Topological Machine Learning

by

Ankan Kar and Pranava Priyanshu

at

IISER Bhopal Summer Internship 2023

Supervised by

Prof. Kuntal Roy

Introduction

Topological machine learning lies at the intersection of topological data analysis (TDA) and machine learning, leveraging topology to encode the overall shape of data. While geometric features capture localized and rigid properties, topological features excel at capturing multi-scale, global, and intrinsic characteristics of datasets. This capability has been increasingly recognized with the advancement of TDA, and topological information is now widely acknowledged as relevant in data analysis.

The recent expansion of TDA includes the integration of topological techniques to enhance and complement classical machine learning methods and deep learning models. By incorporating topological features, machine learning models can gain a deeper understanding of complex and high-dimensional data, leading to improved performance and insights in various applications. This growing synergy between TDA and machine learning opens up new opportunities for tackling challenging data analysis tasks and enriches the landscape of modern data science.

Indeed, many data analysis algorithms, including those in topological data analysis (TDA), often involve setting multiple parameters. Selecting appropriate parameter values can be challenging, especially when prior domain knowledge is limited. Persistent homology offers a valuable insight by considering information from all parameter values and encoding it into a comprehensible and easily representable form.

The fundamental idea behind persistent homology is to capture the essential features of the data that persist across a wide range of parameter values. By doing so, it avoids the need to make arbitrary choices for specific parameter values and instead provides a more robust and stable representation of the underlying data structure. The information obtained from the persistence of topological features, such as homology groups, allows for a meaningful and interpretable characterization of the data.

In the context of TDA, features that persist for a wide range of parameter values are considered to be more robust and significant, as they are likely to be genuine and meaningful features of the data rather than artifacts of specific parameter choices or noise. This property makes persistent homology a powerful tool for identifying and understanding essential topological features in complex datasets, even in the absence of prior knowledge about the data domain.

Basic Concepts

Some of the concepts that will be widely used are described here:

- **Point Cloud :** It is often defined as a finite set of points in some Euclidean space. But may be taken to be any finite metric space.
- **Nerve Complex :** Nerve complex of a set family is an abstract complex that records the pattern intersections between the sets in the family.
- **Čech Complex :** The Čech complex of a point cloud is the nerve complex of balls of a fixed radius around each point in the cloud.
- **Persistence Module :** A persistence module $\mathbb{U}$ indexed by $\mathbb{Z}$ is a vector space U_t for each $t \in \mathbb{Z}$ and a linear map $u_t^s : U_s \to U_t$ whenever $s \leq t$, such that $u_t^t = 1$ for all t and $u_t^s u_s^r = u_t^r$ whenever $r \leq s \leq t$. An equivalent definition is a functor from $\mathbb{Z}$ considered as a partially ordered set to the category of vector spaces.

- **Persistent Homology Group :** The persistent homology group PH of a point cloud is the persistence module defined as $PH_k(X) = \prod H_k(X_r)$ where X_r is the Čech complex of radius r of the point cloud X and H_k is the homology group.
- **Persistence Barcode :** It is a multiset of intervals of $\mathbb{R}$.
- **Persistence Diagram :** It is multiset of points in $\Delta(:= \{(u, v) \in \mathbb{R}^2 | u, v \geq 0, u \leq v\})$
- **Wasserstein Distance :** It is defined between two persistence diagram X and Y as $W_p[L_q](X, Y) := \underset{\phi : X \to Y}{inf} [\sum_{x \in X} (||x - \phi(x)||_q)^p]^{1/p}$ where $1 \leq p, q \leq \infty$ and ϕ ranges over bijections between X and Y.
- **Bottleneck Distance :** It is the Wasserstein distance when $p = \infty$,i.e, $W_\infty[L_q](X, Y) := \underset{\phi : X \to Y}{inf} \underset{x \in X}{sup}(||x - \phi(x)||_q)$.

Background on Algebraic Topology

In the realm of data analysis, the “manifold hypothesis” posits that data possesses a underlying shape or structure, sampled from an underlying manifold. Rather than solely relying on traditional statistical descriptors, topological data analysis (TDA) takes a distinct approach by exploring this underlying manifold structure from an algebraic standpoint.A central problem in algebraic topology is classification, specifically distinguishing between different manifolds.

