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)

ICMC USP
SÃO CARLOS

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. 37 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. 37 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

LISTA DE ALGORITMOS

LISTA DE CÓDIGOS-FONTE

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

CAPÍTULO

2

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 divido 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.

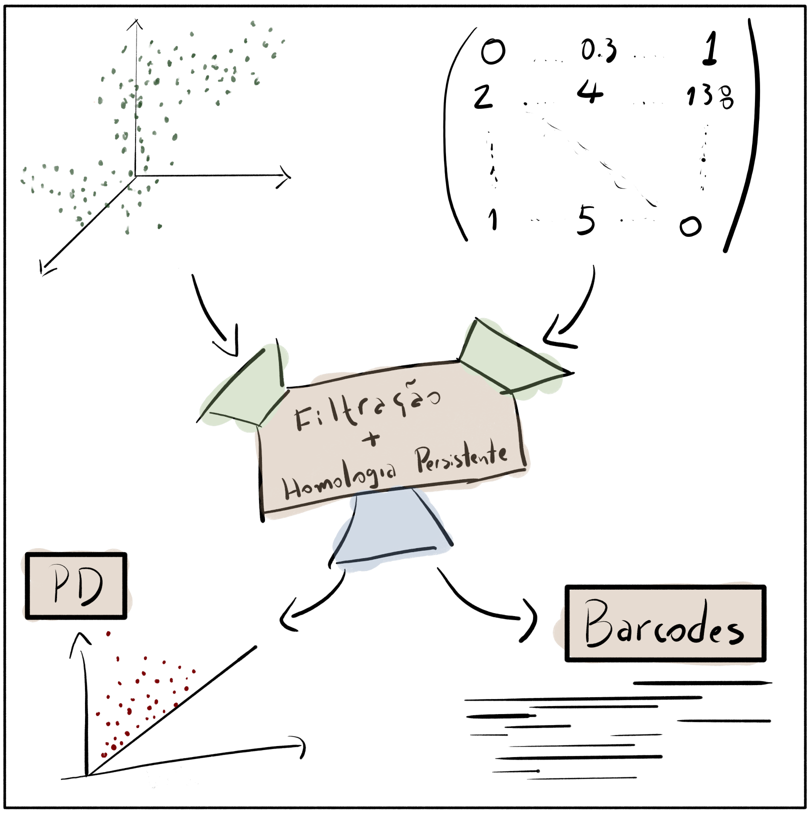

Figura 1 – Representação do pipeline para a utilização da homologia persistente com um conjunto de dados.

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

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

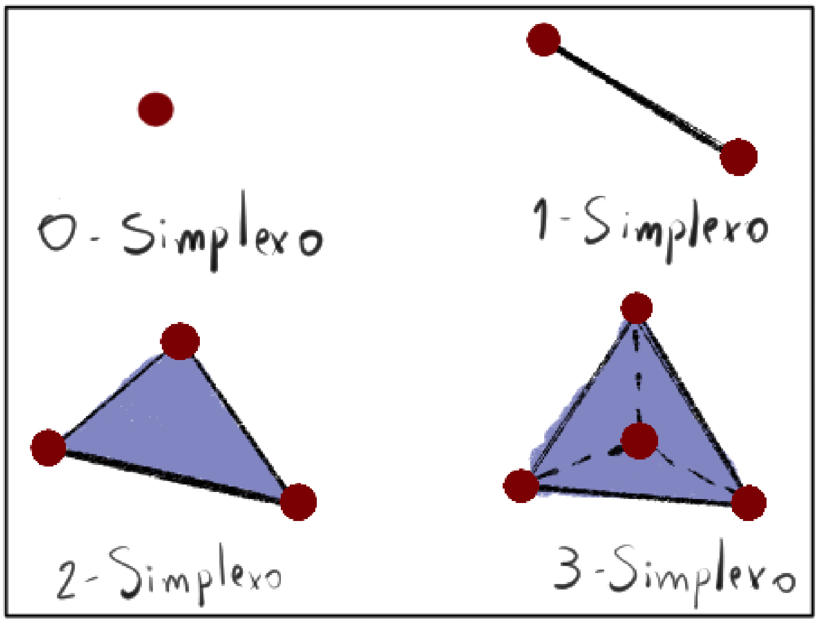

Figura 2 – Exemplos de k-simplexos para $k \in \{0,1,2,3\}$.

um conjunto linearmente independente. O k-simplexo definido pelos pontos acima, chamados de vértices, é a envoltória convexa, definida na Equação 2.1.

$$\left\{ \sum_{i=0}^{k} \lambda_i v_i \;\middle|\; \sum_{i=0}^{k} \lambda_i = 1 \text{ e } \lambda_i \geq 0,\ \forall i \right\}. \tag{2.1}$$

Denotamos o k-simplexo por $< v_0, \ldots, v_k >$.

