

COC800

Introdução à Ciência de Dados

Aprendizado de máquina



- Contextualização e preliminares
- O aprendizado: tipos e tarefas
- Estratégias de avaliação de desempenho
- Preparação de dados



Motivação

NETFLIX

Populares na Netflix



Em alta



Séries excêntricas dos EUA aclamadas pela crítica



ScienceDirect

Journals & Books

Register Sign in

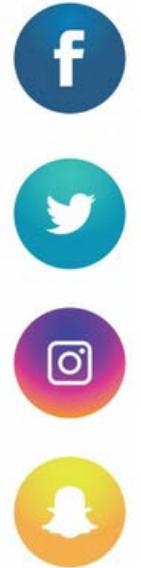
amazon.com.br Olá Selecionar o endereço Todos

Todos Prime Mais Vendidos Ofertas do Dia Livros Moda da Amazon Eletrônicos Lançamentos Computadores + 1

Clientes que compraram este item também compraram



Inspirado pelo seu histórico de navegação

Outline

Abstract

Keywords

1. Introduction

2. Background

3. Publication recommender system

4. Experimental results and analysis

5. Conclusions

Acknowledgments

Appendix A. Supplementary materials

Research Data

References

ELSEVIER

Knowledge-Based Systems

Volume 157, 1 October 2018, Pages 1-9

A content-based recommender system for computer science publications

Donghui Wang ^a, Yanchun Liang ^{a, b, c}, Dong Xu ^{a, c}, Xiaoyue Feng ^a, Renchu Guan ^{a, b}

Show more

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.knosys.2018.05.001>

Get rights and content

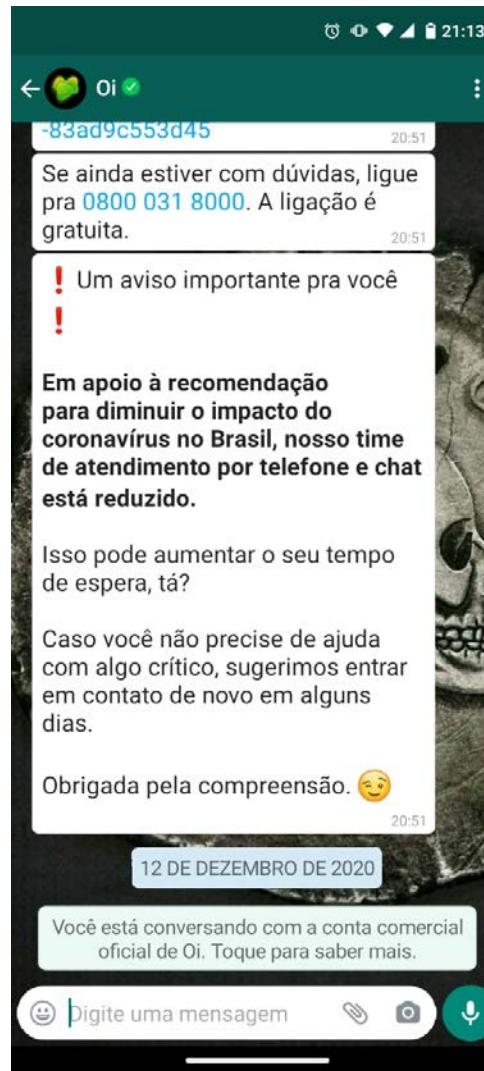
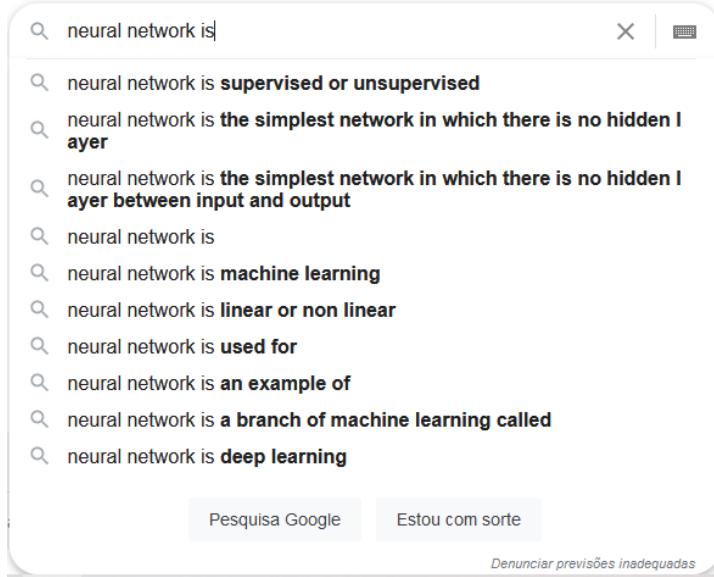
ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