To achieve this, algebraic topology employs the tools of mathematics to associate computable algebraic structures, such as groups or vector spaces, with a given manifold, preserving invariance under homeomorphisms. Key algebraic invariants used for this purpose are the homology groups, which effectively capture a wealth of information while remaining efficiently computable in many instances. These homology groups emerge from combinatorial representations of the manifold, referred to as chain complexes.

TDA, building upon the foundations of algebraic topology, leverages the notion of homology groups to unveil intrinsic patterns and features in data sets, enabling a deeper understanding of their underlying structure beyond traditional statistical analyses. By investigating the topological aspects of data, TDA provides valuable insights into complex datasets, especially in scenarios where traditional statistical methods may be limited.

Chain Complexes and Homology

A chain complex is an algebraic structure that consists of a sequence of abelian groups (or modules) and a sequence of homomorphisms between consecutive groups such that the image of each homomorphism is included in the kernel of the next. Associated to a chain complex is its homology, which describes how the images are included in the kernels.

Definition: A chain complex (A_*, d_*) is a sequence of abelian groups or modules ..., $A_0, A_1, A_2, A_3, ...$ connected by homomorphisms (called boundary operators or differentials) $d_n : A_n \to A_{n-1}$, such that the compositions of any two consecutive maps is the zero map, i.e, $d_n \circ d_{n+1} = 0$.

The standard k-simplex Δ^k is defined as the convex hull of the standard basis vectors in $\mathbb{R}^{k+1}$, i.e., $\Delta^k := \{(x_0, ..., x_k) \in \mathbb{R}^{k+1} | \sum_{i=0}^{k} x_i = 1, x_i \geq 0 \forall i\}$. Similarly, a general k-simplex $[v_0, ..., v_k]$ is the convex hull of k+1 affinely independent points $v_0, ..., v_k$ in a Euclidean space. Simplices are the basic building blocks of chain complexes that are used in algebraic topology for the computation of homological invariants. Any topological manifold X can be topologically modelled using simplices.

A singular k-simplex in X is a continuous map $\sigma : \Delta^k \to X$. The inclusion of the i-th face of Δ^k is an important singular simplex in Δ^k, which we will denote by $F_i^k : \Delta^{k-1} \to \Delta^k$. We will be working here in $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ for simplicity.

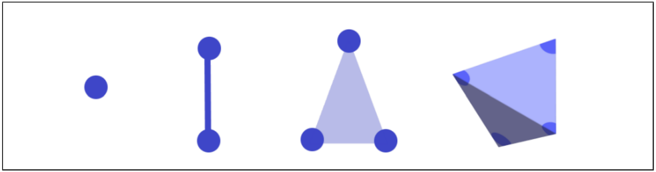


Figure 1: Simplices in increasing dimensions from 0-dimensional to 3-dimensional connected components. From left to right: a 0-dimensional point to a 3-dimensional tetrahedron.

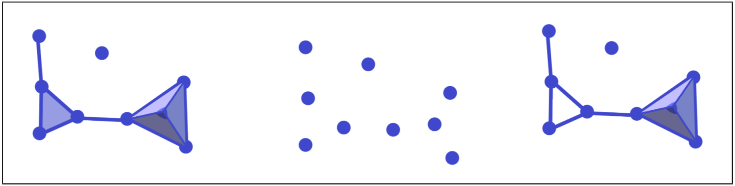


Figure 2: Data in the center and their potential simplicial complex-representations on either side. The left complex consisting of 1 0-dimensional connected component, 2 1-dimensional connected components, 1 2-dimensional connected component and 1 3-dimensional connected component. In the right complex, the 2-dimensional face is not yet fully connected, and so represents a 2-dimensional loop.

Given any space X its singular k-chains are elements of $\mathbb{F}_2$ vector space $C_k(X)$ generated by the set of all singular k-simplices in X. The singular chain complex (C(X), ∂) of X is the sequence of spaces with boundary maps $\partial_k : C_k(X) \to C_{k-1}(X)$ given by $\partial(\sigma) := \sum_i \sigma \circ F_i^k$.

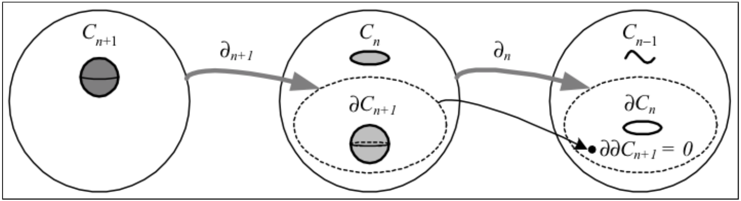


Figure 3: Illustration of Chain Complex with C_n as the n-chains (abelian groups)

Here elements of $Z_k(X) := ker(\partial_k)$ are k-cycles and those of $B_k(X) := im(\partial_{k+1})$ are called k-boundaries and their well defined quotient $H_k(X) := Z_k(X)/B_k(X)$ is the k-th singular homology group of X.