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 envoltória convexa de qualquer subconjunto dos vértices de um simplexo σ é chamado de face de σ.

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.

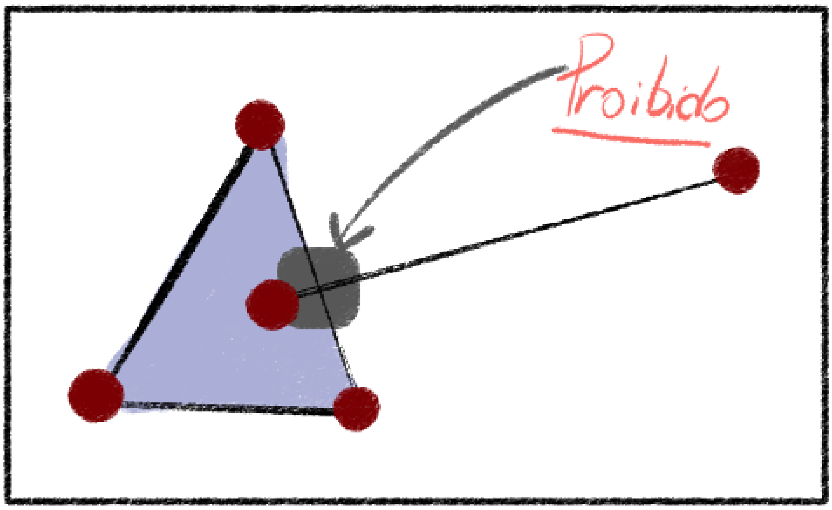

Figura 3 – Exemplo em que a interseção de dois simplexos não é um simplexo.

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

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

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

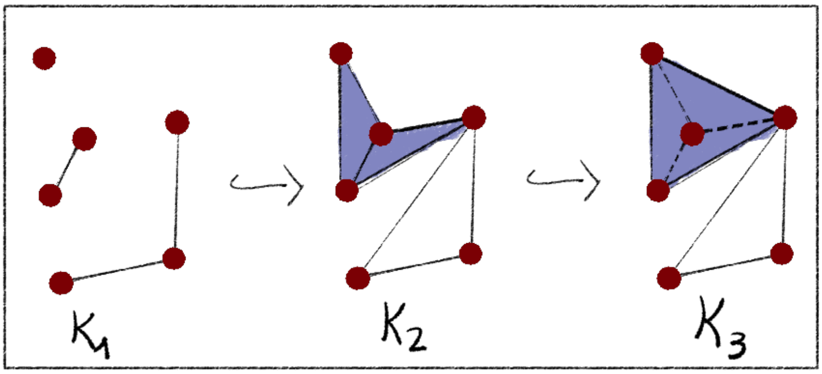

Figura 4 – Exemplo de filtração para um complexo simplicial K.

2.1.1 Complexo de Čech

Para construir complexos simpliciais a partir dos dados, precisamos abstrair a noção de um simplexo simplicial. Na definição dada anteriormente, temos uma representação geométrica do que é um simplexo, mas podemos abstrair tal noção dando origem aos *complexos simpliciais abstratos*.

Figura 5 – Exemplo de um complexo simplicial abstrato e sua realização geométrica

Definição 2.1.4. Seja X um conjunto finito com pontos quaisquer. Seja F um conjunto de subconjuntos não-vazios de X. Dizemos que F é um complexo simplicial abstrato de X se a seguinte condição é satisfeita.

- Se para todo $\sigma \in F$, temos que para todo subconjunto $\sigma' \subset \sigma$ está em F também.

Cada elemento $\sigma \in F$ é chamado de simplexo.

Exemplo 2.1.1. Seja $X = \{a, b, c\}$ e considere $F = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\}$. Precisamos mostrar que F é um complexo simplicial abstrato. Seja $\sigma = \{a, c\}$. Note que seus subconjuntos são $\{a\}$ e $\{c\}$, além disso ambos pertencem a F. De forma análoga, mostramos que para qualquer outro simplexo, suas faces (subconjuntos) estão em F.

Podemos realizar os complexos simpliciais abstratos geometricamente, ou seja, apesar de trabalharmos com conjuntos de elementos quaiser, podemos incluir esses complexos em algum $\mathbb{R}^n$ e assim visualiza-los. Para obtermos o complexo simplicial *geométrico*, associamos a cada simplexo abstrato σ um simplexo geométrico. Por exemplo, se adotarmos o complexo simplicial abstrato F acima mostrado, teriamos que sua realização geométrica seria um triângulo sem preenchimento, como é mostrado na Figura 5.

