Fundação Universidade Federal do ABC
Pró reitoria de pesquisa

Projeto de Iniciação Científica submetido para avaliação
Edital Nº 4/2022 - PROPES (11.01.07).

Título do projeto: Estrutura em Dados

Palavras-chave do projeto: TDA; homology; data analysis; persistence

Área do conhecimento do projeto: topologia algébrica e aplicações.

Sumário

1 Resumo

O projeto visa introduzir o estudante à teoria da topologia algébrica em relação com técnicas atuais de análise de dados. Para isto, será desenvolvida a teoria e prática sobre complexos simpliciais, homologia e persistência associada a uma filtração.

2 Introdução

A base empírica da ciência começa no ato de recoleição e sistematização de dados finitos. As caraterísticas particulares dos eventos registrados dão uma noção de semelhança ou dissimilitude entre pares de ocorrências. Assim, desde um ponto de vista matemático, o resultado dessa prática pode ser modelado como um espaço métrico finito (X, d), i.e., um conjunto com uma função de distância entre pares de elementos. Porém, com frequência esses dados são considerados como meros indícios de estruturas subjacentes às quais só acedemos por um processo de amostragem, estruturas que tipicamente são concebidas como contínuas. Aqui a metáfora visual é a de procurar entender uma figura concreta do espaço a partir de alguns pontos conhecidos. Essa figura pode ser um grafo, uma variedade, ou mais geralmente um complexo simplicial (finito). A vantagem destes últimos é que, enquanto objetos puramente combinatórios, admitem tratamento computacional.

Um caso de interesse que iremos considerar é o do complexo de Vietoris-Rips associado a um espaço métrico (X, d) dado, denotado por $\mathcal{VR}_\epsilon(X)$, aonde $\epsilon > 0$ é um parâmetro de escala. Uma vez que (X, d) pode ser pensado como um grafo métrico, $\mathcal{VR}_\epsilon(X)$ é definido como o complexo de bandeira, "flag complex", associado ao grafo filtrado segundo o nível $\epsilon > 0$ dado.

Em uma segunda etapa de conceitualização e resumo, resulta crítico definir caraterísticas que iremos observar pela sua relevância ao problema. É aqui aonde optamos por departir do ponto de vista puramente geométrico e adotar alguns dos invariantes discretos que fornece a topologia algébrica. Concretamente iremos considerar os números de Betti, i.e., a dimensões dos espaços da homologia com coeficientes em um corpo escolhido. As razões são múltiplas mas, principalmente, podemos destacar a invariância respeito a deformações continuas, a invariância respeito à escolha de outra estrutura simplicial (quando o complexo está associado a uma variedade subjacente), e a possibilidade de identificar estruturas em qualquer dimensão $n \in \mathbb{N}_0$. Este último ponto resulta crítico quando os dados que estão sendo analisados/reconstruídos são de natureza mais abstrata, e permite trabalhar com dados em áreas tão diversas como finanças, neurociências ou diagnostico clínico, veja [1, 3].

Finalmente, um ponto técnico de interesse é que o parâmetro de escala $\epsilon > 0$ anteri-

ormente mencionado não será estimado nem fixado por considerações auxiliares. A ideia é trabalhar com o espaço (X, d) sem mais escolhas, considerando assim não apenas um complexo mas a família a um parâmetro de complexos: $\{\mathcal{VR}_\epsilon(X)\}_\epsilon$. No nível da homologia, teremos então um módulo de persistência (quiver linear) dado pelos espaços de n-homologia, para cada $n \in \mathbb{N}_0$ fixo, variando (discretamente) o parámetro $\epsilon > 0$. Tal módulo é caraterizado completamente pelo raconto dos eventos de "nascimento"e "morte"de uma classe de homologia que aporte dimensão, ou seja, que contribua ao n-número de Betti considerado. Esta construção é um luxo computacional introduzido em [2] que, desde o ponto de vista teórico é possível graças à estrutura linear da persistência -veja o clássico teorema de Gabriel sobre representações de quivers.