The homology groups are topological invariants, i.e., they remain invariant under homeomorphisms and therefore encode intrinsic information on the topology of X. Thus, homology groups and simpler invariants derived from them, such as the Betti-numbers $\beta_k := dim(H_k(X))$ are useful in studying classification. For example, the 0-th Betti number β_0 is a count of the connected components of a space, while β_1 is a count of the number of cycles.

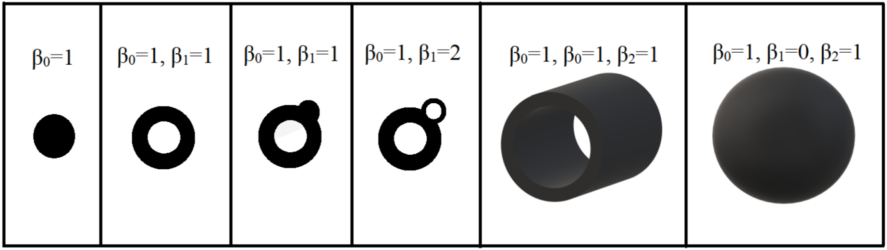


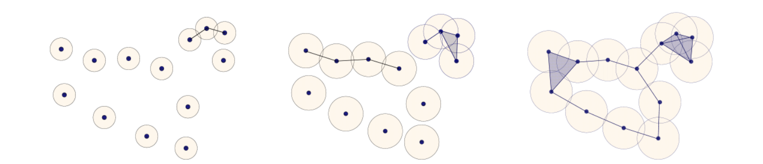
Figure 4: Examples of betti numbers used to characterize topological holes in 0 to 3-dimensional space.

Persistence Homology

Persistent homology is a method for computing topological features of a space at different spatial resolutions. More persistent features are detected over a wide range of spatial scales and are deemed more likely to represent true features of the underlying space rather than artifacts of sampling, noise, or particular choice of parameters.

Persistent homology is an extension of homology to the setting of filtered chain complexes. A filtered chain complex is a (not-necessarily strictly) ascending sequence of chain complexes $C^{\varepsilon_0} \subset C^{\varepsilon_1} \subset C^{\varepsilon_2} \subset C^{\varepsilon_3} \subset ...$ with inclusion maps $\iota^i : C^{\varepsilon_i} \hookrightarrow C^{\varepsilon_{i+1}}$ and $\iota^{i,j} := \iota^j \circ \iota^{j-1} \circ ... \circ \iota^i : C^{\varepsilon_i} \hookrightarrow C^{\varepsilon_j}$ for $i < j$. Filtered chain complexes naturally arise in situations where we have a sequence of inclusions of spaces $X^{\varepsilon_0} \subset X^{\varepsilon_1} \subset X^{\varepsilon_2} \subset X^{\varepsilon_3} \subset ....$

Such cases, for instance, occur if we consider the sublevel sets $X\varepsilon := f^{-1}(R_{<\varepsilon})$ of a so-called filtration function $f : X \to R$, or if we consider a point cloud Y in a metric space (M, d) and set $Y^\varepsilon := \bigcup_{y \in Y} B_\varepsilon(y) = g^{-1}(\mathbb{R}_{<\varepsilon})$ with filtration function $g : M \to \mathbb{R}$ given by $g(m) := \underset{y \in Y}{inf}\{d(m, y)\}$. Here $B_\varepsilon(y)$ denotes the open ball of radius ε centred at y and we implicitly identify $\varepsilon \simeq \varepsilon'$ if X^ε (resp. Y^ε) is canonically homeomorphic to X^δ (resp. Y^δ) for all $\delta \in [\varepsilon, \varepsilon']$.



(a) Rips complex at some scale (b) ..at a larger scale. (c) ...and at an even larger scale.

