11 EMMSB 2024

Conceitos e aplicações da aprendizagem de máquina

Eduardo Krempser
eduardo.krempser@fiocruz.br

Matheus Müller
matheusp@posgrad.lncc.br

FIOCRUZ

PIBSS
Plataforma Institucional
Biodiversidade e Saúde Silvestre

Introdução

Introdução

- O que é aprendizagem
  - “Assim, a aprendizagem pode ser definida como uma modificação sistemática do comportamento, por efeito da prática ou experiência, com um sentido de progressiva adaptação ou ajustamento.”
    (Psicologia da Aprendizagem. Campos, D. M. S; 1971)

Introdução

- O que é aprendizagem de máquina
  - Disciplina que estuda como um computador pode aprender sem ser explicitamente programado. (Arthur Samuel, 1959)
- Aprendizagem no contexto computacional é adaptação!
  - Os programas de computador melhoram o seu desempenho em uma determinada tarefa a partir da coleta de dados.
- Deixa de se ter uma solução conhecida e previamente programada (sabe-se o como fazer) para um contexto em que apenas sabe-se avaliar a qualidade de uma possível solução (o que fazer)

Introdução

- Hoje está em todas as partes, até mais do que podemos perceber
  - Segurança
  - Recomendação de compras e notícias
  - Atendimento ao consumidor
  - Engenharia
  - **Pesquisa Científica**
  - ...
- Aprendizagem no contexto computacional é adaptação!
  - Os programas de computador melhoram o seu desempenho em uma determinada tarefa a partir da coleta de dados.
- Deixa de se ter uma solução conhecida e previamente programada (sabe-se o como fazer) para um contexto em que apenas sabe-se avaliar a qualidade de uma possível solução (o que fazer)

Introdução

- A Aprendizagem de Máquina é uma subárea da Inteligência Artificial, hoje envolvida por um emaranhado de nomes e novas áreas
  - Aprendizagem de máquina
  - Reconhecimento de padrões
  - Mineração de dados
  - Modelagem a partir de dados
  - Aprendizagem estatística
  - Ciência de Dados
  - Deep Learning
  - BI
  - ...

Introdução

- Entre diversas maneiras de “dividir” a Aprendizagem de Máquina, podemos considerar umas das mais clássicas:
  - Aprendizagem Supervisionada
  - Aprendizagem não supervisionada
  - Aprendizagem por reforço

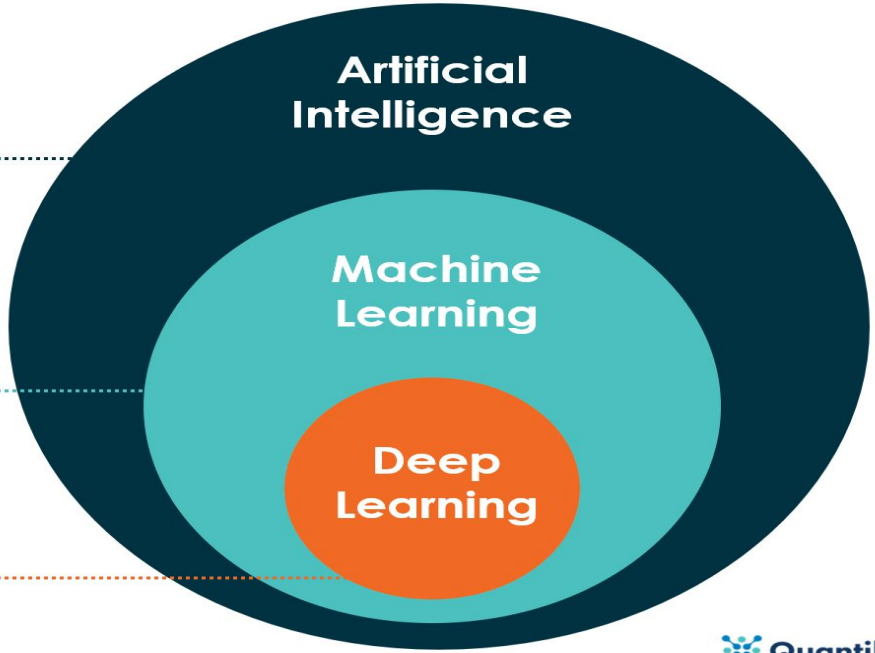
Introdução

- E o que é a Inteligência Artificial?

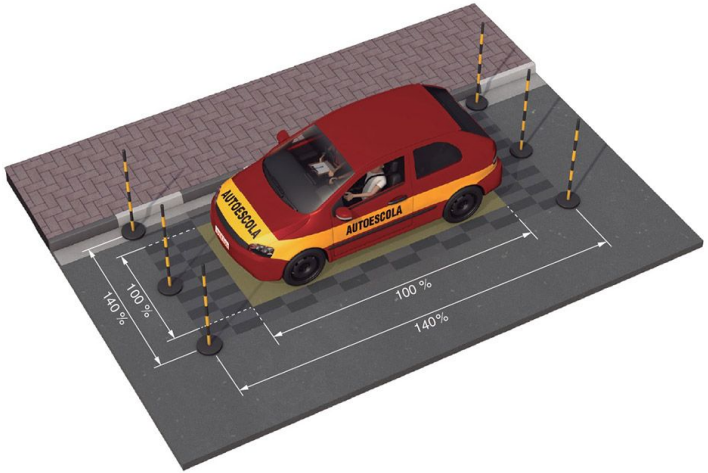
ARTIFICIAL INTELLIGENCE (AI)
Programming systems to perform tasks which usually require human intelligence.

MACHINE LEARNING (ML)
Training algorithms to solve tasks by pattern recognition instead of specifically programming them how to solve the task.

DEEP LEARNING (DL)
Training algorithms by using deep neural networks with multiple layers.

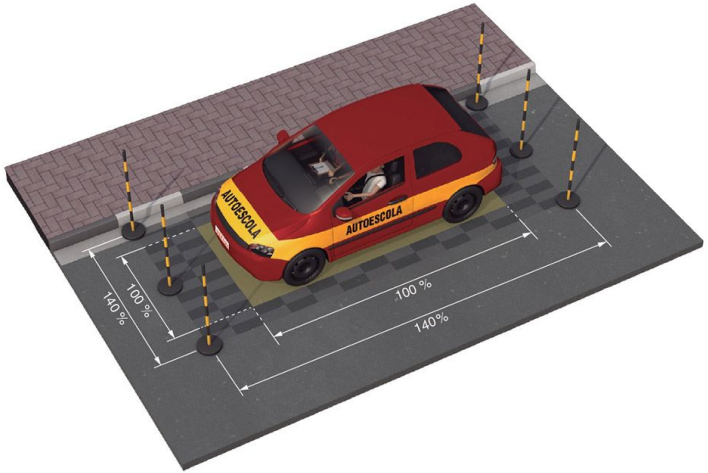


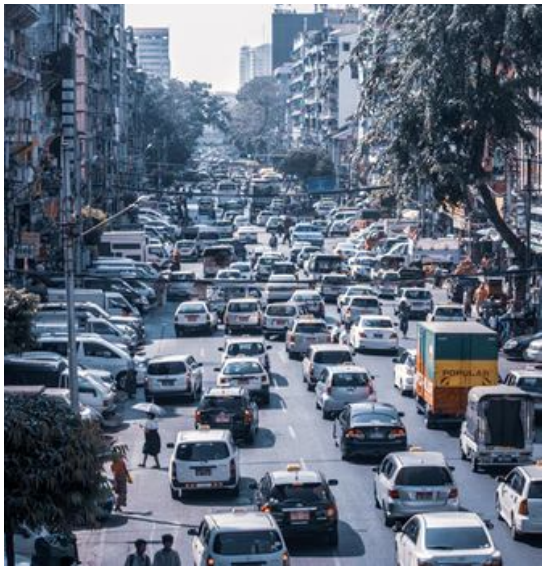
Introdução



Imagens:
https://www.autoescolaonline.net/como-fazer-uma-baliza-perfeita-em-10-passos
https://unsplash.com/photos/q88ZVP2f2fg

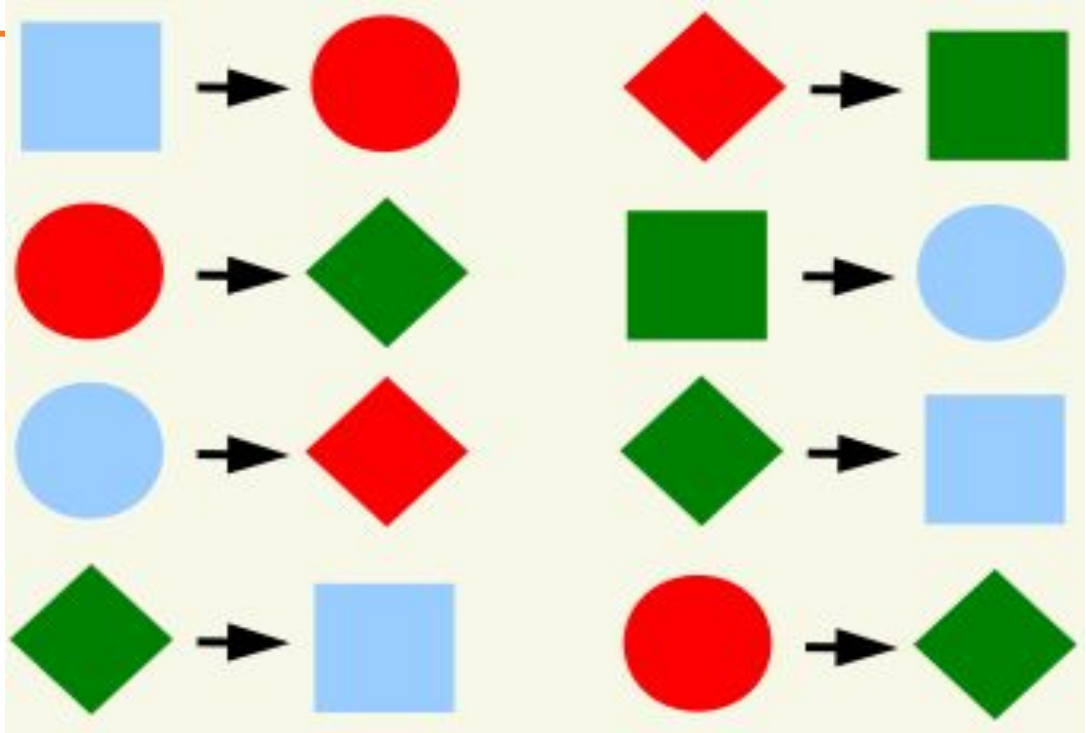
Introdução





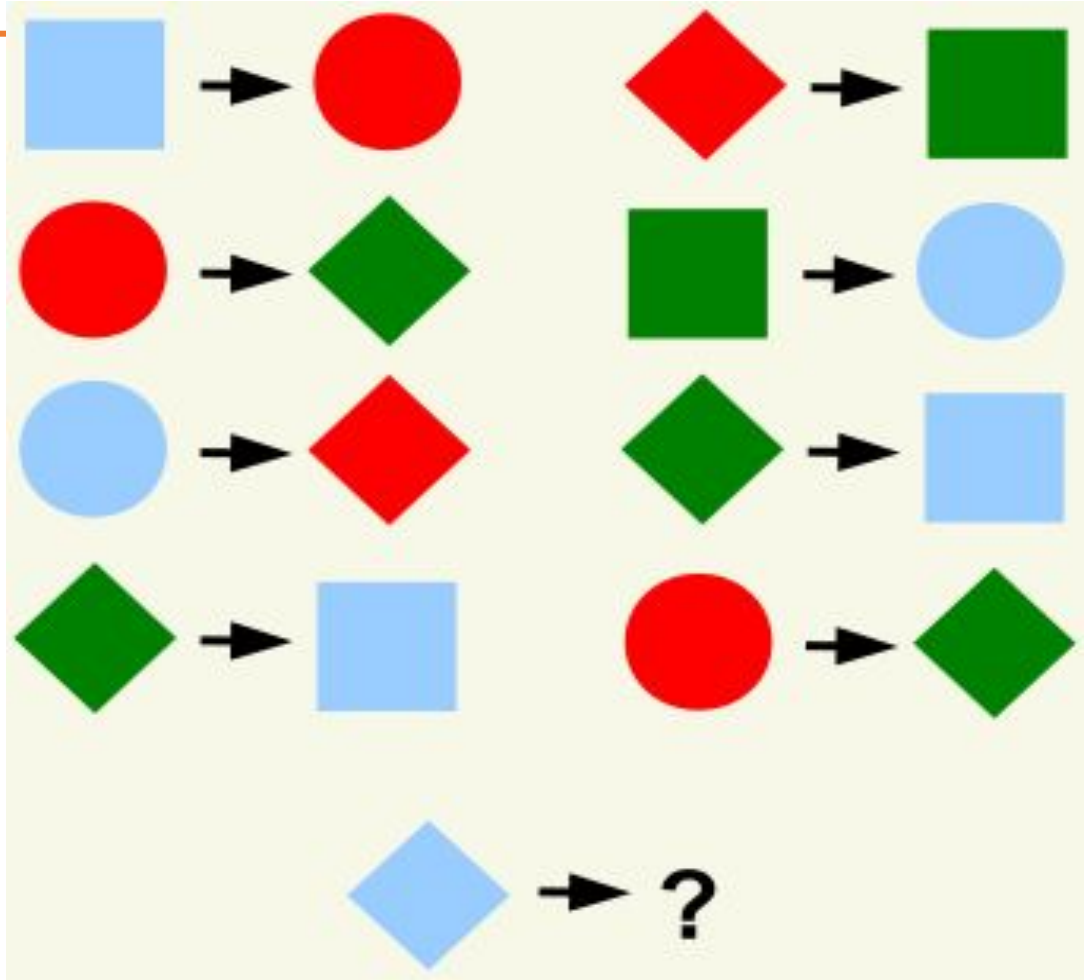
Imagens:
https://www.autoescolaonline.net/como-fazer-uma-baliza-perfeita-em-10-passos
https://unsplash.com/photos/q88ZVP2f2fg

