CS 6170: Computational Topology, Spring 2019
Lecture 17

Topological Data Analysis for Data Scientists

Dr. Bei Wang

School of Computing
Scientific Computing and Imaging Institute (SCI)
University of Utah
www.sci.utah.edu/~beiwang
beiwang@sci.utah.edu

March 5, 2019

Stability of Persistence Diagrams: Continued

Edelsbrunner and Harer (2010), C.VIII

Tame functions

- A *triangulation* of a topological space $\mathbb{X}$ is a simplicial complex K together with a homeomorphism between $\mathbb{X}$ and $|K|$, the support of K.
- Let $\mathbb{X}$ be *triangulable* (i.e., if it has a triangulation) and $f : \mathbb{X} \to \mathbb{R}$ continuous.
- Define *sublevel set*

$$\mathbb{X}_a = f^{-1}(-\infty, a],$$

  for $a \in \mathbb{R}$ and for $a \leq b$

$$f_p^{a,b} : \mathsf{H}_p(\mathbb{X}_a) \to \mathsf{H}_p(\mathbb{X}_b).$$

- The *p-th persistent homology group* is defined to be

$$\mathsf{H}_p^{a,b} = \mathrm{im}\ f_p^{a,b}.$$

- The *p-th persistent Betti number* is

$$\beta_p^{a,b} = \mathrm{rank}\ \mathsf{H}_p^{i,j}.$$

Tame functions

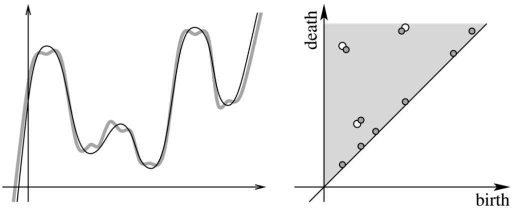
- A a group *isomorphism* is a function between two groups that sets up a one-to-one correspondence between the elements of the groups that respects the given group operations/relations among the elements.
- Greek: *iso* means "equal", and *morphosis* means "to shape".
- $a \in \mathbb{R}$ is a *homological critical value* if there is no $\epsilon > 0$ for which $f_p^{a-\epsilon,a+\epsilon}$ is an isomorphism for each p.
- f is *tame* if it has only finitely many homological critical values and all homology groups of all sub level sets have finite rank.

Bottleneck Stability for Tame Functions

Theorem (Stability Theorem for Filtrations)

Let $\mathbb{X}$ be a triangulable topological space, $f, g : \mathbb{X} \to \mathbb{R}$ two tame functions. For each dimension p, the bottleneck distance between the diagrams $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$ is bounded from above by the L_∞ distance between the functions (Edelsbrunner and Harer, 2010, Page 183), that is,

$$W_\infty(X, Y) \leq ||f - g||_\infty.$$



(Edelsbrunner and Harer, 2010, Page 183)

Degree-q Wasserstein distance

- Given two persistence diagrams X and Y
- The *degree-q Wasserstein distance* is

$$W_q(X, Y) = \left[\inf_{\eta : X \to Y} \sum_{x \in X} ||x - \eta(x)||_\infty^q \right]^{1/q}$$

- Think about assignment problem
- Hungarian algorithm: find a perfect matching (in a bipartite graph) with a minimum total cost
- Software: `https://bitbucket.org/grey_narn/hera`
- Kerber et al. (2016)

Stability with Wasserstein distance

- A function $f : \mathbb{X} \to \mathbb{R}$ is *Lipschitz* if there is a constant C such that
$$|f(x) - f(y)| \leq c||x - y||$$
for all points $x, y \in \mathbb{X}$.
- mesh: max distance between two points in $\sigma \in K$
- $N(r)$: minimum number of simplices whose $\mathrm{mesh} \leq r$.
- A triangulation of $\mathbb{X}$ *grows polynomially* if there are constants c and j such that $N(r) \leq \frac{c}{r^j}$.

Stability Theorem for Lipschitz Functions

Theorem (Stability Theorem for Lipschitz Functions)

Let $f, g : \mathbb{X} \to \mathbb{R}$ be two tame Lipschitz functions on a metric space whose triangulations grow polynomially with constant j. Then there are constants C and $k > j$ no smaller than 1 such that the degree-q Wasserstein distance between $X = \mathrm{Dgm}_p(f)$ and $Y = \mathrm{Dgm}_p(g)$ is

$$W_q(X, Y) \leq C \cdot ||f - g||_\infty^{1-k/q}$$

for every $q \geq k$.