Figure 5: Example of a Vietoris-Rips filtration on point cloud data

An important property of (singular) homology is that it is functorial, i.e., the inclusion maps $\iota^{i,j}$ induce maps on the respective homology groups $H_k(\iota^{i,j}) : H_k(C^{\varepsilon_i}) \to H_k(C^{\varepsilon_j})$. The figure below depicts the Vietoris–Rips complex construction based on a distance filtration, a standard construction in TDA.

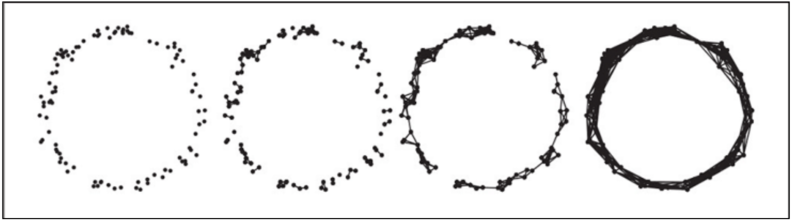


Figure 6: Different stages of a Vietoris–Rips filtration for a simple "circle" point cloud. From left to right, connectivity of the underlying simplicial complex increases as ε increases.

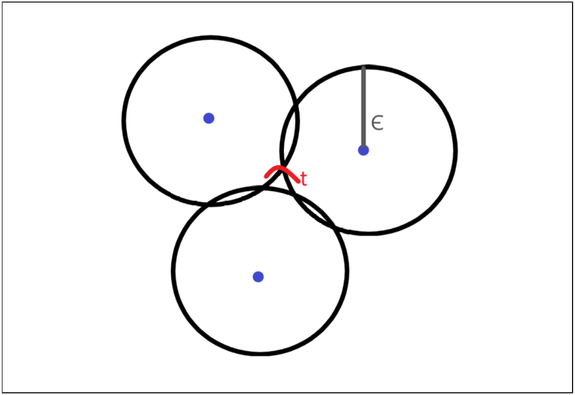


Figure 7: Example of a union-of-spheres at which the Čech and VR filtrations would differ. In this example, if t is sufficiently small, than the VR complex would consider these to be a 3-way connected component, where the Čech filtration would leave them as pairwise connected with a hole between them.

The k-th persistent homology groups are the images of these inclusions, i.e., $H_k^{i,j} := im(H_k(\iota^{i,j})) = Z_k(C^{\varepsilon_i})/(B_k(C^{\varepsilon_i}) \cap Z_k(C^{\varepsilon_i}))$ and thus precisely consist of the k-th homology classes of C^{ε_i} that still exist after taking the inclusion $H_k(\iota^{i,j})$. A homology class $\alpha \in H_k(C^{\varepsilon_i})$ is said to be born at C^{ε_i} if $\alpha \notin H_k^{i-1,i}$.

If α is born at C^{ε_i} then it is said to die at C^{ε_j} if $H_k(\iota^{i,j-1})(\alpha) \notin H_k^{i-1,j-1}$ and $H_k(\iota^{i,j})(\alpha) \in H_k^{i-1,j}$. The persistence of α is given by $\varepsilon_j - \varepsilon_i$ and is set to infinity if it never dies. The persistent Betti-numbers, defined by $\beta_k^{i,j} := dim(H_k^{i,j})$, carry information on how the homology changes across filtration. This information can be captured in a so-called persistence diagram, which is a multiset $\bar{\mathbb{R}}^2 := \mathbb{R} \cup \mathbb{R} \times \{\infty\}$. Specifically, the persistence diagram of (homological) dimension k is given by the points $(\varepsilon_i, \varepsilon_j) \in \bar{\mathbb{R}}^2$ with multiplicity $\mu_k^{i,j} := (\beta_k^{i,j-1} - \beta_k i, j) - (\beta_k^{i-1,j-1} - \beta_k^{i-1,j})$ for all $i < j$. The multiplicity $\mu_k^{i,j}$ counts the number of k-th homology classes that are born at C^{ε_i} and die at C^{ε_j}.

The Figure 8 below is a persistent diagram calculated from the Vietoris–Rips complex in Figure 6. The axes of this diagram correspond to the ε values at which topological features are created and destroyed, respectively. The single point of high persistence corresponds to the primary topological feature of the point cloud, namely its circular shape. Other topological features occur at smaller scales—lower values of ε—and hence form a small dense cluster in the lower-left corner of the persistence diagram. The persistent Betti-numbers can be recovered from the persistence diagram itself.