Introdução

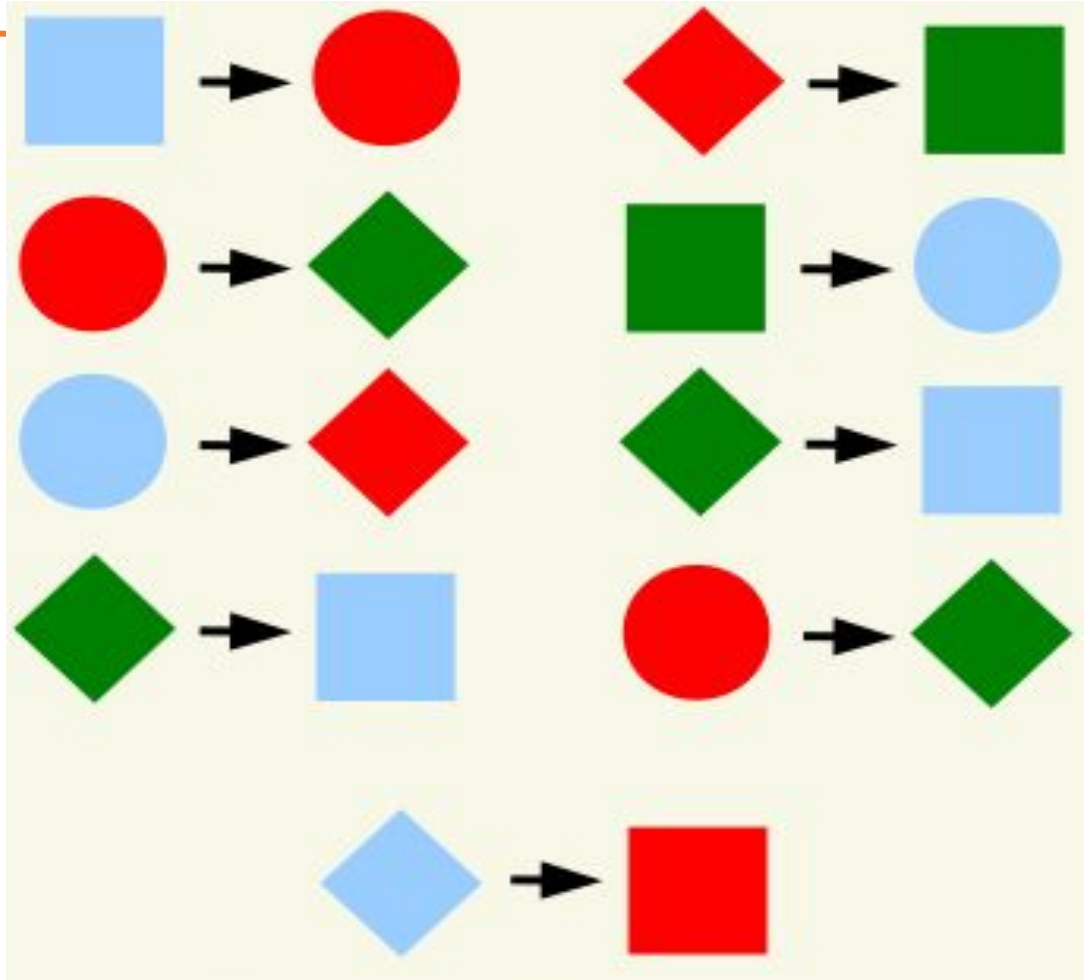


“Notas de Aula: Introdução à Aprendizagem de Máquina”; Barreto, A. M. S.

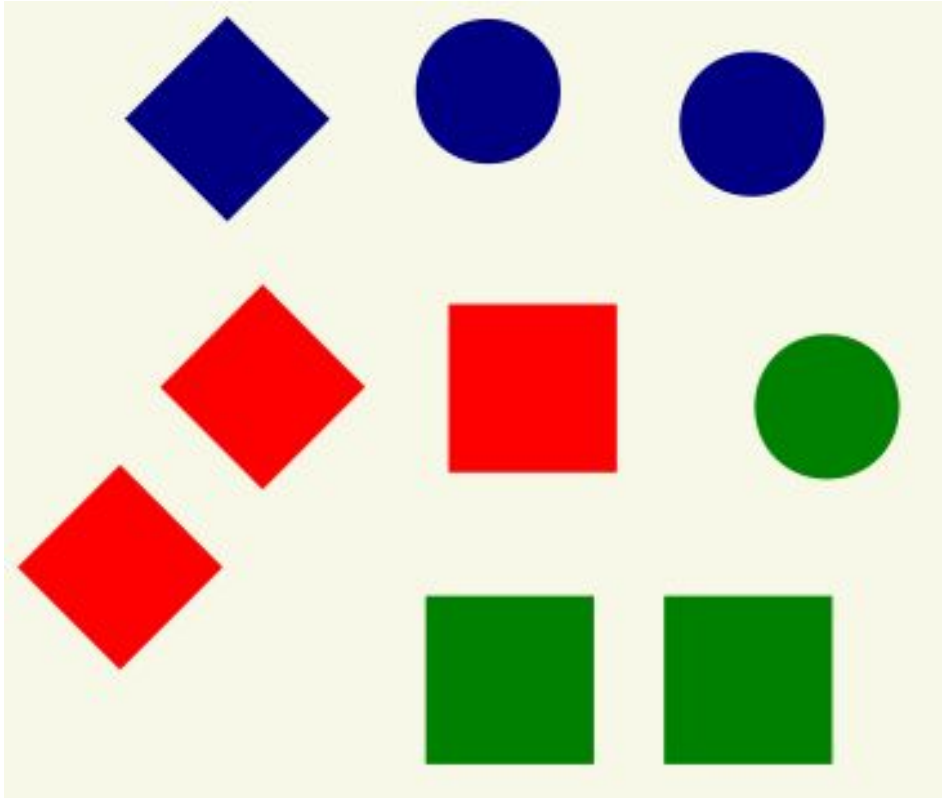
Introdução



Introdução

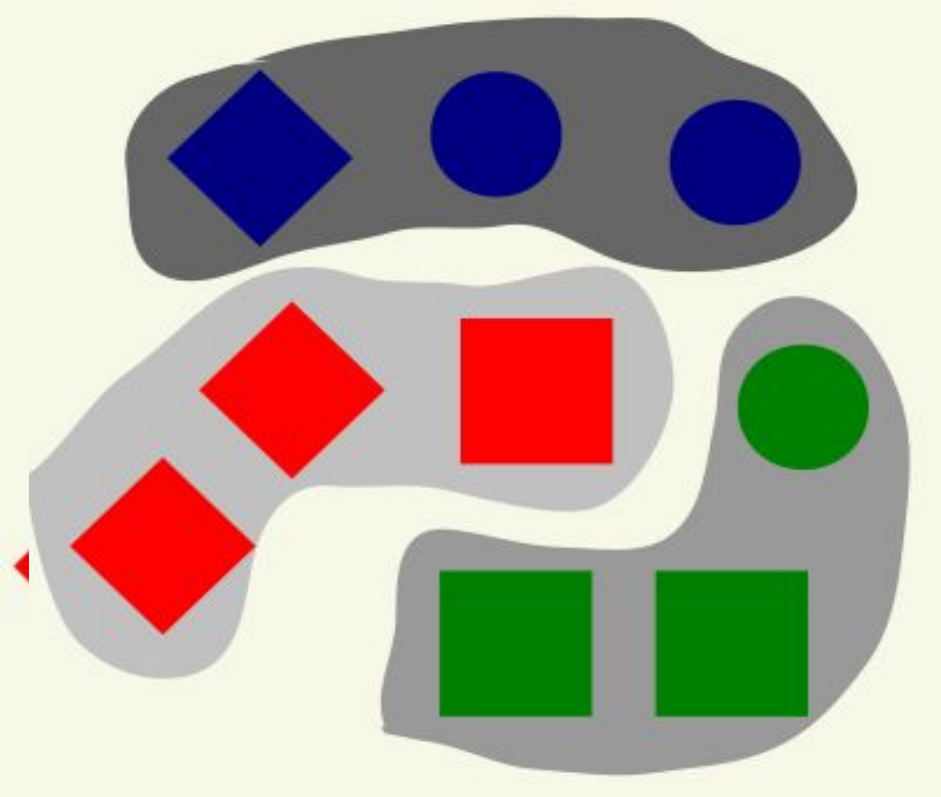


Introdução



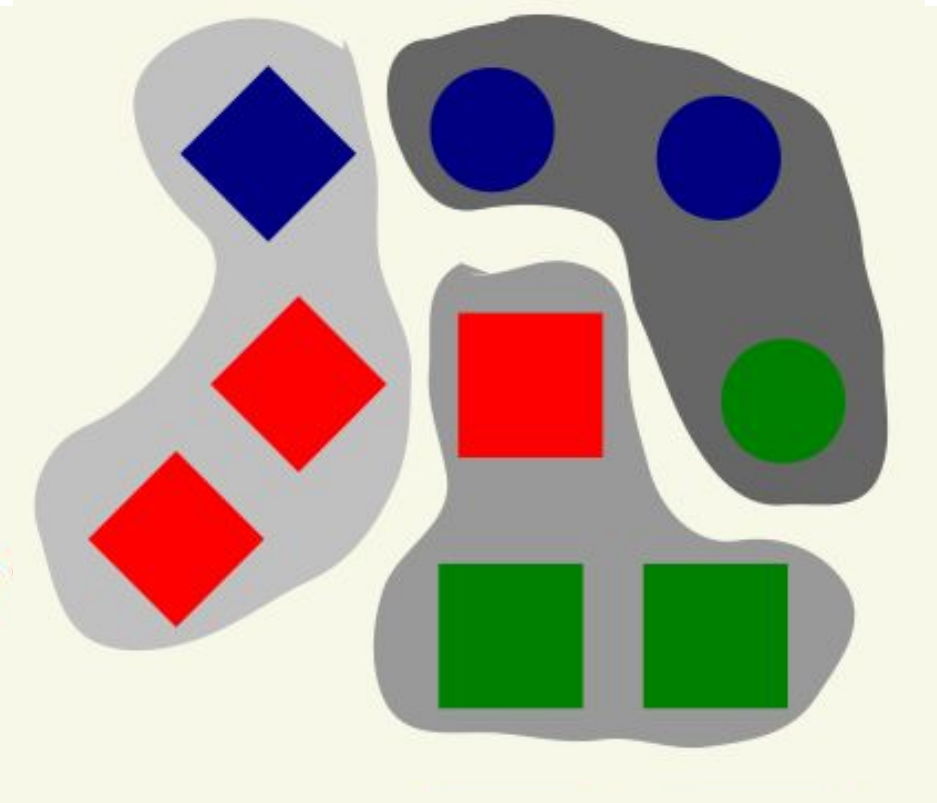
“Notas de Aula: Introdução à Aprendizagem de Máquina”; Barreto, A. M. S.

Introdução



"Notas de Aula: Introdução à Aprendizagem de Máquina"; Barreto, A. M. S.

