Topological Data Analysis: extracting insights from the “shape” of data

55e Journées de statistique de la SFdS
27 May 2024
What is topology?
Topology is...

- the mathematics of shape;
Topology is...

- the mathematics of shape;
- the mathematics of connectivity;
Topology is...

- the mathematics of **shape**;
- the mathematics of **connectivity**;
- the mathematics of **emergence of global structure from local constraints**.

D. Benbennick, CC BY-SA 3.0
Application to Data Science

The shape of a data set, described by a topological signature encoding its multi-scale structure, can reveal important relations among the data points, with the help of machine learning.

Topological Data Analysis (TDA)
## Topological analytical tools

<table>
<thead>
<tr>
<th>Method</th>
<th>Appropriate data types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapper</td>
<td>Clinical data, metabolomics, genomics, etc.</td>
</tr>
<tr>
<td>Two-tier Mapper</td>
<td>Gene expression data, single-cell transcriptomics</td>
</tr>
<tr>
<td>Persistent homology</td>
<td>Connectivity data, high-dimensional point cloud data</td>
</tr>
<tr>
<td>Graph signal processing</td>
<td>Connectivity data + “signal”</td>
</tr>
</tbody>
</table>
Mapper
Overview

• (Mostly) **unsupervised** mutivariate pattern analysis of high-dimensional data, retains **more information than PCA**

• Produces a **compressed visual representation** of the data, providing a strong indication of where to look for **meaningful clustering** and encoding relations between clusters

• Numerous remarkably successful applications

• Input:
  • Data set X equipped with notion of “distance” between points
  • Real-valued “measurements” on X
  • Decomposition of the real line into overlapping subsets
  • Choice of clustering algorithm
Mapper output: synthetic data
Mapper output: gene expression data

Mapper output: fMRI data

Persistent homology
The basic persistence workflow

Key notion: Filtration
Step 1: Data to Point Cloud
Step 2: Point cloud to nested complexes

L. Munch, 2019.
Step 2: Point cloud to nested complexes

Radius $r = 0.5$

L. Munch, 2019.
Step 2: Point cloud to nested complexes

L. Munch, 2019.
Step 2: Point cloud to nested complexes

Radius $r = 0.9$
Step 2: Point cloud to nested complexes
Step 2: Point cloud to nested complexes
Step 2: Point cloud to nested complexes

L. Munch, 2019.
Step 3: Nested complexes to barcode

Otter et al., arXiv, 2016.
Barcodes vs persistence diagrams (PD)
Stability

• The set of barcodes/persistence diagrams can be equipped with a variety of earthmover-type distances: the Wasserstein distances of $L_p$-type and the bottleneck distance of $L_\infty$-type.

• Most reasonable known instantiations of the TDA pipeline are Lipschitz continuous with respect to Hausdorff distance on point clouds and bottleneck distance on persistence diagrams.
Practicalities

• There are extensive libraries of software, mostly open source, for TDA computations (e.g., GUDHI, Ripser, Flagser, Giotto-TDA,...).

• There exist “inverse analysis” tools for interpreting results of TDA computations (e.g., work of Hiraoka et al.).
From one to many parameters

• In real data, there are often several parameters along which it would be natural to filter (e.g., some notion of density or time).

• Generalization from one to many parameters poses serious problems, for reasons of both theory and implementation: in general, there is no analogue of barcodes or persistence diagrams.

• Common approaches for two parameters
  • Restrict to lines in the plane determined by the two parameters: fibered barcode.
  • Focus on decompositions into blocks (instead of bars) when possible.
Static TDA input to ML
Strategies for vectorization/featurization

• **Problem:** Cannot compute statistics in the space of barcodes or the space of persistence diagrams.

• **Solution:**
  • Define a Lipschitz-continuous mapping from the space of barcodes/persistence diagrams to a vector space $\mathcal{V}$ equipped with an inner product.
  • Compute statistics in $\mathcal{V}$!
  • Two main types:
    • Embeddings into finite-dimensional Euclidean spaces
    • Kernel methods: defining generalized scalar product on PD, i.e., see PD as elements of a Hilbert space

Few trainable parameters

Expensive
Cavities

By Fashionslide at English Wikipedia, CC BY-SA 4.0
Betti curves

Bar code for cavities of dimension $k$

Betti $k$ curve
Nested complex to Betti curve

Extracting numerical features

Persistence landscapes

- Barcodes also give rise to persistence landscapes.