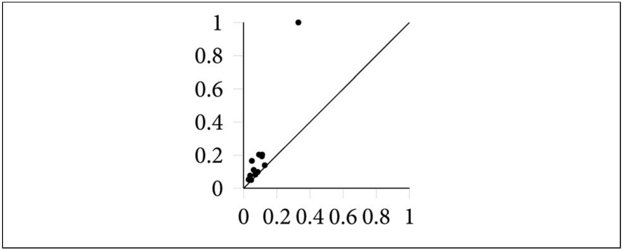


Figure 8: A persistence diagram containing 1-dimensional topological features (cycles)

$\mathbf{d_c^p}$ **Distance :** Let D_x and D_y be two persistence diagrams with cardinalities n and m respectively such that $n \leq m$ and denote $D_x = \{x_1, ..., x_n\}$, $D_y = \{y_1, ..., y_m\}$. Let $c > 0$ and $1 \leq p \leq \infty$ be fixed parameters. The d_c^p distance between two persistence diagrams D_x and D_y is

$$d_c^p(D_x, D_y) = (\frac{1}{m}(\min_{\pi \in \prod_m} \sum_{l=1}^{n} min(c, ||x_l - y_{\pi(l)}||_\infty)^p + c^p|m-n|))^{\frac{1}{p}}$$

where $\prod_m$ is the set of permutations of (1, ..., m).

If $m < n$ then define $d_c^p(D_x, D_y) := d_c^p(D_y, D_x)$.

Various methods for comparing persistence homology exist beyond the traditional persistence diagram comparisons. One widely used approach, offering convenient statistical properties, is the landscape distance. In this method, the (birth, death) coordinates are transformed into a system of orthogonal lambda functions, represented as (m, h) coordinates, to construct what is known as the persistence landscape. This transformation allows for a more manageable representation of the persistence information.

The persistence landscape is demonstrated to form a Hilbert space, a concept from functional analysis, which opens up opportunities for applying powerful mathematical tools. Notably, estimates of the mean in this Hilbert space conform to the central limit theorem, providing statistical reliability in analyzing the data.

Moreover, the stability of the persistence landscape is established, making it resistant to perturbations and noise in the data. This stability property is crucial in ensuring robustness and consistent performance across different data sets and scenarios. Additionally, the landscape distance enables the computation of Z-scores, confidence intervals, and other statistical measures, facilitating thorough and interpretable analyses.

Furthermore, it is demonstrated that the distance between two landscape diagrams serves as a stable lower bound on the Wasserstein bottleneck distance, a measure used to quantify the similarity between probability distributions. This additional insight strengthens the credibility of the landscape distance as a reliable metric for comparing persistence homology.

The landscape distance's various desirable properties, including its stability, simplicity, and statistical applicability, make it a valuable tool for statistical applications involving topological data analysis and persistence homology.

Persistence Landscape

$m = \frac{death+birth}{2}$, $h = \frac{death-birth}{2}$. After persistence points have have been mapped to (m, h) there is a rescaled rank function $\lambda : \mathbb{R}^2 \to \bar{\mathbb{R}}$ defined as

$$\lambda(m,h) = \begin{cases} \beta^{t-m,t+m}, & \text{if } h \geq 0 \\ 0, & otherwise \end{cases} \tag{1}$$

The persistence landscape is then given by $\lambda : \mathbb{N} \times \mathbb{R} \to \bar{\mathbb{R}}$ defined as $\lambda(k,t) = sup(m \geq 0 | \beta^{t-m,t+m} \geq k)$.

The functions of λ can then be grouped into sequences $\lambda_k : \mathbb{R} \to \bar{\mathbb{R}}$. The set of all k is then the persistence landscape. The set of all such λ_k is called the persistence landscape.

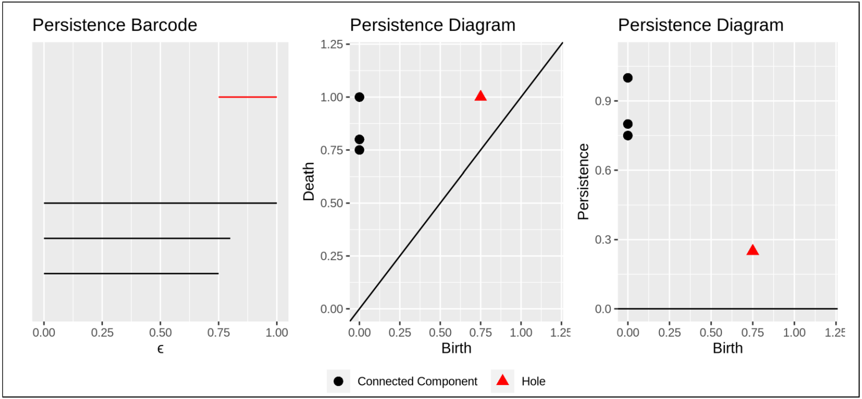


Figure 9: A persistence barcode and two forms of the persistence diagram. These data belong to the simplices in the union-of-spheres example in Figure 7. Notice that the right is simply the rotation of the line birth = death on the left to the x-axis on the right.

