An Introduction and Application of Topological Data Analysis

Olumayowa Olowomeye

Texas A&M University

2019

Topological Data Analysis

Applying topological principles to a set of data. To find the "shape" of the data and analyze it using persistent homology

Simplicial Complexes

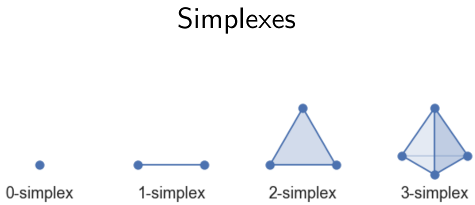


Figure: Source: https://umap-learn.readthedocs.io/en/latest/how$_u$*map*$_w$*orks.html*

- k-simplexes: Nodes, edges, triangles, tetrahedron, so on into higher dimensions

Free Abelian Groups

Definition: An Abelian group is a set with “kind of” addition and additive inverse.
Definition: A free abelian group is an abelian group with “basis.”
For example, if $B = \{b_1, \cdots, b_n\}$ is a set with n elements,

$$\mathbb{Z}B := \{a_1 b_1 + \cdots + a_n b_n : a_i \in \mathbb{Z}, b_i \in B\}.$$

Example: We can assign a free abelian group generated by set of (same dimensional) faces.

Chain complexes and Homomorphism

A chain complex of the simplicial complex is a collection of free abelian group over the set of n-dimensional simplices with the boundary homomorphism between them.
For example, the 2-simplex has 1 face, 3 edges, and 3 vertices. Thus, its chain complex is

$$0 \to \mathbb{Z}F \to \mathbb{Z}e_1 \oplus \mathbb{Z}e_2 \oplus \mathbb{Z}e_3 \to \mathbb{Z}v_1 \oplus \mathbb{Z}v_2 \oplus \mathbb{Z}v_3 \to 0$$

Boundaries

Each simplex is spanned by a lower dimension simplices; call it Boundary operator

$$\delta\triangle^n = \sum_{i=0}^{n} (-1)^n [v_0, ...\hat{v}_i...v_n] \tag{1}$$

$$\delta^2 = \delta(\delta\triangle^n) = 0 \tag{2}$$

An example of boundary operator

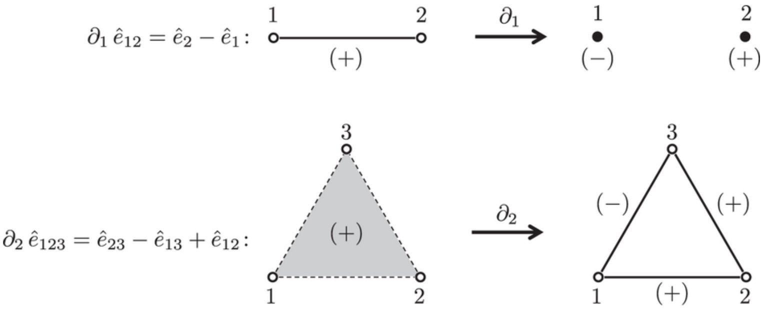


Figure: Source: (Hayden et al,, 2016)

Kernel and Image

Let $\delta : C_k \to C_{k-1}$ be an abelian group homomorphism.
Then, the kernel is defined as $\{x \in C_k : \delta(x) = 0\}$.
An image of δ is defined as $\{$y
$\in C_{k-1} : \delta(x) = y$ for some $x \in C_k\}$.