- The $L_2$-landscape distance between barcodes $B$ and $B'$ with associated landscapes $\lambda$ and $\lambda'$:

$$\Lambda(B, B') = \|\lambda - \lambda'\|_2 = \sum_{k=1}^{\infty} \left(\int |\lambda_k(t) - \lambda'_k(t)|^2 dt\right)^{1/2}$$

Bubenik, JMLR 2015
Dlotko & Bubenik, J Symbolic Comp 2017
Persistence curves

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>$\psi(b, d, t)$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betti</td>
<td>$\beta(D)$</td>
<td>$1$</td>
<td>sum</td>
</tr>
<tr>
<td>Midlife</td>
<td>$ml(D)$</td>
<td>$(b + d)/2$</td>
<td>sum</td>
</tr>
<tr>
<td>Life</td>
<td>$\ell(D)$</td>
<td>$d - b$</td>
<td>sum</td>
</tr>
<tr>
<td>Multiplicative Life</td>
<td>$mul(D)$</td>
<td>$d/b$</td>
<td>sum</td>
</tr>
<tr>
<td>Life Entropy [2]</td>
<td>le($D$)</td>
<td>$-\frac{d - b}{\sum(d-b)} \log \frac{d - b}{\sum(d-b)}$</td>
<td>sum</td>
</tr>
<tr>
<td>Midlife Entropy</td>
<td>mle($D$)</td>
<td>$-\frac{d + b}{\sum(d+b)} \log \frac{d + b}{\sum(d+b)}$</td>
<td>sum</td>
</tr>
<tr>
<td>Mult. Life Entropy</td>
<td>mule($D$)</td>
<td>$-\frac{d/b}{\sum(d/b)} \log \frac{d/b}{\sum(d/b)}$</td>
<td>sum</td>
</tr>
<tr>
<td>$k$-th Landscape [5]</td>
<td>$\lambda_k(D)$</td>
<td>$\min{t - b, d - t}$</td>
<td>$\max_k$</td>
</tr>
</tbody>
</table>

Simultaneous generalization of Betti curves and persistence landscapes. Robust to input noise, efficient to compute, interpretable, and allowing weighting of relative importance of different regions in the PD.

For each $t$, compute $T(\{ \psi(b,d,t) \mid b \leq t, d > t \})$.

Chung and Lawson, arXiv, 2019
Persistence images

• Smooth the PD: replace each point by a Gaussian kernel, then sum
• Discretize

(Image from Kanari, et al., Neuroinformatics, 2018.)
ML methods applied to vectorized TDA

- Decision tree
- Random forest
- Support Vector Machine
- CNN
- GNN

Also possible to integrate a TDA layer into an ML model!
Applications
Idea: Starting at the leaves and descending recursively to the root, decompose the tree into branches, while respecting the Elder Rule, i.e., at any bifurcation, the elder (longer) branch survives and the younger branch is broken off.

Integrate the topology of the tree and the geometry of its embedding in space into a surprisingly powerful global descriptor.

Kanari et al., Neuroinformatics 2017
Kanari et al., Cerebral Cortex 2019
Classification of neuron morphologies

Kanari et al., Neuroinformatics 2017
Kanari et al., Cerebral Cortex 2019
Classification of neuron morphologies

Kanari et al., Neuroinformatics 2017
Kanari et al., Cerebral Cortex 2019
Classification of neuron morphologies

Kanari et al., Cerebral Cortex 2017
Classification of neuron morphologies

Kanari et al., Cerebral Cortex 2017
Classification of microglia

Classification of microglia

Sexual dichotomy in larval fruitflies

Jiao et al, eLife, 2022
Classification of neural dynamics

• For a range of activity parameters, associate to an active Brunel network a weighted graph, to which we apply tools of persistent homology.
• Extract simple topological features of each dynamic regime.
• Use these to train a (highly accurate!) classifier.

Classification of neural dynamics

Automated classification of network dynamics

• For a range of activity parameters, associate to an active Brunel network a weighted graph, to which we apply tools of persistent homology.
• Extract simple topological features of each dynamic regime.
• Use these to train a (highly accurate!) classifier.

Classification of nanoporous crystalline materials

Lee et al., Nature Communications 2017
Lee et al., J Chem Thy Comput 2018
Classification of nanoporous crystalline materials

Lee et al., Nature Communications 2017
Lee et al., J Chem Thy Comput 2018
Detection of gene cascades in single-cell data

- **Idea:** Generate biological hypotheses about closed processes in single-cell RNA seq data using topology and geometry

Maggs, Nguyen, Youssef, in progress
Detection of gene cascades in single-cell data

• Idea: Generate biological hypotheses about closed processes in single-cell RNA seq data using topology and geometry

Maggs, Nguyen, Youssef, in progress
Detection of gene cascades in single-cell data

• **Idea:** Generate biological hypotheses about closed processes in single-cell RNA seq data using topology and geometry.
Detection of gene cascades in single-cell data

- Closed biological process
- Gene expression cascade

Maggs, Nguyen, Youssef, in progress
Detection of gene cascades in single-cell data

Maggs, Nguyen, Youssef, in progress
Detection of gene cascades in single-cell data

Maggs, Nguyen, Youssef, in progress
Detection of gene cascades in single-cell data

Maggs, Nguyen, Youssef, in progress
Detection of gene cascades in single-cell data

Maggs, Nguyen, Youssef, in progress
Thank you to our funding agencies

- SNSF
- Innosuisse
- Blue Brain Project

Members of the Laboratoire de Topologie et Neurosciences