Introdução



"Notas de Aula: Introdução à Aprendizagem de Máquina"; Barreto, A. M. S.

Aprendizagem Supervisionada

Aprendizagem Supervisionada

Localização	Quartos	Garagem	...	Aluguel (R$/$m^2$)
Copacabana	2	Não	...	41,96
Copacabana	1	Sim	...	49,15
Centro	2	Não	...	28,05
Barra da Tijuca	3	Sim	...	30,86
...	...	...	...	...

Aprendizagem Supervisionada

Localização	Quartos	Garagem	...	Aluguel (R$/$m^2$)
Copacabana	2	Não	...	41,96
Copacabana	1	Sim	...	49,15
Centro	2	Não	...	28,05
Barra da Tijuca	3	Sim	...	30,86
...	...	...	...	...

Localização	Quartos	Garagem	...	Aluguel (R$/$m^2$)
Centro	1	Sim	...	**?**

Aprendizagem Supervisionada

Localização	Quartos	Garagem	...	Aluguel (R$/$m^2$)	Alugo?
Copacabana	2	Não	...	41,96	Sim
Copacabana	1	Sim	...	49,15	Não
Centro	2	Não	...	28,05	Não
Barra da Tijuca	3	Sim	...	30,86	Sim
...	...	...	...	...	...

Localização	Quartos	Garagem	...	Aluguel (R$/$m^2$)
Centro	1	Sim	...	**?**

Aprendizagem Supervisionada

Regressão

Classificação

Localização	Quartos	Garagem	...	Aluguel (R$/m^2)	Alugo?
Copacabana	2	Não	...	41,96	Sim
Copacabana	1	Sim	...	49,15	Não
Centro	2	Não	...	28,05	Não
Barra da Tijuca	3	Sim	...	30,86	Sim
...	...	...	...	...	...

Localização	Quartos	Garagem	...	Aluguel (R$/m^2)
Centro	1	Sim	...	**?**

Aprendizagem Supervisionada

Localização	Quartos	Garagem	...	Aluguel (R$/$m^2$)
Copacabana	2	Não	...	41,96
Copacabana	1	Sim	...	49,15
Centro	2	Não	...	28,05
Barra da Tijuca	3	Sim	...	30,86
...	...	...	...	...

X **y**

Aprendizagem Supervisionada

Title	ALogP	ALogp2	AMR	BCUTp-1	BCUTp-1h	fragC	apol	naAromAt	nAtom
Acebutolol	-2.1004	4.41168	67.0727	4.75506	9.76616	2152.06	55.7582	6	52
Amoxicillin	-1.874	3.51188	63.5964	4.78119	12.1043	1516.09	51.0391	6	44
Bendroflumethlazide	-0.8075	0.65206	39.6471	4.24449	12.8541	1147.12	49.7141	12	41
Benzocaine	-0.4883	0.23844	20.5043	4.54796	8.52863	397.03	25.8787	6	23
Benzthlazide	-1.3398	1.79506	50.9651	6.69282	12.9721	1114.11	53.1231	10	40
Clozapine	0.0092	8.46E-05	45.0577	5.92358	11.5903	1519.05	50.9291	12	42
Dibucaine	-0.5602	0.31382	57.7981	5.09923	10.5046	2425.05	59.441	10	54
Diethylstibestrol	-0.2302	0.05299	33.7086	4.85094	11.5338	1301.02	46.6199	12	40
Diflunisal	0.2593	0.06724	11.5289	4.03982	9.84339	423.05	31.7343	12	26
Dipyridamole	-3.3772	11.4055	95.8664	4.30728	11.5724	4981.12	80.9197	10	76
Folic_Acid	-3.3182	11.0105	52.3395	3.83215	9.33155	1817.13	58.6211	12	51
Furosemide	-0.9949	0.98983	33.6078	4.6696	12.2524	669.09	39.7447	11	32
Hydrochlorothiazide	-1.5436	2.3827	35.4327	7.26833	12.8018	404.1	32.1423	6	25
Imipramine	0.091	0.00828	39.2562	5.70631	11.5894	1789.02	51.643	12	45
Indomethacin	0.2922	0.08538	37.8739	4.12177	11.086	1249.06	50.5967	12	41
Ketoprofen	-0.2088	0.0436	21.9194	4.29006	9.91827	814.03	39.9011	12	33
Lidocaine	1.166	1.35956	45.8867	5.09304	9.98654	1249.03	42.3114	6	39
Meclofenamic_acid	1.4704	2.16208	27.2526	4.4468	11.3886	619.05	39.0387	12	30

Machine learning methods in chemoinformatics. Mitchell, J. B. O.; 2014

Aluguel (R$/m^2)
41,96
49,15
28,05
30,86
...

y

Aprendizagem Supervisionada

- Componentes da aprendizagem
  - Entrada: **X**
  - Saída: **y**
  - Função alvo:
    - $f: \mathbf{X} \rightarrow \mathbf{y}$
  - Dados: $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$

    
  - Hipótese:
    - $g: \mathbf{X} \rightarrow \mathbf{y}$

Learning from Data. Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.; 2012

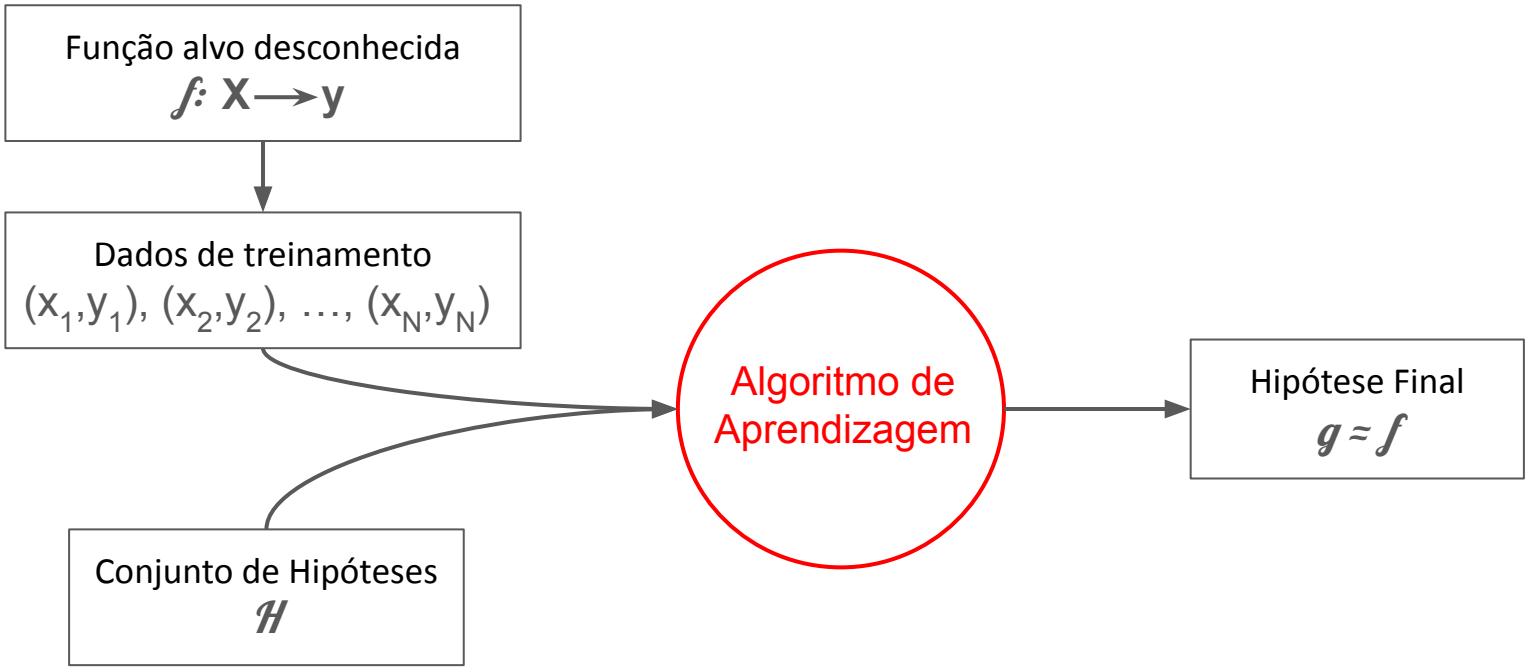
Aprendizagem Supervisionada

- Componentes da aprendizagem
  - Conjunto de hipóteses: $\mathcal{H}$
    - $g \in \mathcal{H}$
  - O Algoritmo de Aprendizagem: $\mathcal{A}$
  - Juntos formam o modelo de aprendizagem

Learning from Data. Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.; 2012

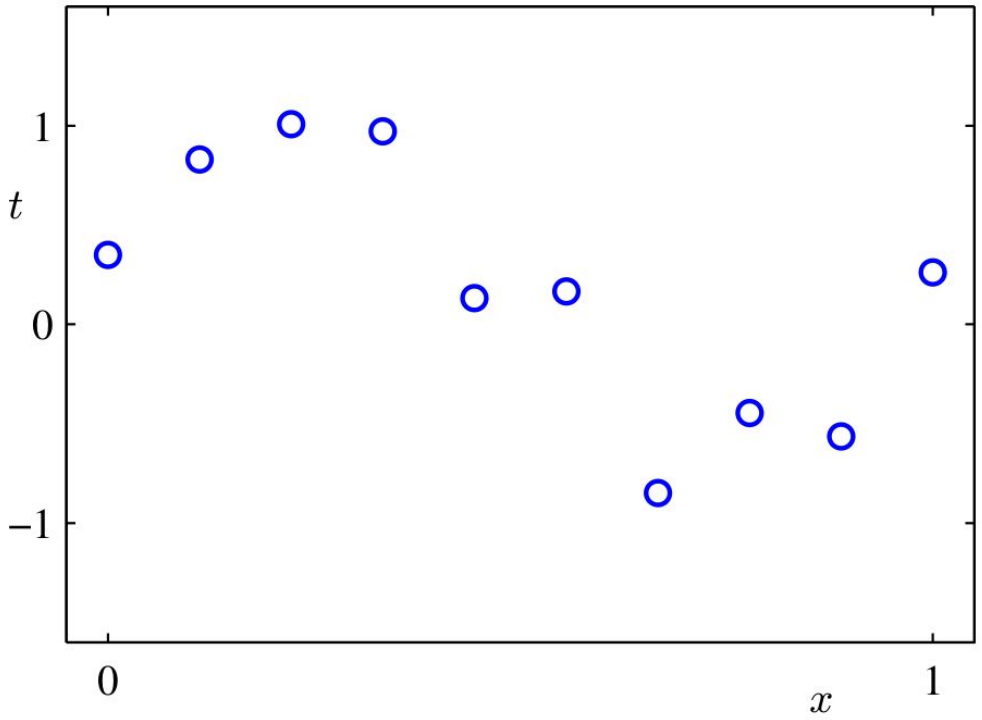
Aprendizagem Supervisionada

- Componentes da aprendizagem



Learning from Data. Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.; 2012

Aprendizagem Supervisionada



Pattern Recognition and Machine Learning. Bishop, C. M.; 2006

Aprendizagem Supervisionada

- Como encontrar um modelo a partir desses dados?
- Uma possibilidade é a obtenção de um modelo paramétrico
- Por exemplo, um polinômio:
  - $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_M x^M$

