

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Geradores de homologia persistente e aplicações

Carlos Henrique Venturi Ronchi

Dissertação de Mestrado do Programa de Pós-Graduação em
Matemática (PPG-Mat)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Carlos Henrique Venturi Ronchi

Geradores de homologia persistente e aplicações

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Matemática. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática

Orientador: Prof. Dr. Marcio Fuzeto Gameiro

USP – São Carlos
Junho de 2018

Carlos Henrique Venturi Ronchi

Persistent homology generators and applications

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Mathematics Graduate Program, for the degree of Master in Science.
EXAMINATION BOARD PRESENTATION COPY

Concentration Area: Mathematics

Advisor: Prof. Dr. Marcio Fuzeto Gameiro

USP – São Carlos
June 2018

RESUMO

RONCHI, C. H. V. **Geradores de homologia persistente e aplicações**. 2018. 33 p. Dissertação (Mestrado em Ciências – Matemática) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

a.

Palavras-chave: Modelo, Monografia de qualificação, Dissertação, Tese, Latex.

ABSTRACT

RONCHI, C. H. V. **Persistent homology generators and applications**. 2018. [33](#) p.
Dissertação (Mestrado em Ciências – Matemática) – Instituto de Ciências Matemáticas e
de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

a.

Keywords: Template, Qualification monograph, Dissertation, Thesis, Latex.

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação do pipeline para a utilização da homologia persistente com um conjunto de dados.	22
Figura 2 – Exemplos de k -simplexos para $k \in \{0, 1, 2, 3\}$	23
Figura 3 – Exemplo em que a interseção de dois simplexos não é um simplexo. . .	24
Figura 4 – Exemplo de filtração para um complexo simplicial K	24
Figura 5 – Esquema de uma rede neural artificial. O número de vértices na camada escondida é determinado pelo tamanho da matriz A_i	30

LISTA DE ALGORITMOS

LISTA DE CÓDIGOS-FONTE

LISTA DE TABELAS

SUMÁRIO

1	INTRODUÇÃO	19
2	HOMOLOGIA PERSISTENTE 101	21
2.1	Filtrações e espaços vetoriais	22
2.2	A matriz de bordo ∂	24
2.3	Redução da matriz	24
3	MÓDULOS DE PERSISTÊNCIA	25
4	GERADORES ÓTIMOS E OUTROS CONCEITOS	27
4.1	Geradores ótimos	27
4.2	Vetorização do diagrama de persistência	27
4.3	Mapper	27
5	APLICAÇÕES	29
5.1	Geradores ótimos em classificadores de imagens	29
5.1.1	<i>Redes Neurais Convolucionais (CNN)</i>	29
5.2	Imagens de persistência aplicadas a proteínas	30
6	CONCLUSÃO	31
	REFERÊNCIAS	33

INTRODUÇÃO

HOMOLOGIA PERSISTENTE 101

A topologia sempre foi vista como uma área de abstração da matemática, sem espaço para aplicações. Ela é usada para o estudo de diversos espaços em sua forma abstrata, auxiliando matemáticos em diversas demonstrações de teoremas e dando uma base fundamental para grande parte da teoria matemática usada no dia a dia ([POINCARÉ, 1895](#)).

Certas propriedades dos espaços topológicos são estudadas através da topologia algébrica, dando algumas informações, como o número de componentes conexas por caminhos de um espaço e buracos. A princípio esta é uma área altamente abstrata da matemática, nos últimos anos esta visão foi mudando, com o desenvolvimento da Homologia Persistente e Análise Topológica de Dados.

Um conjunto de dados, geralmente um subconjunto finito de algum espaço métrico, pode ser estudado através da homologia persistente e assim obtemos informações topológicas do objeto em estudo.