Concept: Homology and Betti numbers

The Homology group is the $\ker(\delta_k)/Image(\delta_{k+1})$
The Betti number (b_k) is simply the rank of this group

- b_0 refers to the number of connected components
- b_1 refers to the number of holes
- b_2 refers to the number of cavities

Vietoris-Rips complex

Method for creating simplicial complexes

- create a ball of $\epsilon > 0$ around each point
- add a d-simplex at pairwise intersections
- makes a manifold that includes all the points and its shape

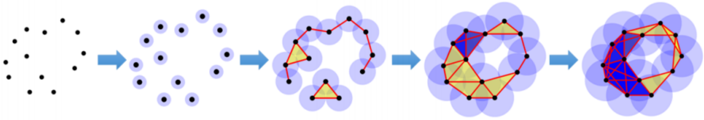
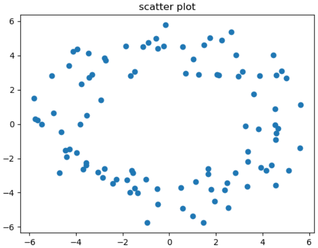


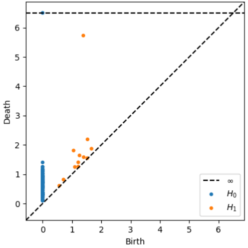
Figure: Source: (Chi et al., 2018)

Persistent Homology

Persistent Homlogy allows us to analyze the Topology of Data sets
As the radius grows what changes...

Example 2d data point in a circular pattern

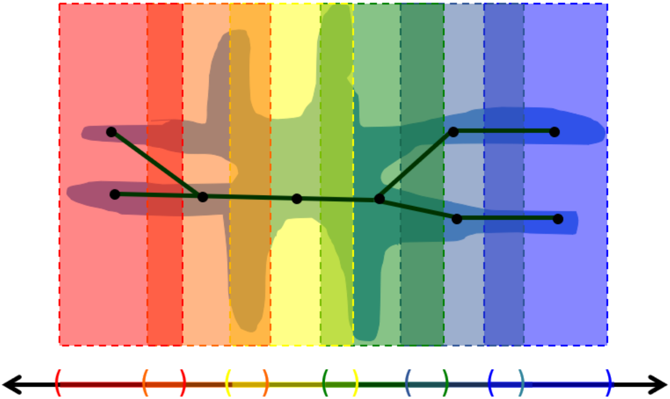




Mapper and Ripser

Ripser: displays resulting Persistent Homology from Vietoris-Rips
Mapper: helps visualize multi-dimensional data and group the data in the simplicial complex generated from Ripser

Mapper process



Source:http://homepage.divms.uiowa.edu/ idarcy/COURSES/TDA/SPRING17/3900mapper2017$_0$1$_2$4.*pptx*

Distances

Cosine distance

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Euclidean distance

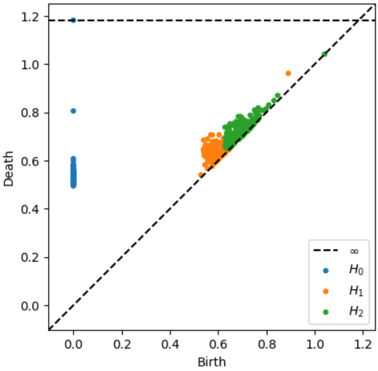
$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Soccer Analysis

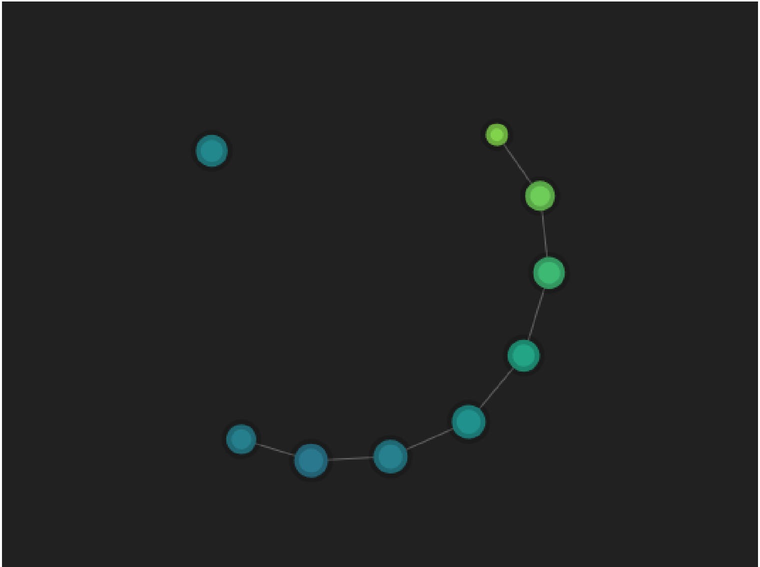
We took data from Sofifa and applied TDA to it
17955 players and 35 values representing their aptitude scores in certain areas such as dribbling, speed,etc.