Send a message...

Motivação

Google



Português ▼

Minha terra tem palmeiras,
onde canta o sabiá.
Os pássaros que aqui gorjeiam,
não gorjeiam como lá.

Gmail

+ Escrever

- 100
- ^ Menos
- 💬 Chats
- 🕒 Programados
- ✉️ Todos os e-mails
- ⚠️ Spam
- ⌧ Lixeira

292

Inglês ▼

My land has palm trees,
where the thrush sings.
The birds that chirp here,
don't chirp like there.

My homeland has many palm-trees
and the thrush-song fills its air;
no bird here can sing as well
as the birds sing over there.

Cadernos de Literatura em Tradução, n. 11, p. 353-372

Aprendizado de máquina	Nauczanie maszynowe
------------------------	---------------------

Uczencie maszynowe

Motivação



ChatGPT

RO

You

Em que o Gemini da Google é melhor que o ChatGPT?

< 2 / 2 >

ChatGPT

O Google Gemini é um sistema de geração de linguagem desenvolvido pela Google que utiliza uma abordagem semelhante ao modelo de linguagem de grande escala (como o ChatGPT). No entanto, é difícil comparar diretamente o Gemini com o ChatGPT, pois as informações detalhadas sobre o Gemini podem não estar disponíveis publicamente até o meu último treinamento em janeiro de 2022.

Dito isso, pode haver algumas áreas em que o Gemini da Google pode se destacar em relação ao ChatGPT:



Em que o ChatGPT é superior ao Gemini?

Mostrar rascunhos ▾



Existem áreas onde o ChatGPT se destaca em relação ao Gemini. Aqui estão alguns exemplos:

- **Criatividade:** Alguns usuários consideram o ChatGPT ligeiramente mais criativo na geração de determinados conteúdos, como convites ou poemas.
- **Acessibilidade:** A versão gratuita do ChatGPT permite mais interações por sessão do que a versão gratuita do Gemini. !

É importante lembrar que essas vantagens são relativas. Veja alguns pontos onde o Gemini pode ser melhor:

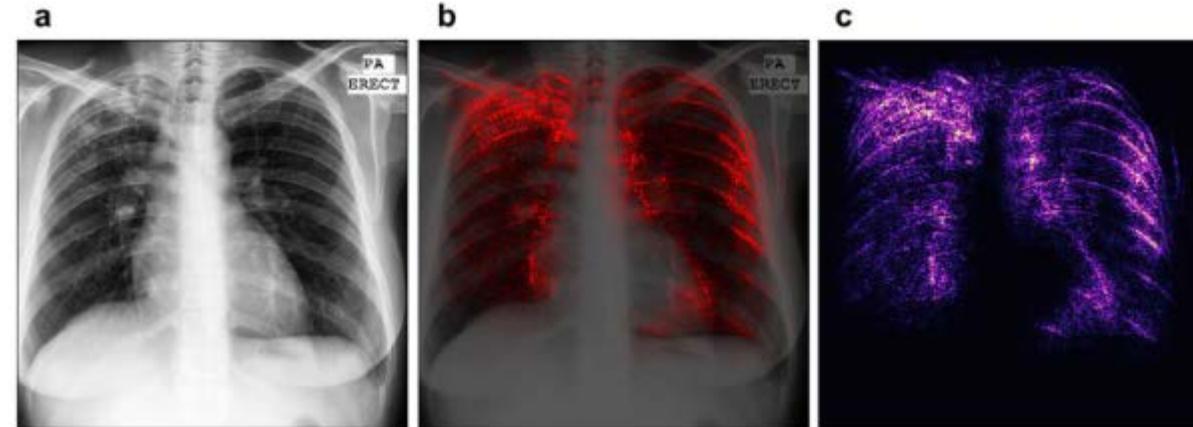
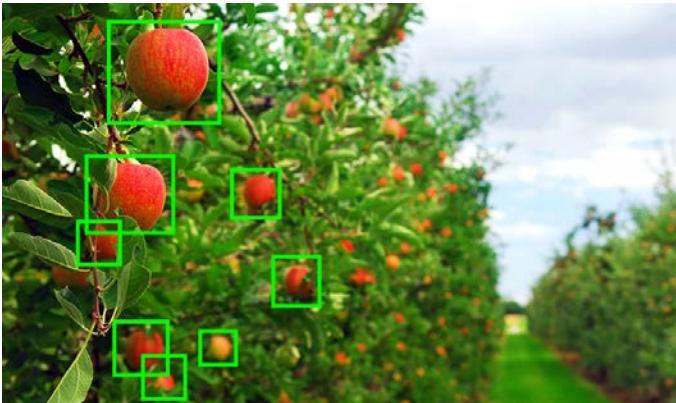
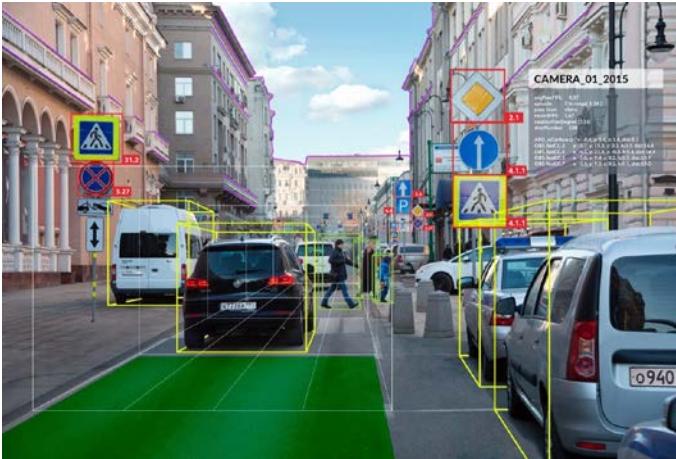
Gemini