O pipeline da análise topológica de dados pode ser dividido nos seguintes passos:

- A entrada do algoritmo pode ser um conjunto de pontos ou alguma matriz de distância/similaridade do conjunto de dados.
- A construção de um objeto combinatorial em cima do conjunto de dados ou da matriz de distância. Geralmente uma filtração ou um complexo simplicial.
- A partir da filtração ou do complexo simplicial é possível extrair informações topológicas e geométricas do conjunto de dados, por exemplo o número de componentes conexas, como um algoritmo de Clustering.
- Por fim a interpretação dos dados obtidos e possível pós processamento para a utilização em outros algoritmos, como os de classificação ou regressão.

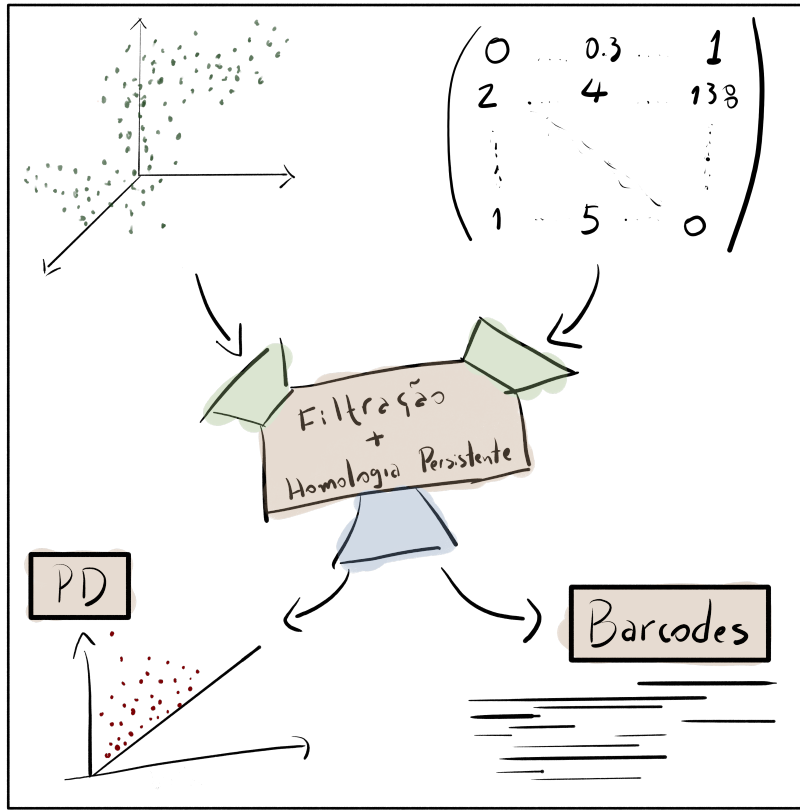


Figura 1 – Representação do pipeline para a utilização da homologia persistente com um conjunto de dados.

Neste capítulo descrevemos de forma ingênua a homologia persistente, começando com filtrações, passando pelos espaços vetoriais associados aos complexos simpliciais e chegando ao algoritmo de homologia persistente. Mostraremos também como interpretar os resultados obtidos. A [Figura 1](#) mostra os passos para utilizar esta ferramenta em um conjunto de dados.

2.1 Filtrações e espaços vetoriais

A filtração de um conjunto de dados é o primeiro passo na nossa sequência apresentada na [Figura 1](#). Dado um conjunto de dados precisamos construir um objeto combinatorial de forma que possa ser analisado do ponto de vista da topologia assim como computacionalmente. A filtração é este objeto que captura as mudanças do conjunto dada uma escala.

Algumas definições se fazem necessárias para entendermos o que é a filtração e qual o seu papel na análise topológica de dados. Começamos definindo um simplexo, primeiro objeto combinatorial que é a base da filtração.