Observe que se o nosso conjunto X for um subconjunto finito de $\mathbb{R}^d$, podemos ter simplexos de dimensão maiores do que d, ou seja, não podem ser realizados (ou visualizados) em $\mathbb{R}^d$ necessariamente. Um exemplo dessa situação pode ser visto na Figura 6 com o conjunto $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^2$ e $n > 3$.

Essa é uma grande diferença entre os complexos simpliciais geométricos e abstratos. Uma vez tendo definido os complexos simpliciais abstratos, podemos definir o *complexo de Čech*.

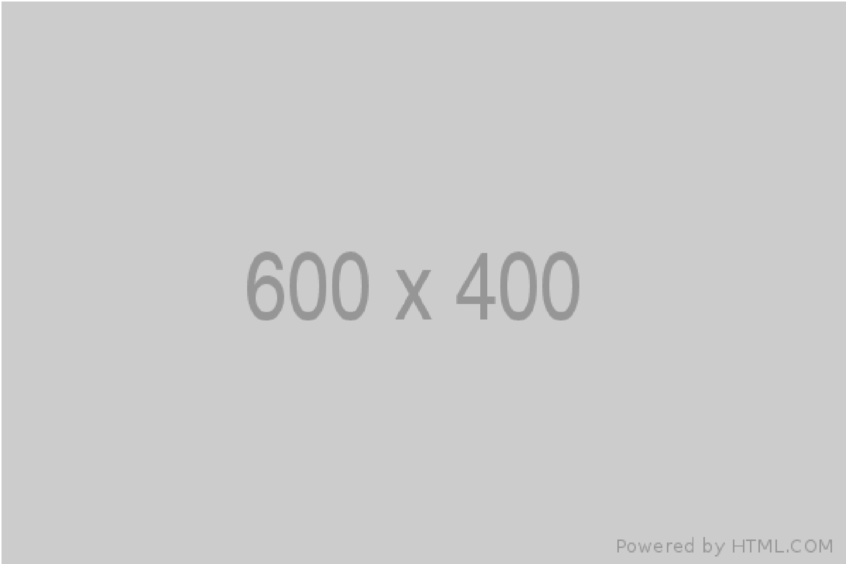

Figura 6 – Cada ponto na imagem corresponde a realização geométrica dos pontos de X. Note que temos um tetraedro neste caso, apesar de estarmos com pontos em $\mathbb{R}^2$.

Definição 2.1.5. Seja X um conjunto de pontos $\{x_1,\ldots,x_n\}$ em $\mathbb{R}^d$. O complexo de Čech de X para um valor real $r > 0$ é o conjunto $C^r(X)$, onde $\sigma =< x_{i_1},\ldots,x_{i_k} >\in C^r(X)$ se, e somente se vale a seguinte condição

$$\bigcap_{j=1}^{k} B(x_{i_j}, r) \neq \varnothing.$$

A definição acima nos diz que quando temos k pontos cujas bolas de raio r centradas neles se intersectam, adicionamos um k simplexo no complexo simplicial abstrato, o que seria apenas o conjunto desses pontos. Geometricamente falando, se duas bolas se intersectam, adicionamos uma aresta. Se três bolas se intersectam, adicionamos um triângulo preenchido, e assim por diante. Na Figura 7 temos um exemplo do complexo simplicial de Čech.

Da mesma forma que definimos a filtração para um complexo simplicial geométrico, o mesmo vale para o caso abstrato.

2.1.2 Complexo de Vietoris-Rips

O complexo de Vietoris-Rips possui uma construção similar ao complexo de Čech, porém computacionalmente é um método mais barato, já que analisa apenas distância entre pontos dois a dois.

Definição 2.1.6. Seja X um conjunto de pontos $\{x_1,\ldots,x_n\}$ em $\mathbb{R}^d$. O complexo de Vietoris-Rips de X para um valor real $r > 0$ é o conjunto $C^r(X)$, onde o simplexo $\sigma =< x_{i_1},\ldots,x_{i_k} >\in V^r(X)$ se, e somente se vale a seguinte condição

$$d(x_{i_k}, x_{i_j}) < r \ \forall j,l \in 1,\ldots,k.$$

Figura 7 – Exemplo de um complexo de Čech para um raio r fixado. Note que temos um tetraedro, apesar dos pontos estarem no plano.

A Figura 8 é um exemplo do complexo de Vietoris-Rips. Uma das diferenças que a construção dos dois complexos já definidos nos dá é que no caso do complexo de Čech temos triângulos preenchidos, e isso não ocorre para Vietoris-Rips.