Pattern Recognition and Machine Learning. Bishop, C. M.; 2006

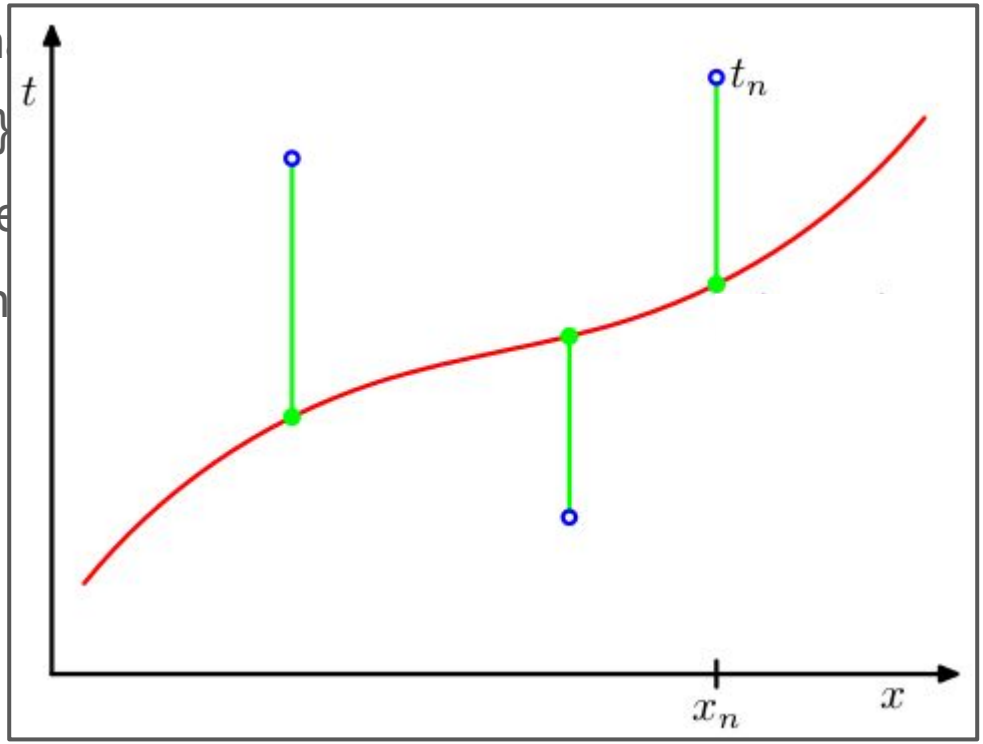
Aprendizagem Supervisionada

- Dado um conjunto de treinamento
  - {(xi, yi): i = 1, 2, …, m}
- O objetivo é encontrar θ de forma que $h_\theta(x_i) \approx y_i$
- Ou seja, minimize uma função de erro, por exemplo:
  - $J(\theta) = \sum_{i=1}^{m}(y_i - h_\theta(x_i))^2$

Pattern Recognition and Machine Learning. Bishop, C. M.; 2006

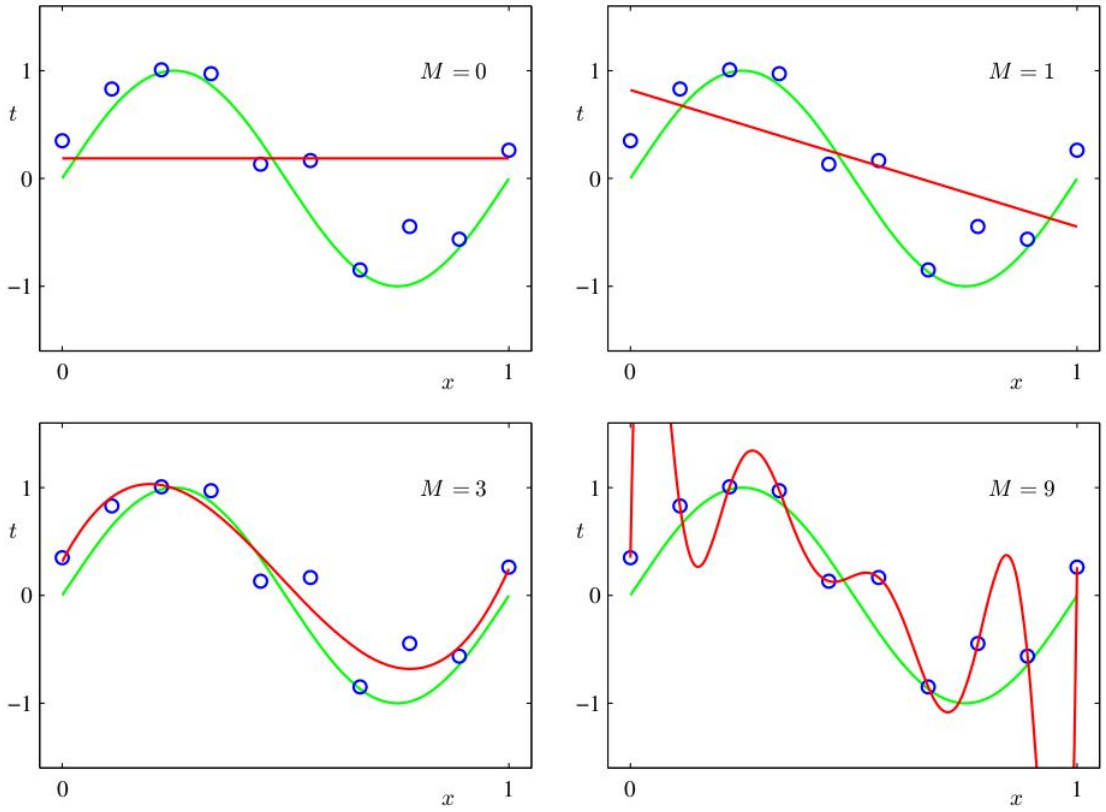
Aprendizagem Supervisionada

- Dado um conjunto de trein
  - {(xi, yi): i = 1, 2, …, m}
- O objetivo é encontrar θ de
- Ou seja, minimize uma fun
  - $J(\theta) = \Sigma_{i=1}^{m}(y_i - h_\theta(x_i))^2$



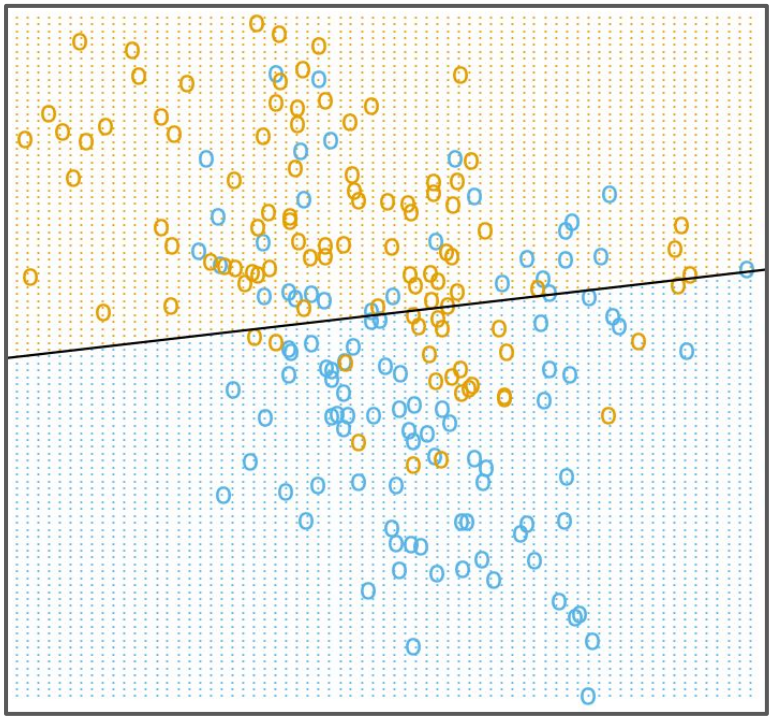
Pattern Recognition and Machine Learning. Bishop, C. M.; 2006

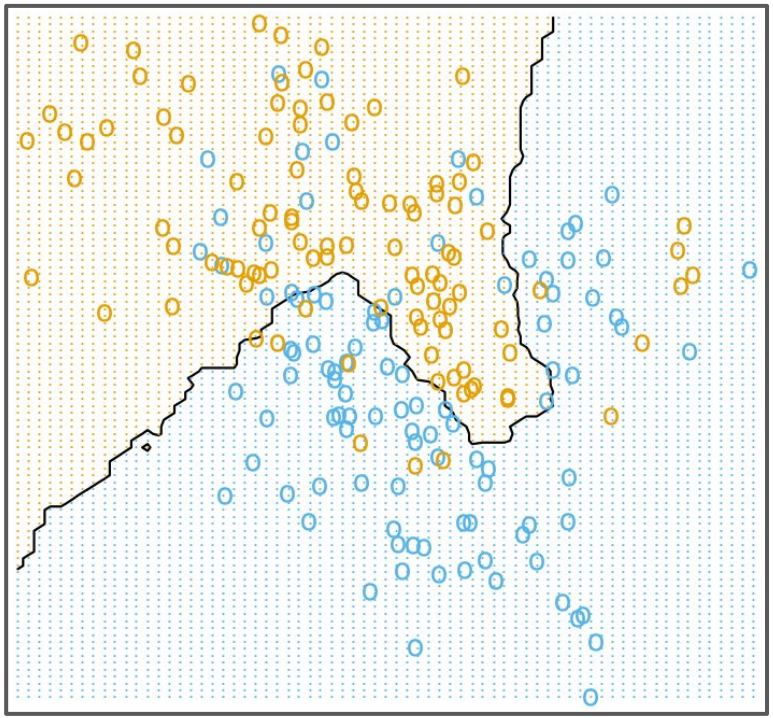
Aprendizagem Supervisionada



Pattern Recognition and Machine Learning. Bishop, C. M.; 2006

Aprendizagem Supervisionada





Pattern Recognition and Machine Learning. Bishop, C. M.; 2006

Aprendizagem Supervisionada

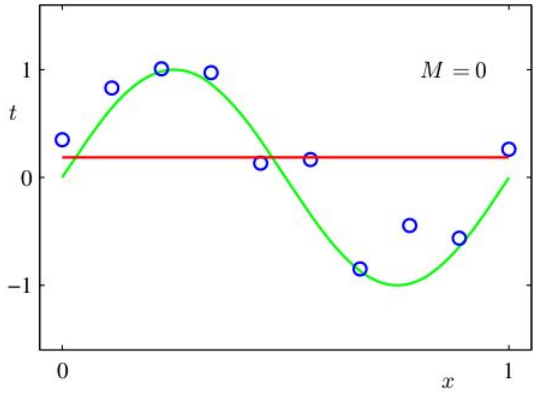
- Qual o objetivo da aprendizagem supervisionada?
- Como escolher entre vários modelos?
- Deve-se buscar modelos que “passem” por todos os pontos conhecidos?

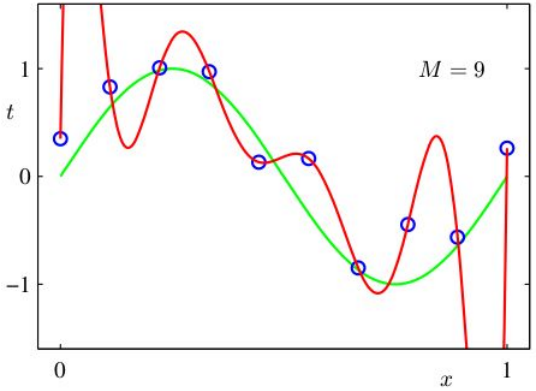
Aprendizagem Supervisionada

- Quais as possíveis origens de ruídos nos dados?
  - Erros na coleta (qualidade de dados)
  - Erros na manipulação dos dados (digitação, arredondamentos,...)
  - Atributos insuficientes para descrever o fenômeno