Definição 2.1.1. Sejam $v_0, v_1, \dots, v_k \in \mathbb{R}^n$ linearmente afins, ou seja $\{v_1 - v_0, \dots, v_k - v_0\}$ é

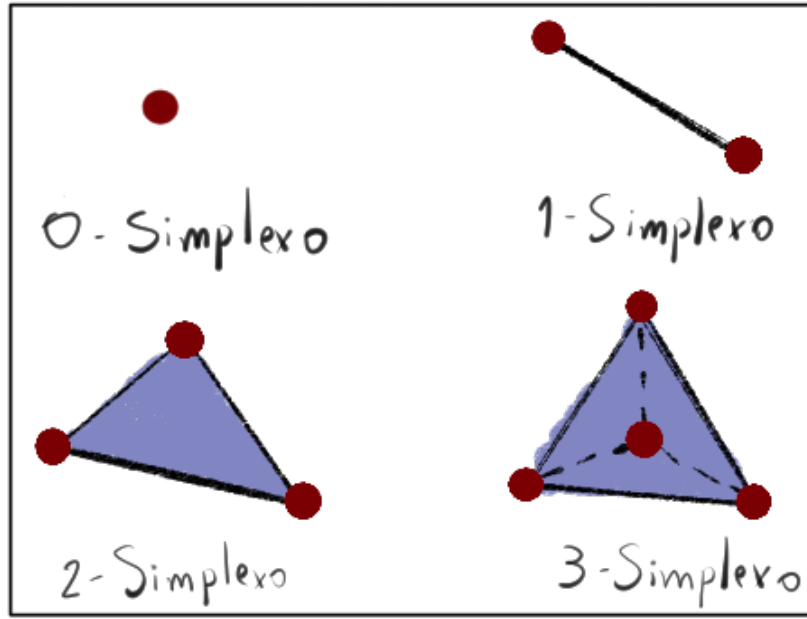


Figura 2 – Exemplos de k -simplexos para $k \in \{0, 1, 2, 3\}$.

um conjunto linearmente independente. O k -simplexo definido pelos pontos acima definidos, chamados de vértices também, é o conjunto

$$\left\{ \sum_{i=0}^k \lambda_i v_i \mid \sum_{i=0}^k \lambda_i = 1 \text{ e } \lambda_i \geq 0, \forall i \right\}.$$

Note que para $k = 0$, temos um único vértice. Para $k = 1$, temos uma reta, já para $k = 2$ temos um triângulo preenchido. E no caso $k = 3$, um tetraedro. Os simplexos podem ser vistos na [Figura 2](#). Além disso, dizemos que a dimensão do k -simplexo é k . A envoltoria convexa de qualquer subconjunto dos vértices de um simplexo S é chamado de face de S .

Tendo definido os k -simplexos, podemos definir o complexo simplicial.

Definição 2.1.2. Um complexo simplicial K é uma coleção de simplexos satisfazendo as seguintes relações:

- Dado $\sigma \in K$, temos que para toda face $\tau \subset \sigma$ vale $\tau \in K$.
- A interseção de dois simplexos é face de ambos os simplexos, em outras palavras, $\sigma, \tau \in K$ implica que $\sigma \cap \tau \subset \sigma$ e $\sigma \cap \tau \subset \tau$.

A segunda condição é necessária para evitar casos patológicos como mostrado na [Figura 3](#)

Dizemos que a dimensão do complexo simplicial K é a maior dimensão dentre os simplexos em K . Podemos definir agora a filtração de um complexo simplicial.