Em resumo temos o seguinte processo: para um valor $\epsilon > 0$ e *input* de dados (X, d) obtemos o complexo (simplicial) de Vietoris-Rips, $\mathcal{VR}_\epsilon(X)$. Sobre este complexo computamos a n-homologia com coeficientes em um corpo $\mathbb{K}$ fixo, $\mathsf{H_n}(\mathcal{VR}_\epsilon(X))$. Iteramos o procedimento para incrementos de ϵ. Ao todo, no nível da homologia temos construído um módulo de persistência. Este objeto é completamente caraterizado por um código de barras, "barcode", uma representação gráfica conveniente dos tempos de nascimento e morte antes descritos. Veja o diagrama seguinte:

$$\left\{ (X,d) \xrightarrow{\overset{itera\ \epsilon}{\longleftarrow}} \mathcal{VR}_\epsilon(X) \longrightarrow \mathsf{H_n}(\mathcal{VR}_\epsilon(X)) \right\} \implies \textbf{barcode}$$

Cabe advertir que as possibilidades em modelagem de dados são bastas e, em essência, uma questão epistemológica profunda. O procedimento aqui apresentado constitui o núcleo do que se conhece como "homologia persistente", uma estrategia matemático-computacional para dar conta de dados provenientes de diversas origens. Esta é a principal ferramenta dentro da análise topológica de dados (TDA), e forma parte das respostas contemporâneas ao advimento do "big data" e a consequente demanda por processamento, veja [5].

3 Objetivos

De modo geral, o objetivo consiste em contribuir para a formação do estudante ao inseri-lo em uma linha ativa de pesquisa em matemática com aplicações concretas à industria. Especificamente, trata-se de introduzir a teoria de complexos simpliciais e a sua homologia seguida da noção de persistência até chegar às implementações computacionais. Como produto final do projeto o estudante poderá optar por redatar uma resenha expositiva, uma demostração/estudo de caso computacional, ou uma forma mista destas.

Elencamos alguns pontos que destacam-se como objetivos:

- despertar a vocação de pesquisa em geometria-topologia a partir de problemas concretos;
- introduzir conceitos avançados de nível mestrado-doutorado ainda na graduação facilitando uma potencial opção por uma pós-graduação em matemática ou computação;
- vincular a formação básica do estudante com um área estratégica de alta demanda, como ser a análise de dados;

4 Metodologia

O núcleo do projeto é essencialmente teórico e será desenvolvido mediante o estudo de referências reconhecidas, discussões e apresentação de seminários internos ao grupo.

Como introdução à área e disparador de perguntas dispomos dos *surveys* de R. Ghrist [3] e G. Carlsson [1]. Para o estudo da teoria básica de homologia via complexos simpliciais, adotamos o livro de P. J. Giblin em [4]. Como referência das aplicações via persistência, com escopo além deste projeto, temos o livro de S.Y. Oudot [6].

5 Viabilidade

Não há restrições ao desenvolvimento de um projeto teórico. As componentes práticas do final do projeto são de caráter computacional e requerem apenas o uso de software livre. Desse modo, deverão bastar os recursos computacionais privados e aqueles disponibilizados pela universidade.

O projeto poderá ser desenvolvido a distância se isso for requerido pelos participantes ou pelas circunstâncias.

6 Cronograma de atividades

- Etapa 1: De dados a complexos
  - Etapa 1.a. Problematização e leitura parcial dos surveys [1, 3].
  - Etapa 1.b. Estudo de complexos simpliciais. Cap. 3 de [4].
  - Etapa 1.c. Continuação Cap. 3.

- Etapa 2 Topologia algébrica
  - Etapa 2.a. Estudo de homologia n-dimensional com coeficientes. em um corpo. Cap.4 de [4].
  - Etapa 2.b. Continuação Cap. 4.
  - Etapa 2.c. Persistência.
- Etapa 3 Prática
  - Etapa 3.a. Implementações em Python.
  - Etapa 3.b. Elaboração do trabalho final.
  - Etapa 3.c. Elaboração da apresentação.

Etapa	Mês											
	01	02	03	04	05	06	07	08	09	10	11	12
1.a	x	x	x									
1.b		x	x	x								
1.c			x	x	x							
2.a				x	x	x						
2.b					x	x	x					
2.c						x	x	x				
3.a							x	x	x			
3.b								x	x	x		
3.c									x	x	x	x

Referências

[1] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[2] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. volume 28, pages 511–533. 2002. Discrete and computational geometry and graph drawing (Columbia, SC, 2001).

[3] Robert Ghrist. Homological algebra and data. In *The mathematics of data*, volume 25 of *IAS/Park City Math. Ser.*, pages 273–325. Amer. Math. Soc., Providence, RI, 2018.

[4] P. J. Giblin. *Graphs, surfaces and homology.* Chapman and Hall Mathematics Series. Chapman and Hall, London; Halsted Press [John Wiley & Sons, Inc.], New York, 1977. An introduction to algebraic topology.

[5] Jürgen Jost. Object oriented models vs. data analysis—is this the right alternative? In *Mathematics as a tool*, volume 327 of *Boston Stud. Philos. Hist. Sci.*, pages 253–286. Springer, Cham, 2017.

[6] Steve Y. Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2015.