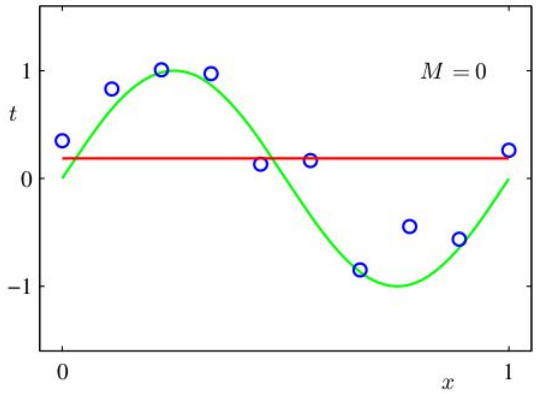
Seleção de Modelos

Seleção de Modelos

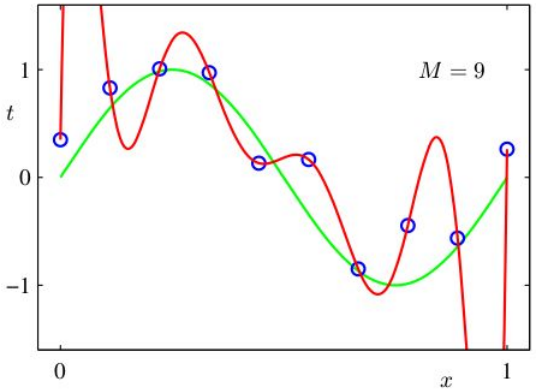




Seleção de Modelos

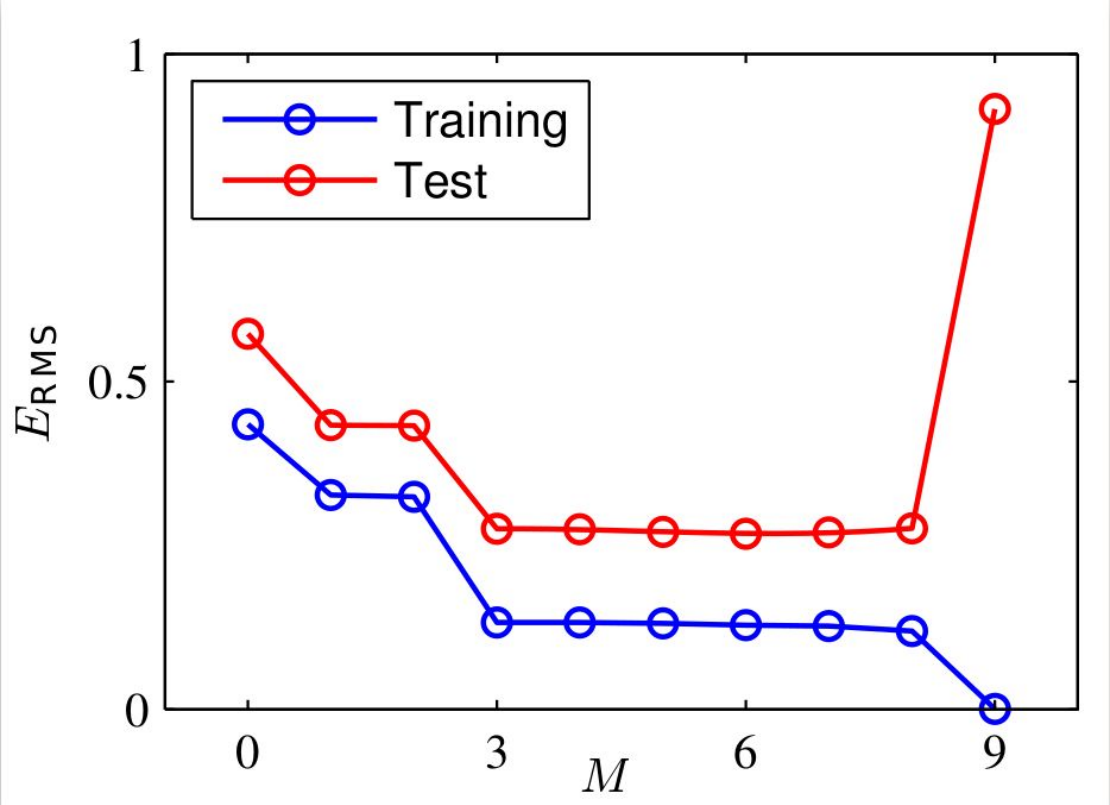


"*underfitting*"
- bias (viés) alto



"*overfitting*"
- variância alta

Seleção de Modelos



Seleção de Modelos

Fonte: https://commons.wikimedia.org/wiki/File:ML_dataset_training_validation_test_sets.png

Muito mais para decidir...

Como equilibrar *bias* e variância?

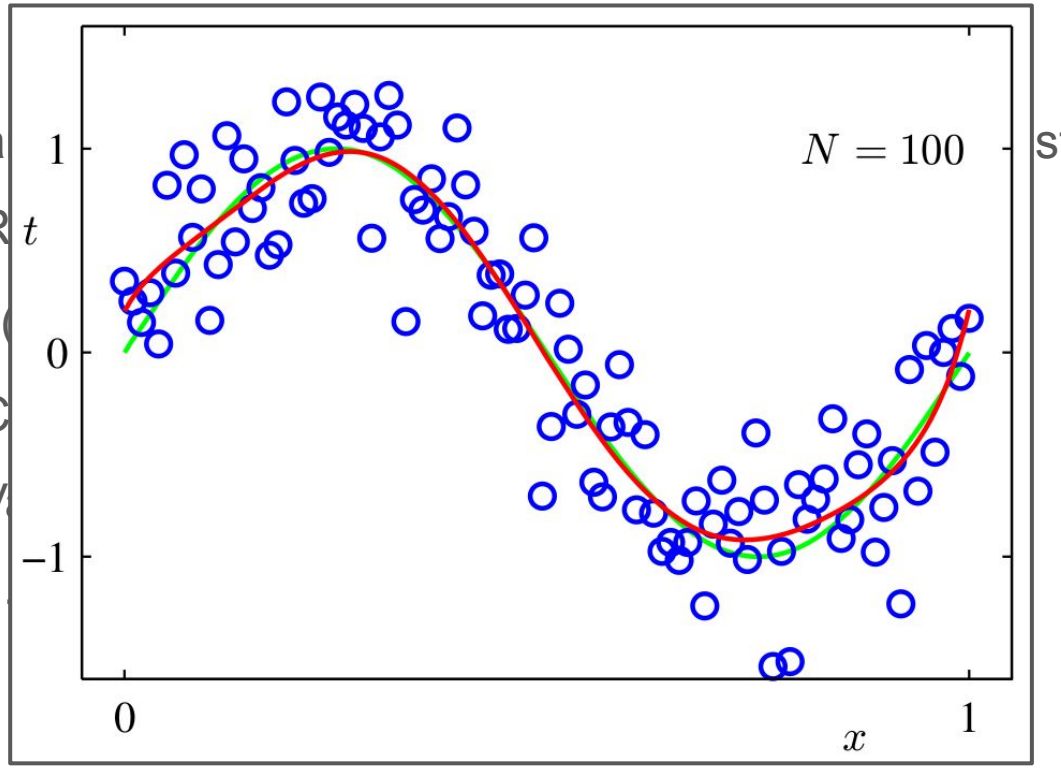
- Coletar o máximo de dados possíveis

- Adotar um modelo complexo e evitar o excesso de ajuste:
  - Regularização

    $J(\theta) = \Sigma_{i=1}^{m}(y_i - h_\theta(x_i))^2 + \lambda\Sigma_{i=1}^{m}\theta_i^2$
  - Comitês de modelo
  - Validação cruzada
  - ...

Como equilibrar *bias* e variância?

- Coletar o máximo de dados possíveis
- Adota ste:
  - R
    J
  - C
  - V
  - ..



Como equilibrar *bias* e variância?

- Coletar o máximo de dados possíveis

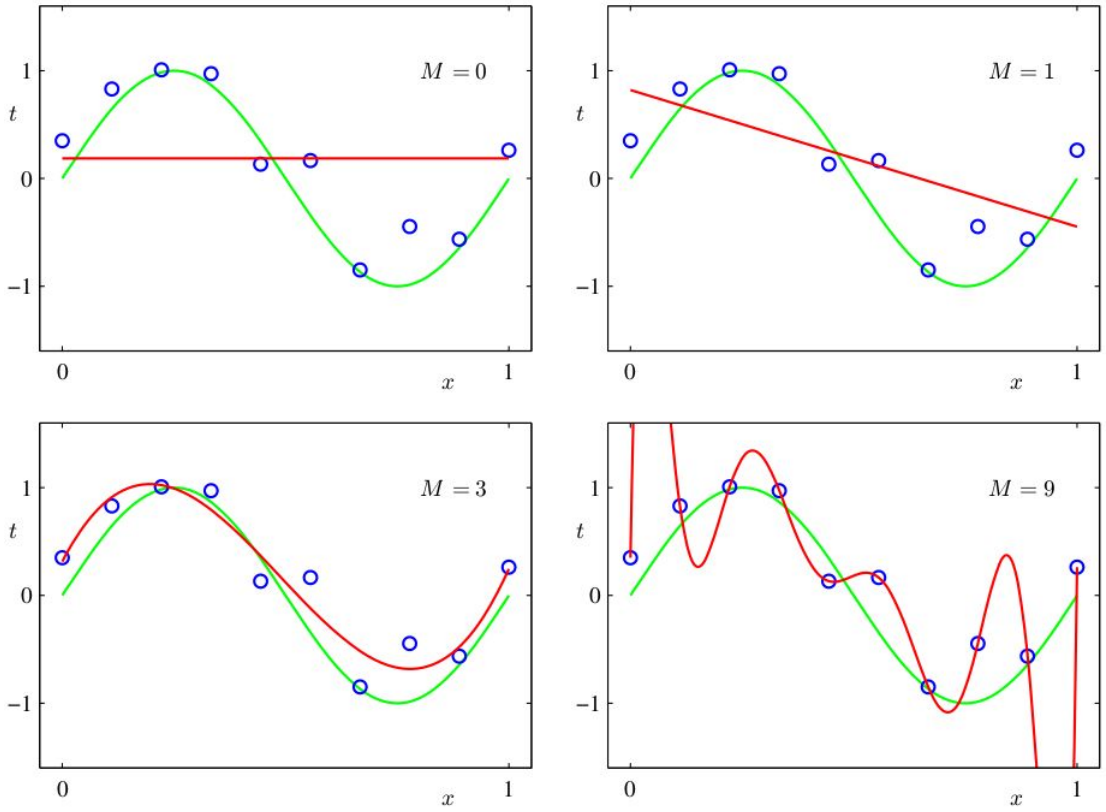
- Adotar um modelo complexo e evitar o excesso de ajuste:
  - Regularização

    $J(\theta) = \Sigma_{i=1}^{m}(y_i - h_\theta(x_i))^2 + \lambda\Sigma_{i=1}^{m}\theta_i^2$
  - Comitês de modelo
  - Validação cruzada
  - ...