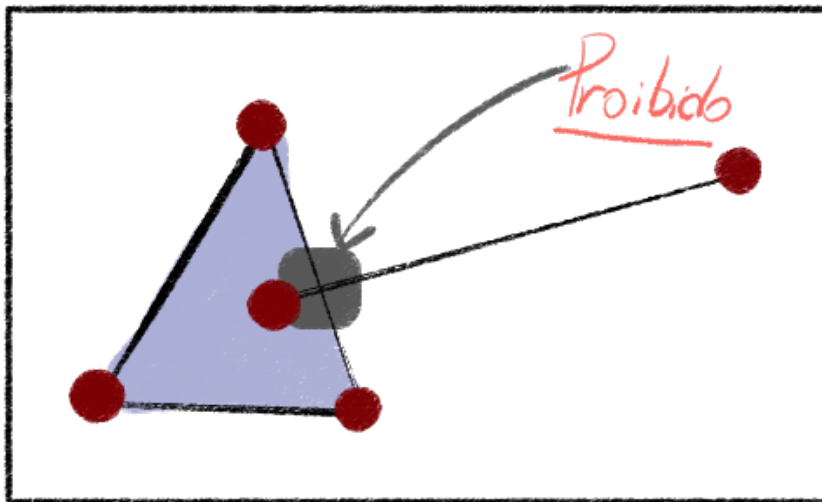


Figura 3 – Exemplo em que a interseção de dois simplexos não é um simplexo.

Definição 2.1.3. Seja K um complexo simplicial. Definimos uma filtração de K sendo uma sequência de subconjuntos $K_i \subset K$, com $i \in \{1, \dots, n\}$, de tal forma que K_i é um complexo simplicial para todo i e vale que

$$K_1 \subset \dots \subset K_{n-1} \subset K_n = K.$$

Na Figura 4 temos um exemplo de filtração para um complexo simplicial.

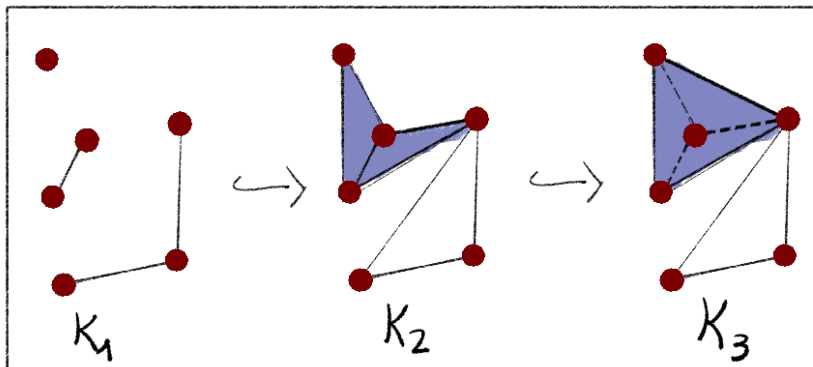


Figura 4 – Exemplo de filtração para um complexo simplicial K .

A pergunta que temos agora é como construir uma filtração dado um conjunto de dados? E felizmente temos várias respostas para isso.

2.2 A matriz de bordo ∂

2.3 Redução da matriz

MÓDULOS DE PERSISTÊNCIA

GERADORES ÓTIMOS E OUTROS CONCEITOS

4.1 Geradores ótimos

4.2 Vetorização do diagrama de persistência

4.3 Mapper

APLICAÇÕES

Neste capítulo serão descritas algumas aplicações utilizando geradores ótimos e imagens de persistência.

5.1 Geradores ótimos em classificadores de imagens

Utilizando imagens e rótulos associados a elas é possível criar classificadores, algoritmos que decidem os rótulos dada uma imagem. Alguns deles são Redes Neurais (MCCULLOCH; PITTS, 1943), SVM (CORTES; VAPNIK, 1995), Redes Neurais Convolucionais (abreviado por CNN, sigla em inglês) (LECUN *et al.*, 1989) e *Generative Adversarial Networks (GAN)* (GOODFELLOW *et al.*, 2014).

Nesta seção será descrito as redes neurais convolucionais e como obteve-se um classificador de imagens utilizando-as. Além disso, será descrito como outros classificadores foram gerados utilizando informações disponibilizadas pelos geradores ótimos para obter-se um classificador com melhor acurácia do que a rede neural convolucional original.