An example of the transformation from point cloud, to persistence diagram, to persistence landscape is illustrated in the Figure 10 below.

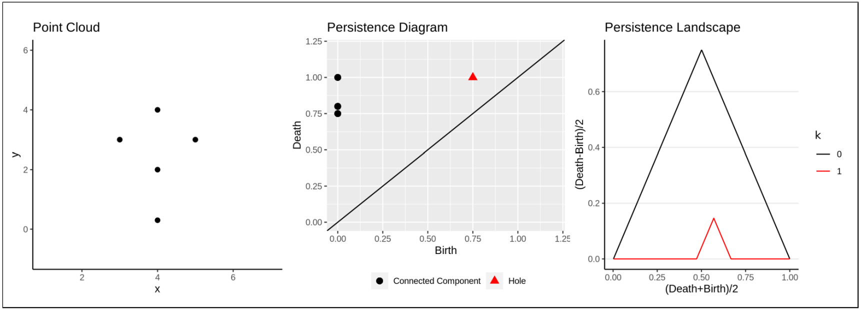


Figure 10: An example of a realization of a persistence landscape from a persistence diagram. Note that λ_0 appears dominated by the VR computation's chosen limit of $\varepsilon \leq 3$ and do to the simplicity of the example.

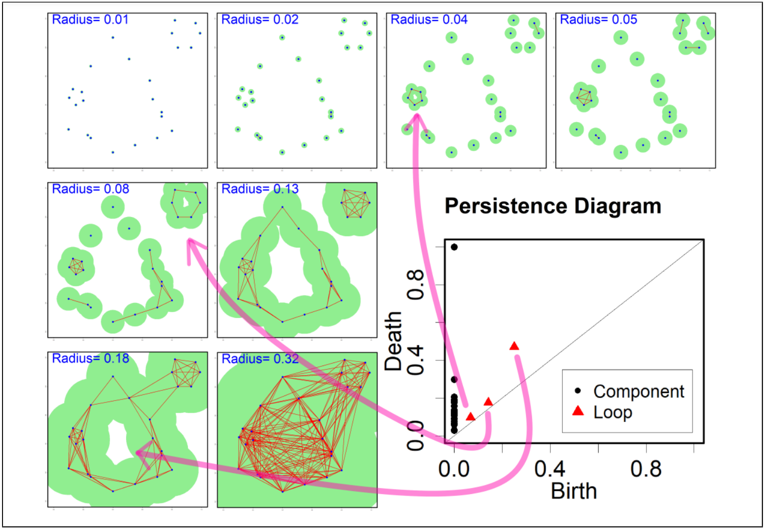


Figure 11: An example of Persistence Diagram Capturing Birth and Death Diameters of Holes

Neural Networks

Neural networks (NNs) have proven to be highly effective in approximating arbitrary, continuous functions on $\mathbb{R}^n$, making them popular for various pattern recognition tasks since the early 2000s. However, two decades ago, researchers faced significant challenges in developing efficient NN formulations due to limited computational power, even for relatively small networks by today's standards.

In the early 1990s, a groundbreaking procedure was introduced for constructing a topologically optimal three-layer NN. This approach draws on foundational concepts from the mathematical literature to estimate the size of the latent dimensions in the data space and constrain the NN to the minimum number of nodes and edges required to accurately model this space. The objective is to find a network with minimal size and complexity, thereby reducing symmetries in the solution space and achieving more efficient approximations. While this procedure is suitable for small networks, its computational feasibility becomes intractable for larger networks and big data sets.

Thus, finding computationally efficient topologies for efficient approximations remains a challenging problem, particularly in modern applications where data size and complexity have grown substantially.