Mesmo com as regras diferentes para a construção de complexos, temos a relação mostrada na **??**.

$$C^r(X) \subset V^r(X) \subset C^{2r}(X) \tag{2.2}$$

2.1.3 Complexo Alpha Shape

E como uma terceira opção para a construção de um complexo simplicial através de pontos no $\mathbb{R}^n$, temos o complexo Alpha Shape. A construção é similar ao complexo de Čech, porém as bolas são uma interseção de bolas no $\mathbb{R}^n$ com conjuntos convexos especiais, as células de Voronoi.

O diagrama de Voronoi é um tipo especial de decomposição de um espaço métrico, um conjunto que possui uma distância associada a ele, em especial o $\mathbb{R}^n$. Dado um subconjunto $X \subset \mathbb{R}^n$ finito, onde $X = \{x_1, \ldots, x_k\}$, definimos a célula de Voronoi associada ao ponto x_i sendo o seguinte conjunto

$$V_i = \left\{ x \in \mathbb{R}^n \,\middle|\, d(x_i, x) \leq d(x_j, x),\ \forall j \in 1, \ldots, k \right\},$$

em que d é a distância euclidiana usual.

Figura 8 – Exemplo do complexo de Vietoris-Rips com os mesmos pontos utilizados para a construção na Figura 7.

2.2 A matriz de bordo ∂

2.3 Redução da matriz

CAPÍTULO

3

MÓDULOS DE PERSISTÊNCIA

CAPÍTULO

4

GERADORES ÓTIMOS E OUTROS CONCEITOS

4.1 Geradores ótimos

4.2 Vetorização do diagrama de persistência

4.3 Mapper

CAPÍTULO

5

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 precurso 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 $Ann : \mathbb{R}^p \to \mathbb{R}^m$

$$Ann(x) = f_n(\ldots f_2(A_2 * f_1(A_1 * x + b_1) + b_2), \tag{5.1}$$

onde A_i é uma matrix de tamanho arbitrário e $b_i \in \mathbb{R}$. Na Figura 9, temos uma imagem clássica para redes neurais.

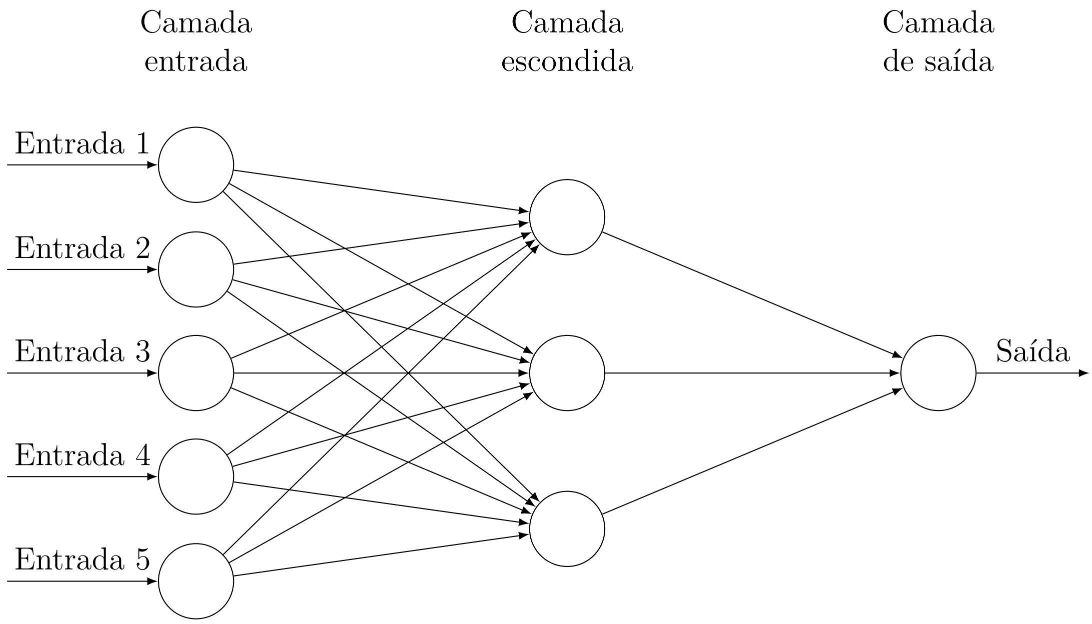

Figura 9 – Esquema de uma rede neural artificial. O número de vértices na camada escondida é determinado pelo tamanho da matriz A_i

5.2 Imagens de persistência aplicadas a proteínas

CAPÍTULO

6

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 33.

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 33.

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 33.

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 33.

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

ICMC USP
SÃO CARLOS