion — Algebraic Topology — NJ Wildberger*. URL: https://www.youtube.com/watch?v=NgrIPPqYKjQ.

Koplik, Gary (Oct. 29, 2019). *Persistent Homology: A Non-Mathy Introduction with Examples*. Medium. URL: https://towardsdatascience.com/persistent-homology-with-examples-1974d4b9c3d0.