RPE
2025

Contextualização e preliminares

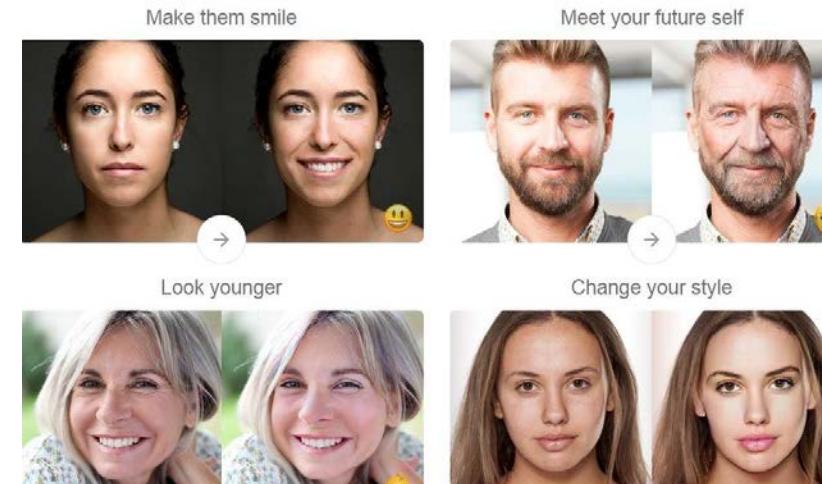
Motivação



Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization

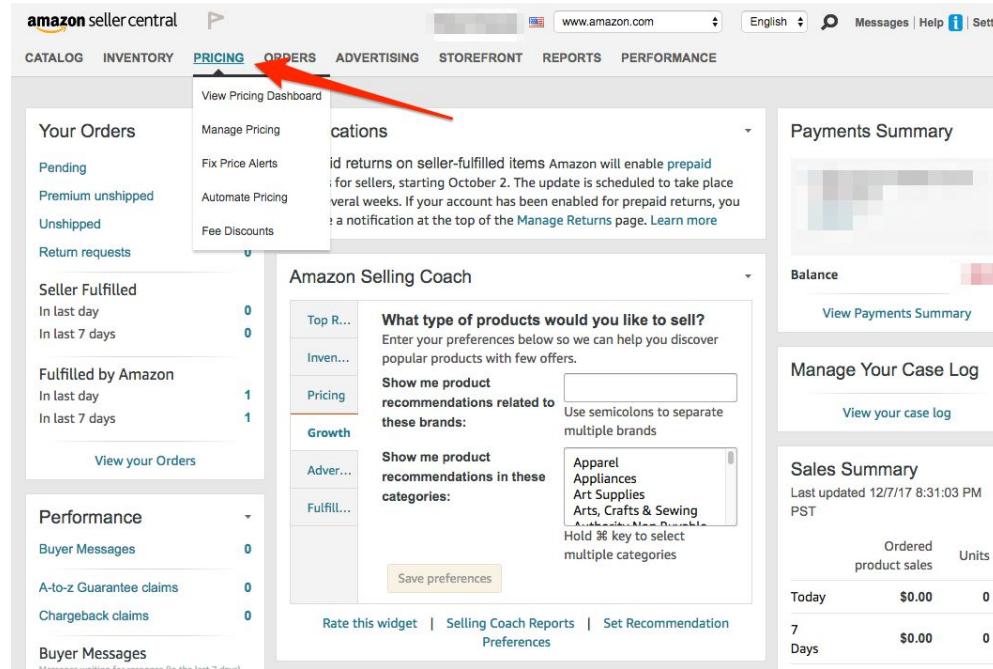
F. Pasa^{1,2}, V. Golikov², F. Pfeiffer^{1,2}, D. Cremers³ & D. Pfeiffer²

SCIENTIFIC REPORTS nature

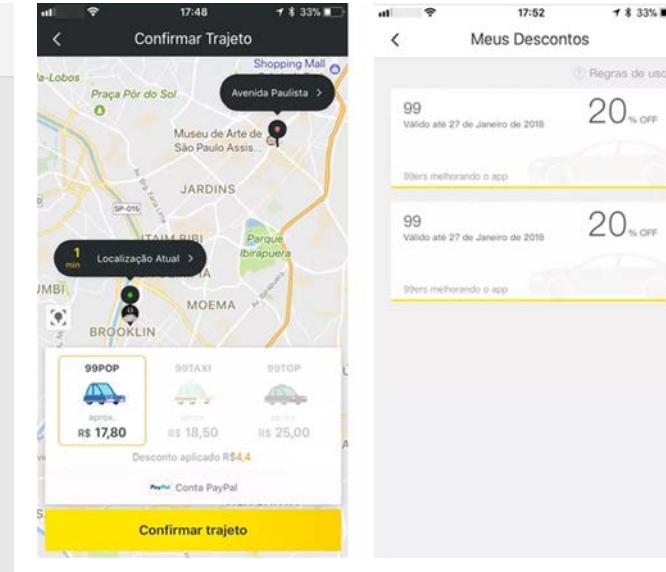


Google caption in 2016: "A baby is eating a piece of paper."