The Assignment Problem

- Given a weighted bipartite graph G with $n + n$ vertices (n vertices on each side), find a *perfect matching* with minimal cost.
- A common cost function is the minimum of the sum of the q-th power of weights of the matching edges for some $q \leq 1$.
- The solution: q-Wasserstein distance
- Kerber et al. (2016): `https://bitbucket.org/grey_narn/hera`
- Bottleneck distance computation: Hopcroft + Karp using k-d tree
- Wasserstein distance computation: Bertsekas using weighted k-d tree

Kernels for barcodes

Inner Product

- Let H be a vector space over $\mathbb{R}$
- A function $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{R}$ is an *inner product* on H if
  - Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_H = \alpha_1 \langle f1, g \rangle_H + \alpha_2 \langle f_2, g \rangle_H$.
  - Symmetric: $\langle f, g \rangle_H = \langle g, h \rangle_H$.
  - $\langle f, f \rangle_H \geq 0$.
  - $\langle f, f \rangle_H = 0$ iff $f = 0$.
- Norm induced by the inner product

$$||f||_H := \sqrt{\langle f, f \rangle_H}.$$

Hilbert space

- Hilbert space: an inner product space that contains a Cauchy sequence.
- Wait a minute...
- A *Hilbert space* is an abstract vector space with the structure of an inner product that allows lengths and angles to be measured.
- A generalizes the notion of Euclidean space.

Kernel

- Given a set X, a function $K : X \times X \to \mathbb{R}$ is a *kernel* if there exists a Hilbert space H called a *feature space* such that

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_H$$

  for all $x, y \in X$.
- Alternatively, K is a kernel if it is symmetric and positive definite.
- A symmetric function $K : X \times X \to \mathbb{R}$ is *positive definite* if $\forall n \geq 1$, $\forall a_1, \cdots, a_n \in \mathbb{R}^n$, $\forall x_1, \cdots, x_n \in X^n$,

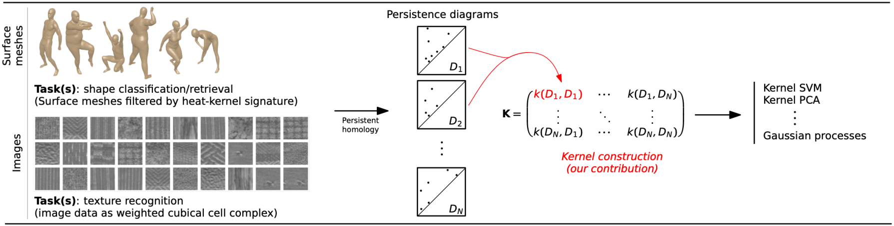
$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0.$$

- Kernels are positive definite
  - Let H be a Hilbert space, X is a nonempty set and $\Phi : \mathbb{X} \to H$, then $K(x, y) := \langle \Phi(x), \Phi(y) \rangle_H$ is positive definite.

TDA Kernels and Vectorizations

- https://github.com/MathieuCarriere/sklearn_tda
- Kernels:
  - Persistence scale space kernel, Reininghaus et al. (2015)
  - Persistence weighted Gaussian kernel
  - Sliced Wasserstein kernel
  - Persistence Fisher kernel
- Vectorizations:
  - Persistence Image, Adams et al. (2017)
  - Persistence landscape
  - Betti Curve
  - Silhouette

TDA Kernels in applications



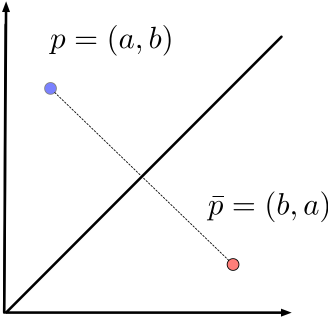
Reininghaus et al. (2015)

Persistence scale space kernel

- Let F, G be two persistence diagram (of a fixed dimension p)
- The *persistence scale space kernel* is

$$K_\sigma(F,G) = \frac{1}{8\pi\sigma} \sum_{p\in F, q\in G} \left(e^{-\frac{||p-q||^2}{8\sigma}} - e^{-\frac{||p-\bar{q}||^2}{8\sigma}} \right)$$

- $\bar{p}$ is p mirrored at the diagonal.



Persistence scale space kernel

- $\mathcal{D}$: set of persistence diagrams
- Parameter: σ
- $L_2(\Omega)$: set of L_2 functions (square integrable) on $\Omega \subset \mathbb{R}^2$
- Feature map: $\Phi_\sigma : \mathcal{D} \to L_2(\Omega)$
- $K_\sigma(F, G) = \langle \Phi_\sigma(F), \Phi_\sigma(G) \rangle_{L_2(\Omega)}$
- Stability of the persistence scale space kernel:

$$||\Phi_\sigma(F) - \Phi_\sigma(G)||_{L_2(\Omega)} \leq \frac{1}{\sigma\sqrt{8\pi}} W_1(F, G)$$

- $W_1(F, G)$: degree-1 Wasserstein distance.

References I

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252.

Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, USA.

Kerber, M., Morozov, D., and Nigmetov, A. (2016). Geometry helps to compare persistence diagrams. *Proceedings of the Workshop on Algorithm Engineering and Experiments*.

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*.