Recently, some studies have delved into leveraging persistent homology to explore topological optimization of networks. The integration of topological methods with NNs offers promising avenues to tackle the complexity and efficiency challenges in modeling data. By harnessing the power of topological data analysis, researchers seek to gain deeper insights into the underlying structure of the data and design more effective and interpretable neural network architectures. These investigations aim to exploit the rich topological information to identify optimal network configurations and enhance the performance of NNs in various application domains.

Persistence Features in Neural Networks

The most commonly used method of incorporating persistence homology is the use of some vectorized form of topological filtration, usually in the form of barcode, landscape or persistence image.

We need to discretize some kind of signal in order to feed the persistence barcode, landscape or diagram to an NN. We have that persistence filtration produce a continuous hierchichal signal which can also be viewed as a point cloud. In order to do this, one must choose a fixed number of bins for the analysis. In the case of barcodes and landscapes, this would be a single choice of width. In the case of persistence landscapes, it would be a two dimensional binning and could therefore have independent choices for width and height of the bins.

In Figure 12 below we can see the demonstratation of several methods for vectorizing persistence representations.

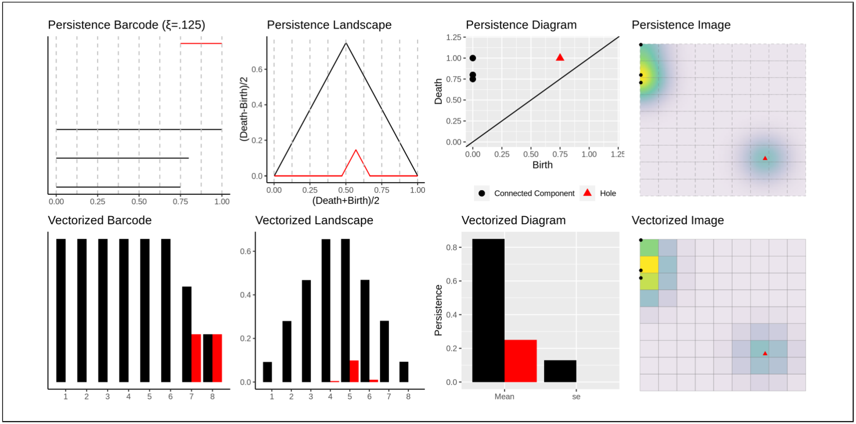


Figure 12: Demonstration of 1 and 2 dimensional binning methods for the vectorization of persistence barcodes, landscapes, diagrams, and images. There are many methods of vectorization, retaining various aspects of the prior persistence information, so these are merely meant to demonstrate several commonly used methods.

After having the vectorized form of persistent information, we can easily use this as input for many algorithms. Support vector machines (SVMs), NNs, clustering algorithms (such as k-means), and others have all been shown effective in application.

In NNs the vectorized input is used as the input to the 0-th layer. If the data are already a single column, such as in the vectorized landscape, this can be input directly. If the vectorized data are greater than 1-dimensional, such as the persistence image, then they must first be unrolled into a single, "flat", column. Also, we can use the signal attribute of some vectorization methods to input the data first to convolutional layers and treat the persistence information as an n-dimensional signal.

Topological Data Analysis in Neural Networks

The integration of Topological Data Analysis (TDA) with Neural Networks (NNs) has led to various applications and strategies in the field of Topological Machine Learning (TML). One common strategy is to incorporate persistence barcode data or persistence diagram data as direct inputs to an NN. These processes may vary widely in applications but a CNN is used primarily to optimize over persistence homologies. A drawback to these studies, from the perspective of a methodologist, is that they make a number of choices in the pre and post processing, bootstrapping, etc. on top of the fact that NN learning is an inherently difficult classifier to understand with a variety of choices in parameterization. This combinatorial array of choices and the complexity of the classifier can make it difficult to generalize results. These methods (with the exception of the vectorized landscape) also rely on a vectorization of persistence information that is known to be unstable, despite its empirically demonstrated success. The vectorization of the barcode and persistence diagram could be problematic, while the vectorized landscape and persistence image have been shown to be stable.

Conclusion

TDA and TML have proven to be valuable tools for analyzing and understanding complex datasets. They provide a fresh lens to explore and interpret the underlying structures of data, uncovering hidden patterns and relationships. By capturing topological features, these approaches offer a more robust and meaningful representation of data, leading to improved machine learning models' performance. As the fields continue to advance, they hold great promise in various scientific, engineering, and real-world applications, contributing to the progress of data analysis and artificial intelligence.

Thank You!!