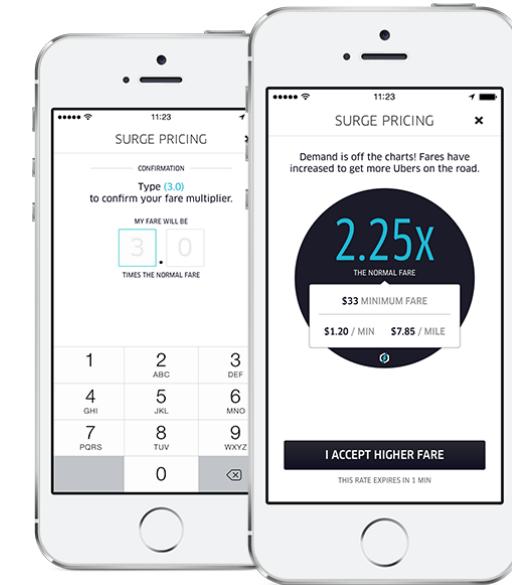
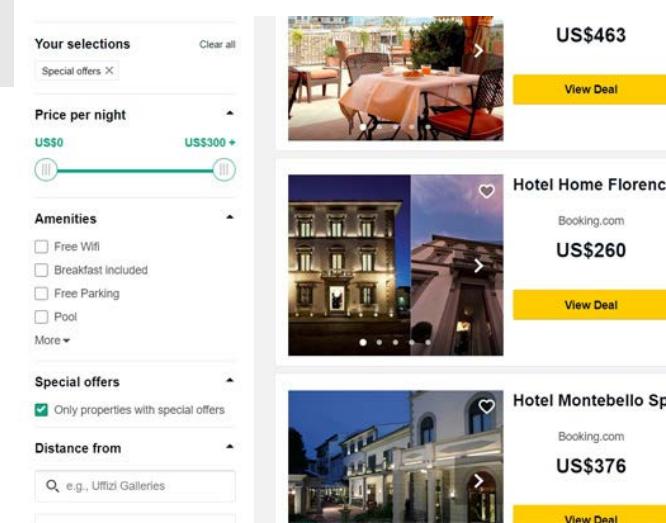
Motivação



The screenshot shows the Amazon Seller Central dashboard. The top navigation bar includes links for Catalog, Inventory, Pricing, Orders, Advertising, Storefront, Reports, and Performance. A red arrow points to the 'PRICING' link. The main content area features sections for Your Orders, Seller Fulfilled, and Fulfilled by Amazon, along with a 'Amazon Selling Coach' widget that suggests product categories to sell.



The first screenshot shows a map-based 'Payments Summary' screen with a 'Confirmar Trajeto' button at the bottom. The second screenshot shows a list of discounts with validity dates and percentage off.

The screenshot shows a search results page for hotel deals. It includes filters for price per night (US\$0 to US\$300), amenities like free wifi and breakfast included, and special offers. Two specific deals are highlighted: Hotel Home Florence at US\$260 and Hotel Montebello Splend at US\$376.