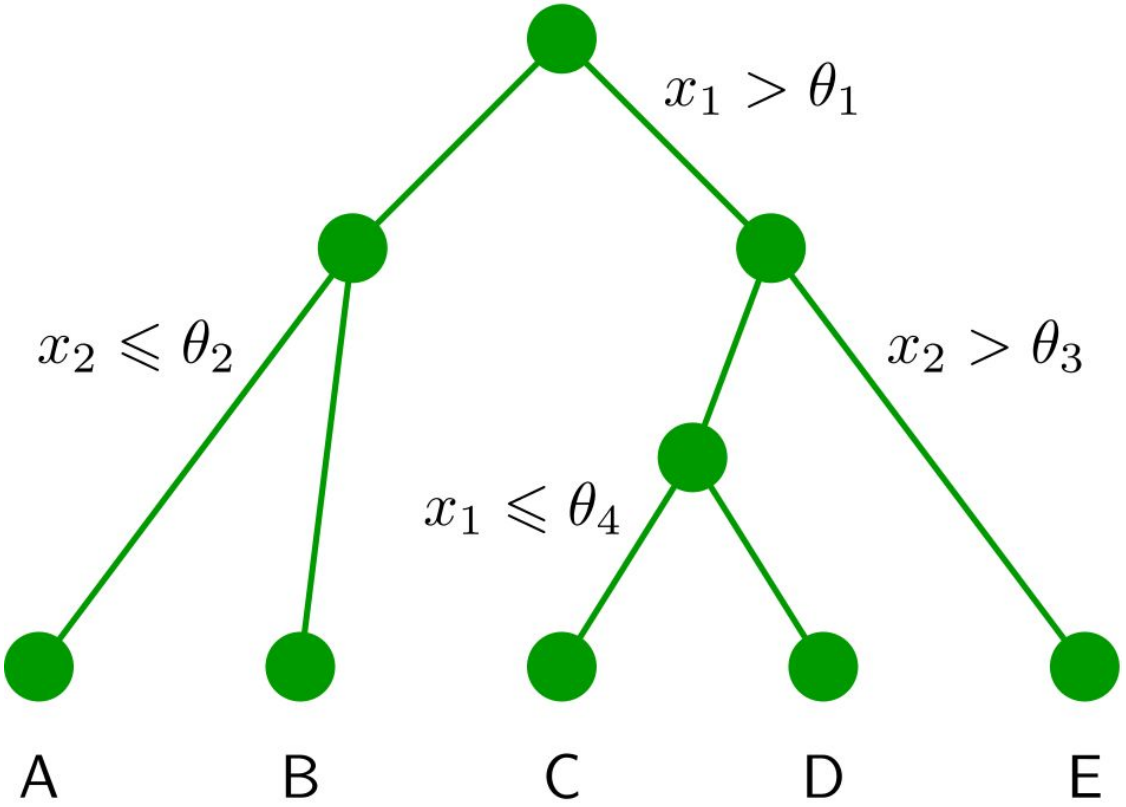
Propriedades desejadas

- Acurácia (generalidade)
- Simplicidade
- Interpretabilidade

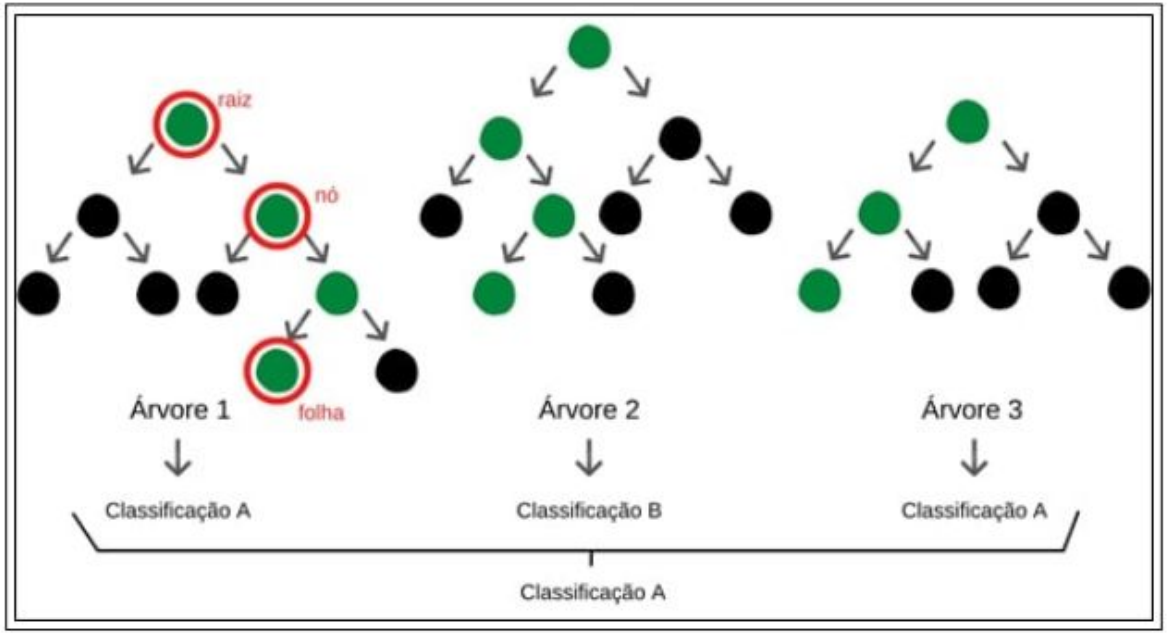
Modelos Polinomiais



Árvores de Decisão

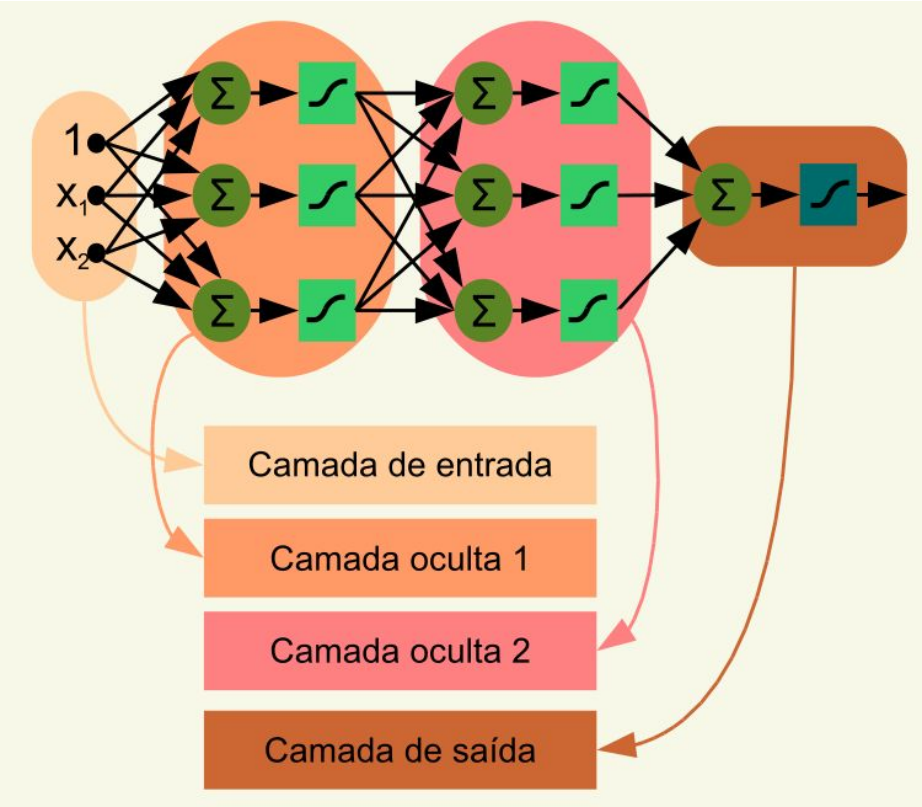


Random Forest



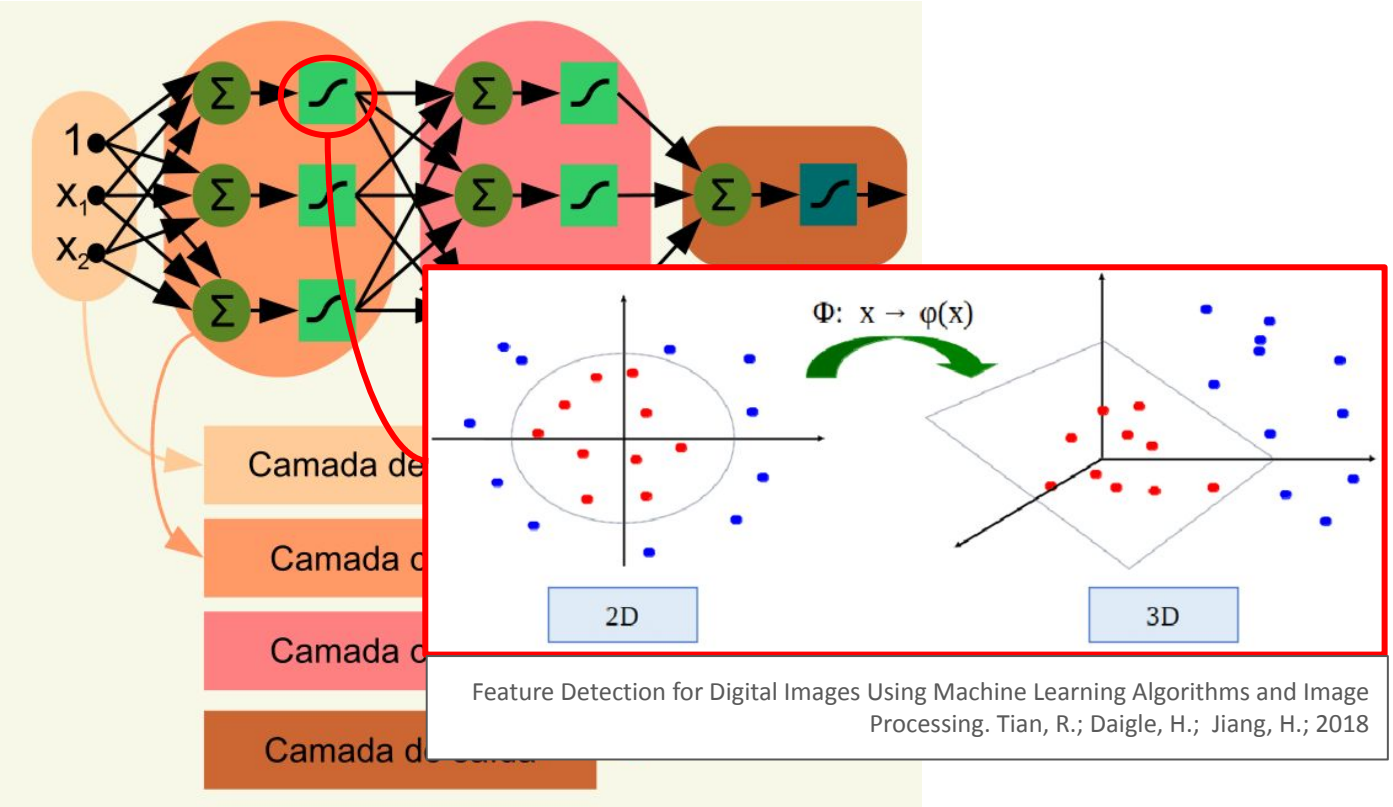
Ariza et al. (2023).

Redes Neuronais



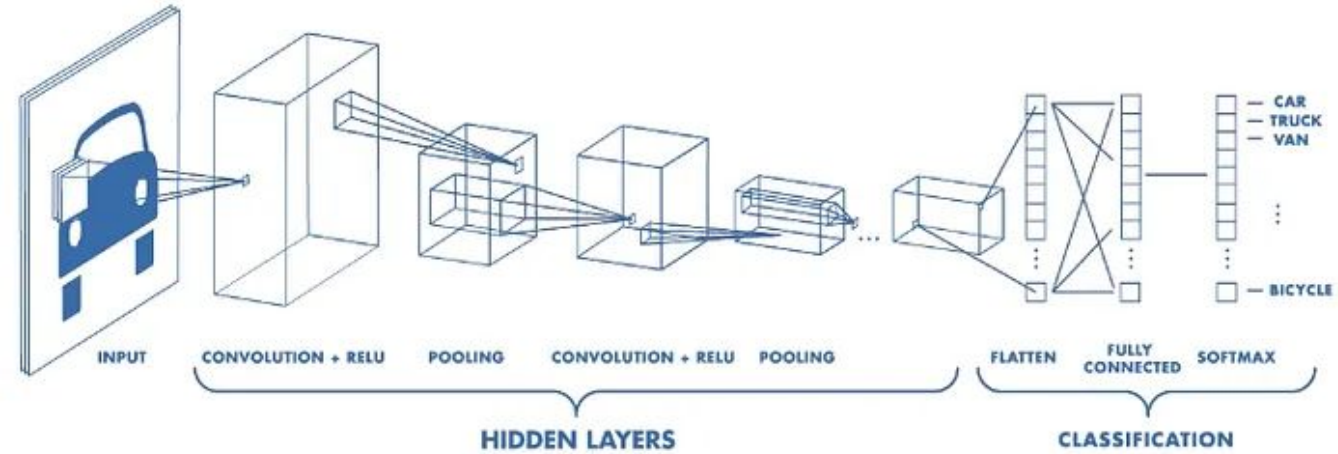
Notas de Aula: Aprendizagem de Máquina. Barreto, A.M.S. LNCC 2009

Redes Neuronais



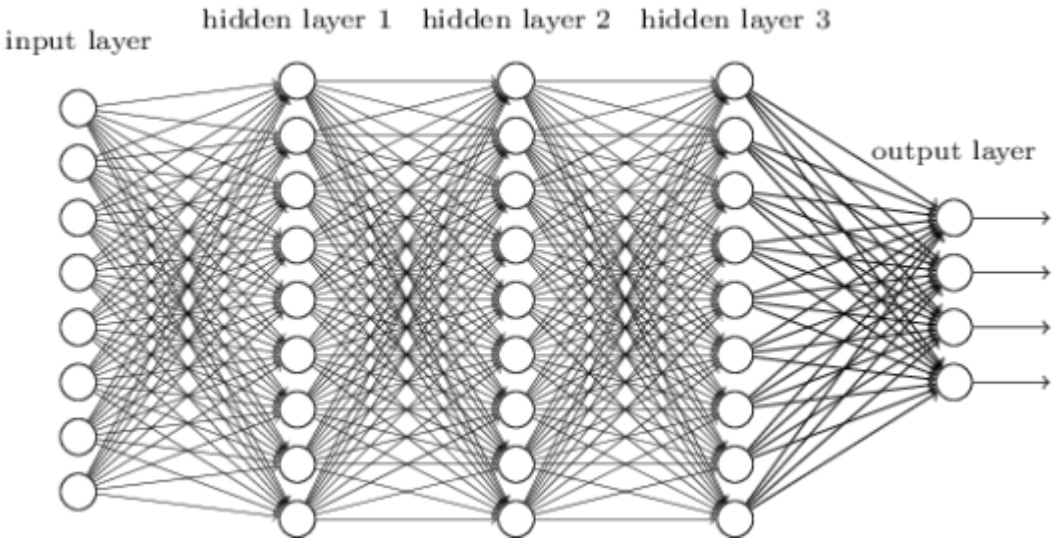
Feature Detection for Digital Images Using Machine Learning Algorithms and Image Processing. Tian, R.; Daigle, H.; Jiang, H.; 2018

Deep Learning



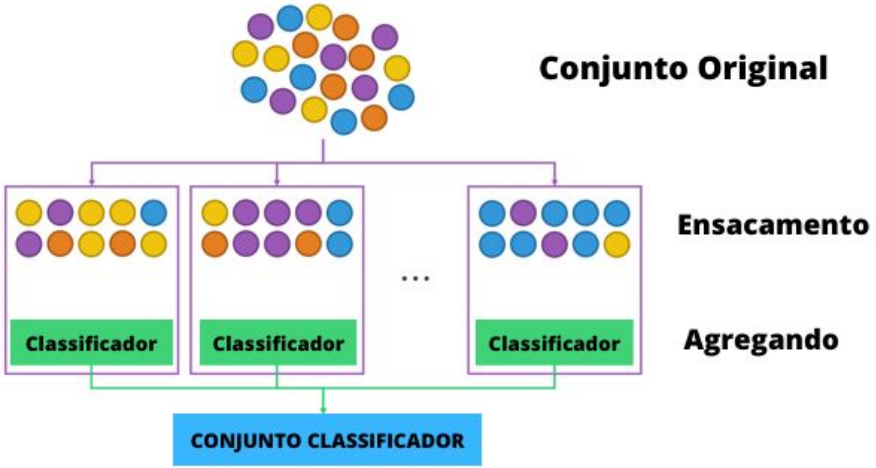
https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

Deep Learning



https://cetax.com.br/o-que-e-deep-learning/

Ensemble



$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

Adaptado de Singhal (2020) por Sousa, A. R. (2023)

Ferramentas

Aprendizagem Não Supervisionada

Aprendizagem Não Supervisionada

Localização	Quartos	Garagem	...	Grupo
Copacabana	2	Não	...	?
Copacabana	1	Sim	...	?
Centro	2	Não	...	?
Barra da Tijuca	3	Sim	...	?
...	...	...	...	...