5.1.1 Redes Neurais Convolucionais (CNN)

O algoritmo de redes neurais artificiais é o precursor da CNN. Um rede neural artificial é uma composição de funções f_n que tem como contra domínio algum \mathbb{R}^m . O seu domínio é dado pela dimensão dos dados disponíveis, por exemplo, se temos uma imagem de tamanho 10x10, a dimensão do domínio é 100. Logo, a rede neural pode ser descrita como uma função $Ann: \mathbb{R}^p \rightarrow \mathbb{R}^m$

$$Ann(x) = f_n(\dots f_2(A_2 * f_1(A_1 * x + b_1) + b_2), \quad (5.1)$$

onde A_i é uma matrix de tamanho arbitrário e $b_i \in \mathbb{R}$. Na Figura 5, temos uma imagem clássica para redes neurais.

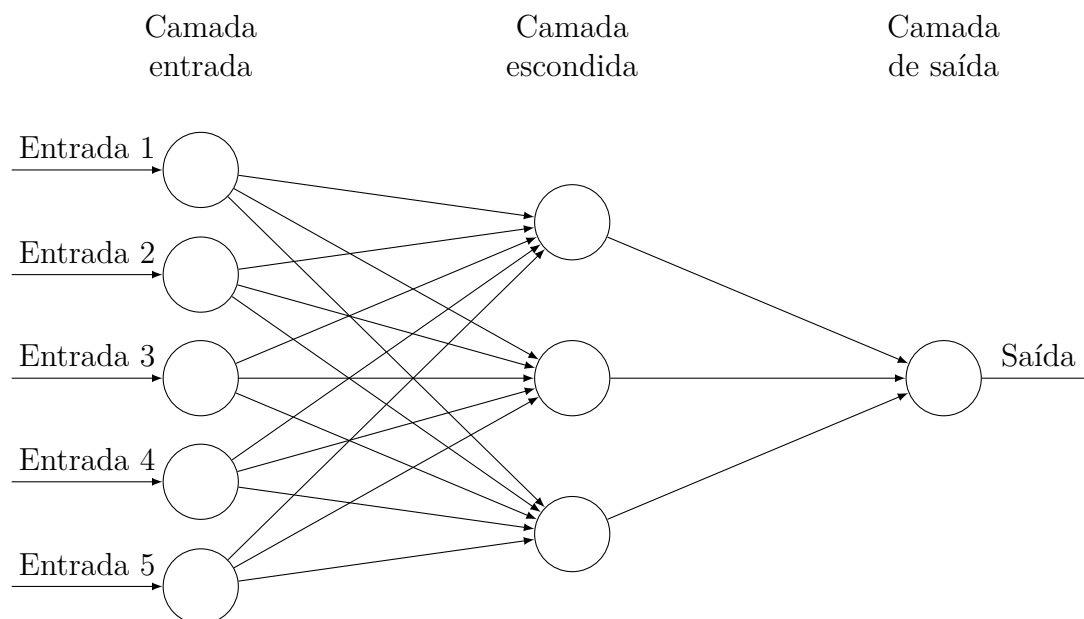


Figura 5 – Esquema de uma rede neural artificial. O número de vértices na camada escondida é determinado pelo tamanho da matriz A_i

5.2 Imagens de persistência aplicadas a proteínas

CONCLUSÃO

REFERÊNCIAS

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, Springer Nature, v. 20, n. 3, p. 273–297, set. 1995. Disponível em: <<https://doi.org/10.1007/bf00994018>>. Citado na página 29.

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 2672–2680. Disponível em: <<http://dl.acm.org/citation.cfm?id=2969033.2969125>>. Citado na página 29.

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. **Neural Computation**, MIT Press - Journals, v. 1, n. 4, p. 541–551, dez. 1989. Disponível em: <<https://doi.org/10.1162/neco.1989.1.4.541>>. Citado na página 29.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, Springer Nature, v. 5, n. 4, p. 115–133, dez. 1943. Disponível em: <<https://doi.org/10.1007/bf02478259>>. Citado na página 29.

POINCARÉ, H. Analysis situs. **Journal de l'École Polytechnique**, p. 1–123, 1895. Citado na página 21.