name	crossing	finishing	heading_accuracy	short_passing	volleys	dribbling	curve
Felipe	34	20	74	62	19	44	49
G. Buffon	13	15	13	37	17	26	20
M. Stekelenburg	18	11	14	39	11	12	13
A. Wilbraham	49	60	75	68	56	46	48
K. Ellison	59	60	61	54	63	65	63
Tarantini	62	64	77	78	69	71	65

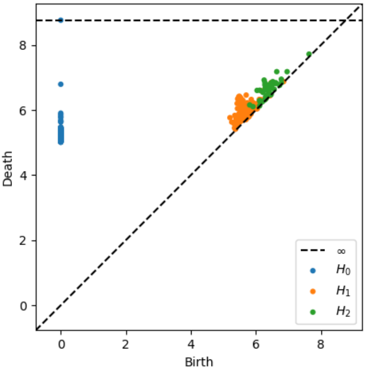
Soccer Analysis: using the cosine distance



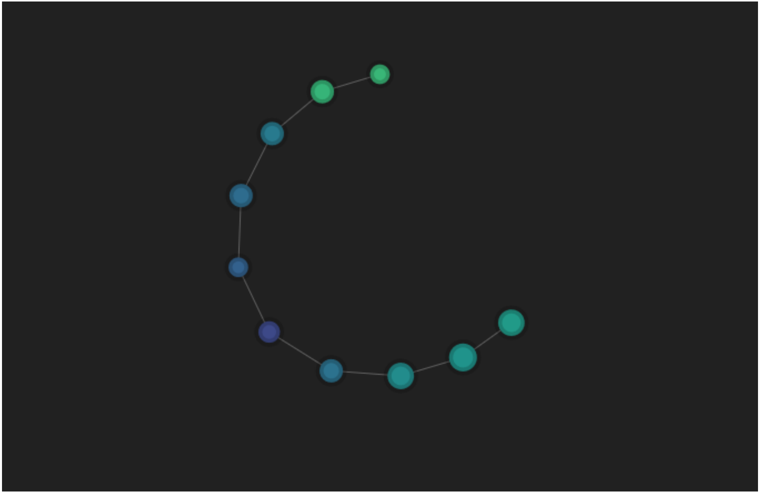
Soccer Analysis: Using the cosine distance



Soccer Analysis: using the Euclidean distance



Soccer Analysis: Using the Euclidean distance



Co-author

Byeoungsu Yu
Texas A&M University

Thank you

Refrences

Saul, Nathaniel and Tralie, Chris. (2019). Scikit-TDA: Topological Data Analysis for Python. Zenodo.
http://doi.org/10.5281/zenodo.2533369

Fraleigh, John B. A First Course in Abstract Algebra. Reading, Mass: Addison-Wesley Pub. Co, 1982. Print.

Adams, Colin Conrad, and Robert Franzosa. Introduction to Topology: Pure and Applied. Dorling Kindersley, 2009.

Chi Seng Pun, Kelin Xia, and Si Xian Lee,
Persistent-Homology-based Machine Learning and its Applications – A Survey, Preprint in Arxiv,
2018

Refrences

Hayden, Patrick and Nezami, Sepehr and Salton, Grant and Sanders, Barry (2016).Spacetime replication of continuous variable quantum information, New Journal of Physics vol 18, doi = 10.1088/1367-2630/18/8/083043