Aprendizagem Não Supervisionada

Localização	Quartos	Garagem	...	Grupo
Copacabana	2	Não	...	?
Copacabana	1	Sim	...	?
Centro	2	Não	...	?
Barra da Tijuca	3	Sim	...	?
...	...	...	...	...

Aprendizagem Não Supervisionada

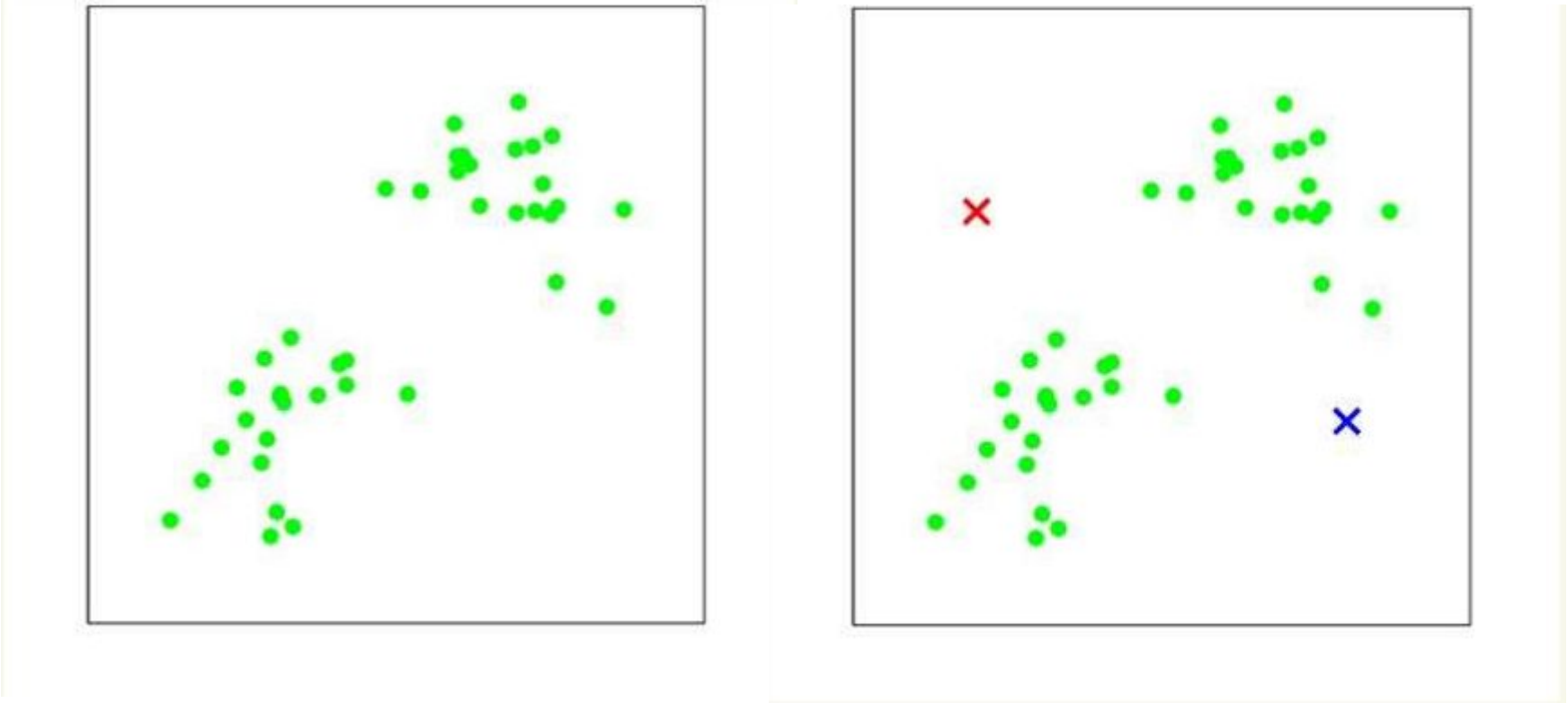
Localização	Quartos	Garagem	...	Grupo
Copacabana	2	Não	...	?
Copacabana	1	Sim	...	?
Centro	2	Não	...	?
Barra da Tijuca	3	Sim	...	?
...	...	...	...	...

Aprendizagem Não Supervisionada

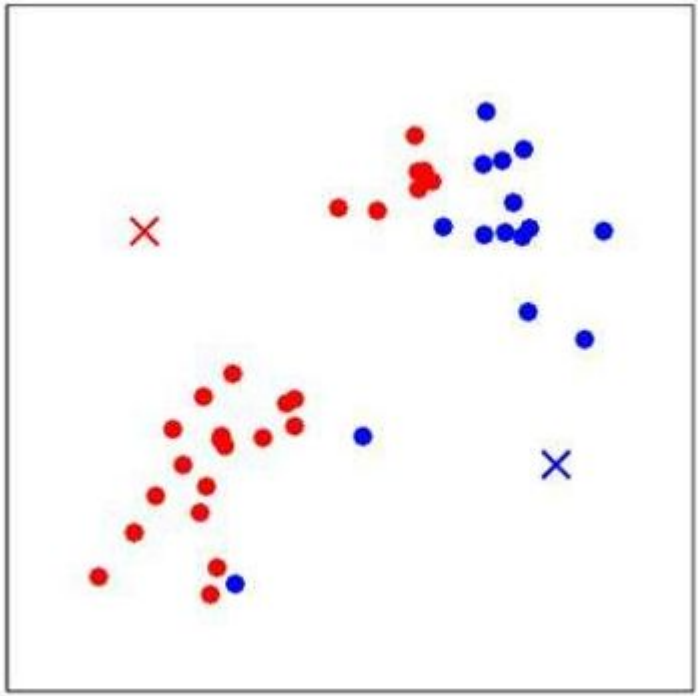
Localização	Quartos	Garagem	...	Grupo
Copacabana	2	Não	...	?
Copacabana	1	Sim	...	?
Centro	2	Não	...	?
Barra da Tijuca	3	Sim	...	?
...	...	...	...	...

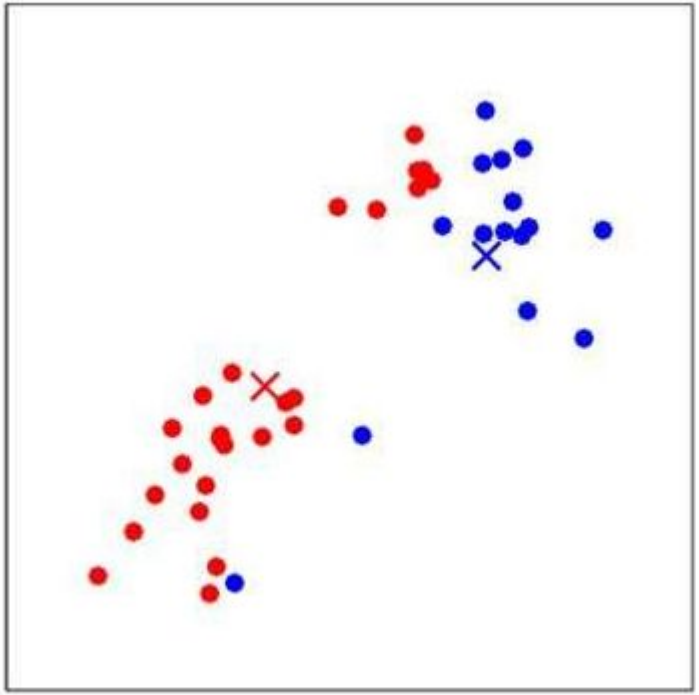
h( )

Aprendizagem Não Supervisionada



Aprendizagem Não Supervisionada



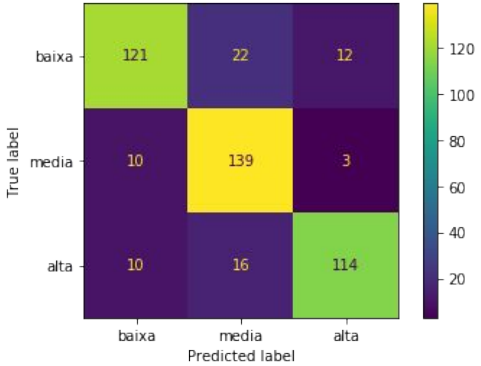


Avaliação de modelos

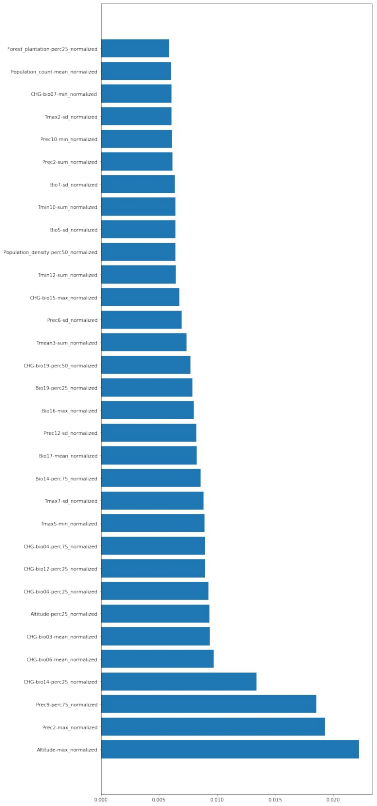
Avaliação de modelos

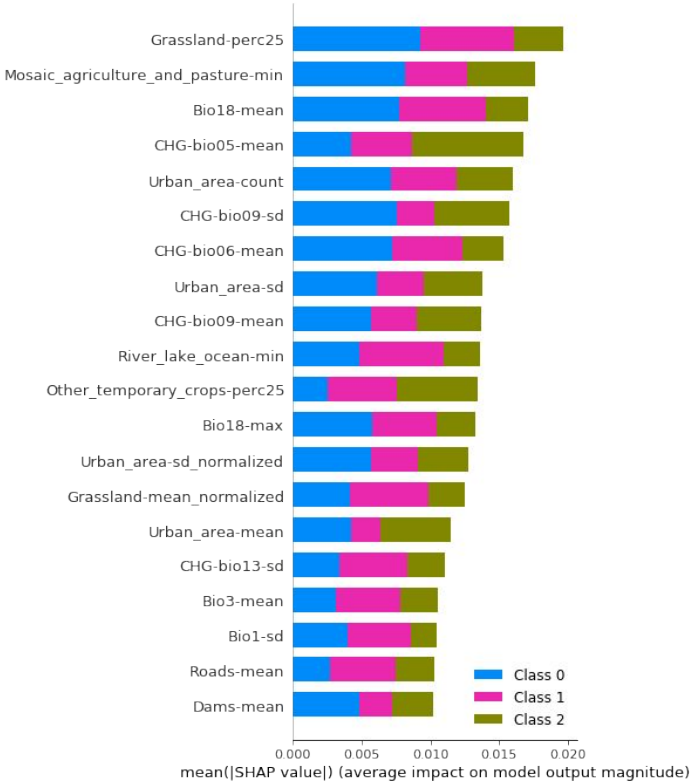
```
In [26]: print(classification_report(y_test, y_pred, labels=[0, 1, 2]))
```

	precision	recall	f1-score	support
0	0.78	0.81	0.80	155
1	0.74	0.86	0.80	152
2	0.87	0.69	0.77	140
accuracy			0.79	447
macro avg	0.80	0.79	0.79	447
weighted avg	0.80	0.79	0.79	447



Avaliação de modelos





E os métodos “generativos”?

Métodos Generativos

- Ganharam fama “estratosférica” com o lançamento de ferramentas como o ChatGPT

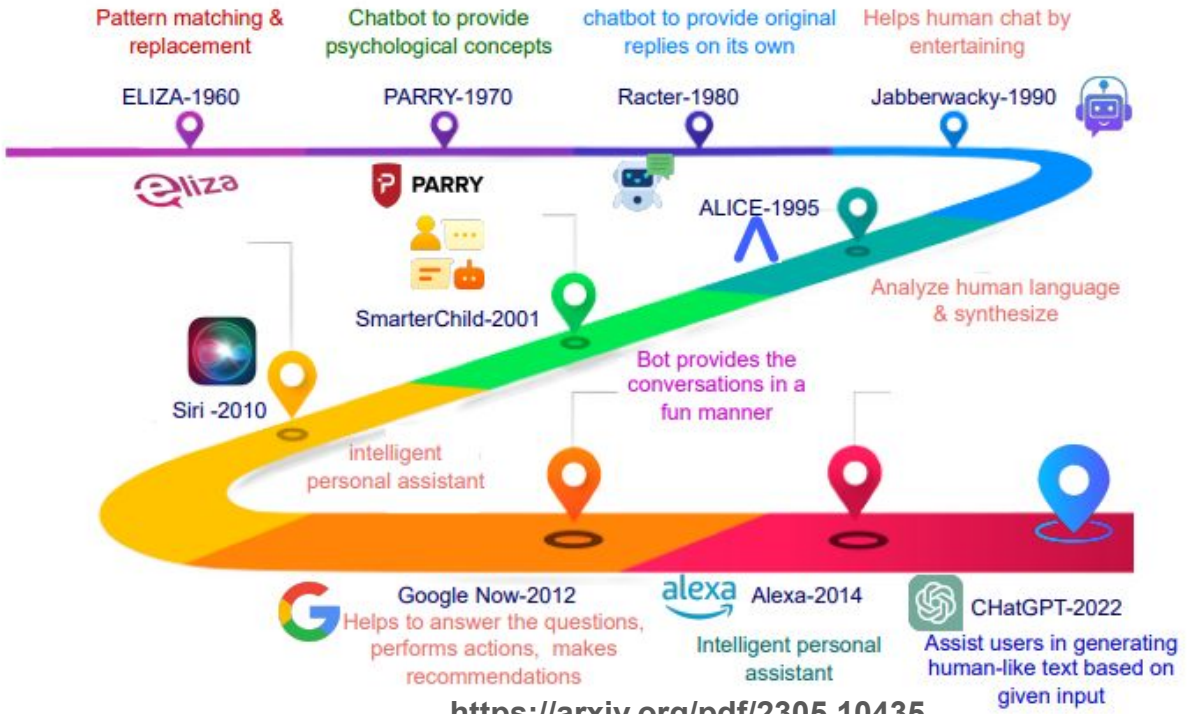
- Qual o princípio?

- **Por que** não usar ou **como** não usar?

Métodos Generativos

- Ganharam fama “estratosférica” com o lançamento de ferramentas como o ChatGPT

Mas essa estrada não é tão nova assim…

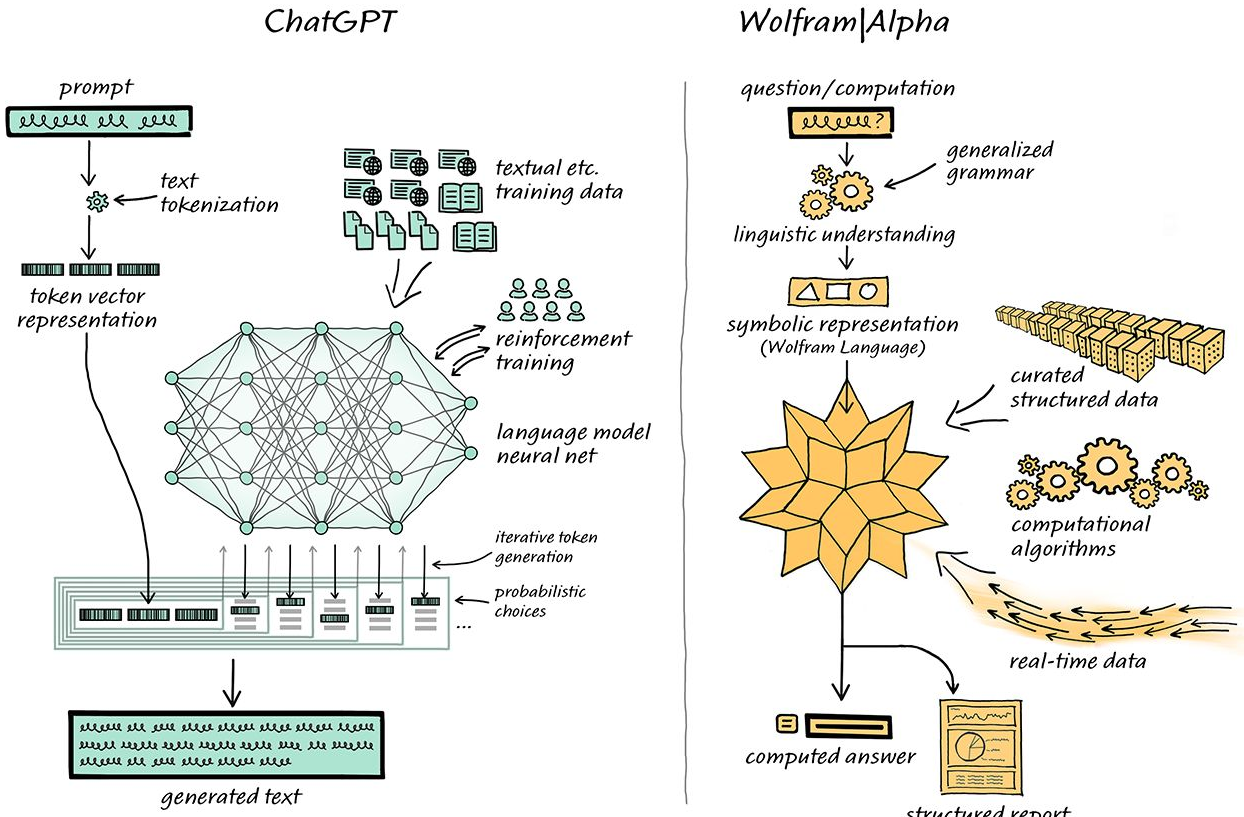


https://arxiv.org/pdf/2305.10435

Métodos Generativos

- Qual o princípio?

Uma combinação de muitas técnicas.



Métodos Generativos

- **Por que** não usar ou **como** não usar?

Science & Society | 1 June 2023 | TRANSPARENT PROCESS

The (mis)use of ChatGPT in science and education: Turing, Djerassi, "athletics" & ethics

David Shaw, Peter Morfeld, and Thomas Erren | AUTHOR INFORMATION

EMBO rep (2023) 24: e57501 | https://doi.org/10.15252/embr.202357501

Métodos Generativos

- **Por que** não usar ou **como** não usar?

Science & Society | 1 June 2023 | TRANSPARENT PROCESS

The (mis)use of ChatGPT in science and education: Turing, Djerassi, "athletics" & ethics

David Shaw, Peter Morfeld, and Thomas Erren | AUTHOR INFORMATION

EMBO rep (2023) 24: e57501 | https://doi.org/10.15252/embr.202357501

ChatGPT is fun, but not an author

H. HOLDEN THORP Authors Info & Affiliations

SCIENCE · 26 Jan 2023 · Vol 379, Issue 6630 · p. 313 · DOI: 10.1126/science.adg7879

143.874 60

Métodos Generativos

- **Por que** não usar ou **como** não usar?

Science & Society | 1 June 2023 |

TRANSPARENT PROCESS

The (mis)use of ChatGPT in science and education: Turing, Djerassi, "athletics" & ethics

David Shaw, Peter Morfeld, and Tl

EMBO rep (2023) 24: e57501 | https://dc

Should researchers really be worried about ChatGPT?

Creado por Richard de Grijs | 9 de Febrero de 2023 | Artificial Intelligence

Share on

ChatGPT is fur., but not an author.

H. HOLDEN THORP Authors Info & Affiliations

SCIENCE · 26 Jan 2023 · Vol 379, Issue 6630 · p. 313 · DOI: 10.1126/science.adg7879

143.874 60

Comentários finais

Por que merecem tanta atenção?

- Permite trilhar caminhos na descoberta de conhecimento
- Produz soluções factíveis para problemas até então sem soluções
- Mecanismos altamente paralelizáveis (CPU e GPU)
- Além de permitir a integração com especialistas na formulação/avaliação
    - Aplicação intrinsecamente interdisciplinar

Por que merecem tanta atenção?

- Permite trilhar caminhos na descoberta de conhecimento
- Produz soluções factíveis para problemas até então sem soluções
- Mecanismos altamente paralelizáveis (CPU e GPU)
- Além de permitir a integração com especialistas na formulação/avaliação
  - Aplicação intrinsecamente interdisciplinar

Mas... não é a solução de todos os seus problemas!

Princípios da Aprendizagem de Máquina

- O modelo mais simples que ajusta ao dados é também o mais plausível (Navalha de Occam)

- Se o dado é amostrado de maneira ”enviesada” (biased way), o aprendizado terá um resultado similarmente ”enviesado”

- Se o conjunto de dados afetou qualquer etapa do processo de aprendizagem, sua capacidade de avaliar o resultado foi comprometida

Referências

Referências

- Psicologia da Aprendizagem. Campos, D. M. S; 1971
- **Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kuma, 2018**
- Prediction of protein function using a deep convolutional neural network ensemble. Zacharaki, E. I, 2017
- **Learning from Data. Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.; 2012**
- **Pattern Recognition and Machine Learning. Bishop, C. M.; 2006**
- Cross validation: https://scikit-learn.org/stable/modules/cross_validation.html
- Phishing Websites Classification using Hybrid SVM and KNN Approach. Taha, A.A; 2017

Referências

- The Kernel Trick in Support Vector Classification:
  - https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f
- Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kuma, 2018:
  - https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap4_ann.pdf
- RMDL: Random Multimodel Deep Learning for Classification, Kowsari, K. et al., 2018
- New machine learning and physics-based scoring functions for drug discovery. Guedes, I. A.; Barreto, A. M. S. ; Marinho, D. ; Krempser, E.; Kuenemann, M. A. ; Sperandio, O.; Dardenne, L. E.; Miteva, M. A. Scientific Reports, v. 11, 2021
- In silico studies of the HIV-1 integrase: mutational patterns, resistance mechanisms, and strategies to search for new drug candidates. Tese de Doutorado em Biologia Computacional e Sistemas - IOC|Fiocruz. Lucas de Almeida Machado. Orientadora: Ana Carolina Ramos Guimarães. 2020.

Referências

- Cursos on-line:
  - Andrew Ng: https://pt.coursera.org/learn/machine-learning
  - Yaser Abu-Mostafa: https://work.caltech.edu/telecourse
  - Eduardo Krempser: https://youtube.com/playlist?list=PLIJhet1J-_l82wpFzn2ZOyCIKy-FotuND&si=dHYGGzV7z8OL4r2E

Prática

- Python 3 (https://www.python.org/downloads)
  - pip / pip3 (https://pypi.org/project/pip)
  - Scikit-Learn (https://scikit-learn.org)
  - Jupyter (https://jupyter.org/install)
  - Pandas (https://pandas.pydata.org)
  - Numpy (https://numpy.org)
  - Matplotlib (https://matplotlib.org)
- Weka
  - https://www.cs.waikato.ac.nz/ml/weka

Prática

- Python 3 (https://www.python.org/downloads)
  - pip / pip3 (https://pypi.org/project/pip)
  - Scikit-L
  - Jupyter
  - Pandas
  - Numpy
  - Matplotlib (https://matplotlib.org)
- Weka
  - https://www.cs.waikato.ac.nz/ml/weka

pip3 install jupyter

pip3 install pandas

pip3 install scikit-learn

pip3 install matplotlib

11 EMMSB 2024

Conceitos e aplicações da aprendizagem de máquina

Eduardo Krempser
eduardo.krempser@fiocruz.br

Matheus Müller
matheus.mullerps@gmail.com

FIOCRUZ

PIBSS
Plataforma Institucional
Biodiversidade e Saúde Silvestre