Rappi

ifood

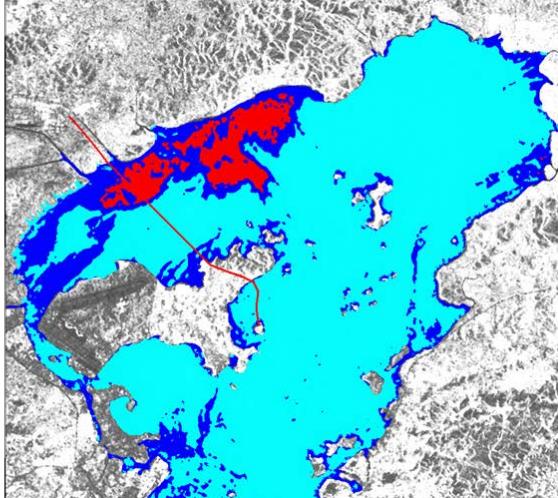
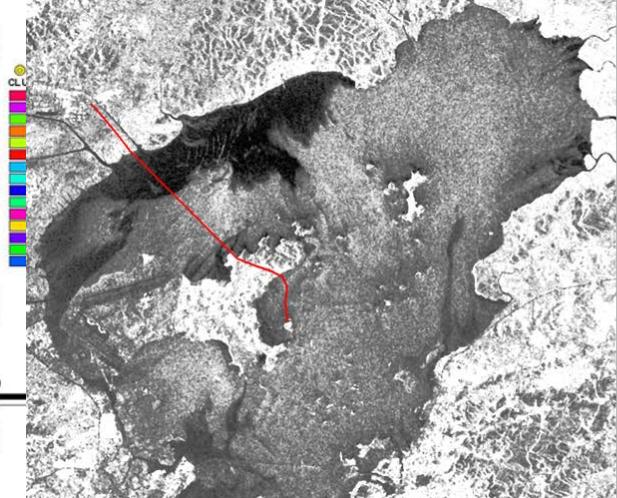
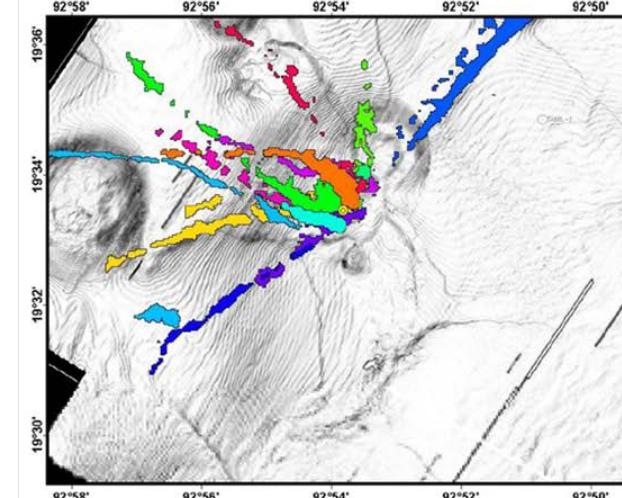
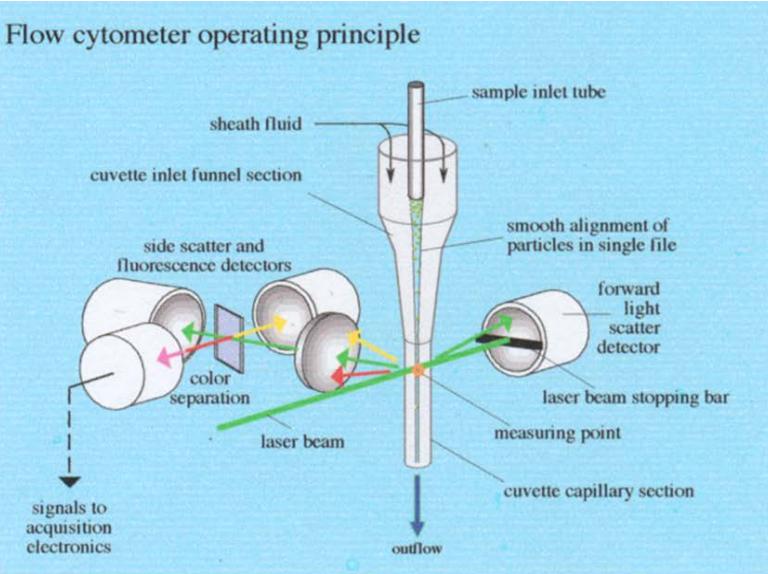
Glovo  **Uber Eats**



RPE
2025

Contextualização e preliminares

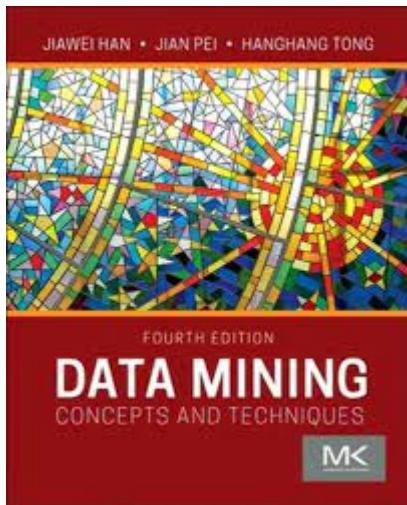
Motivação



Mineração de Dados

Um processo evolutivo (Han, Kamber & Tong):

- coleta e formação do BD;
- gerenciamento dos dados;
- análise e compreensão dos dados.



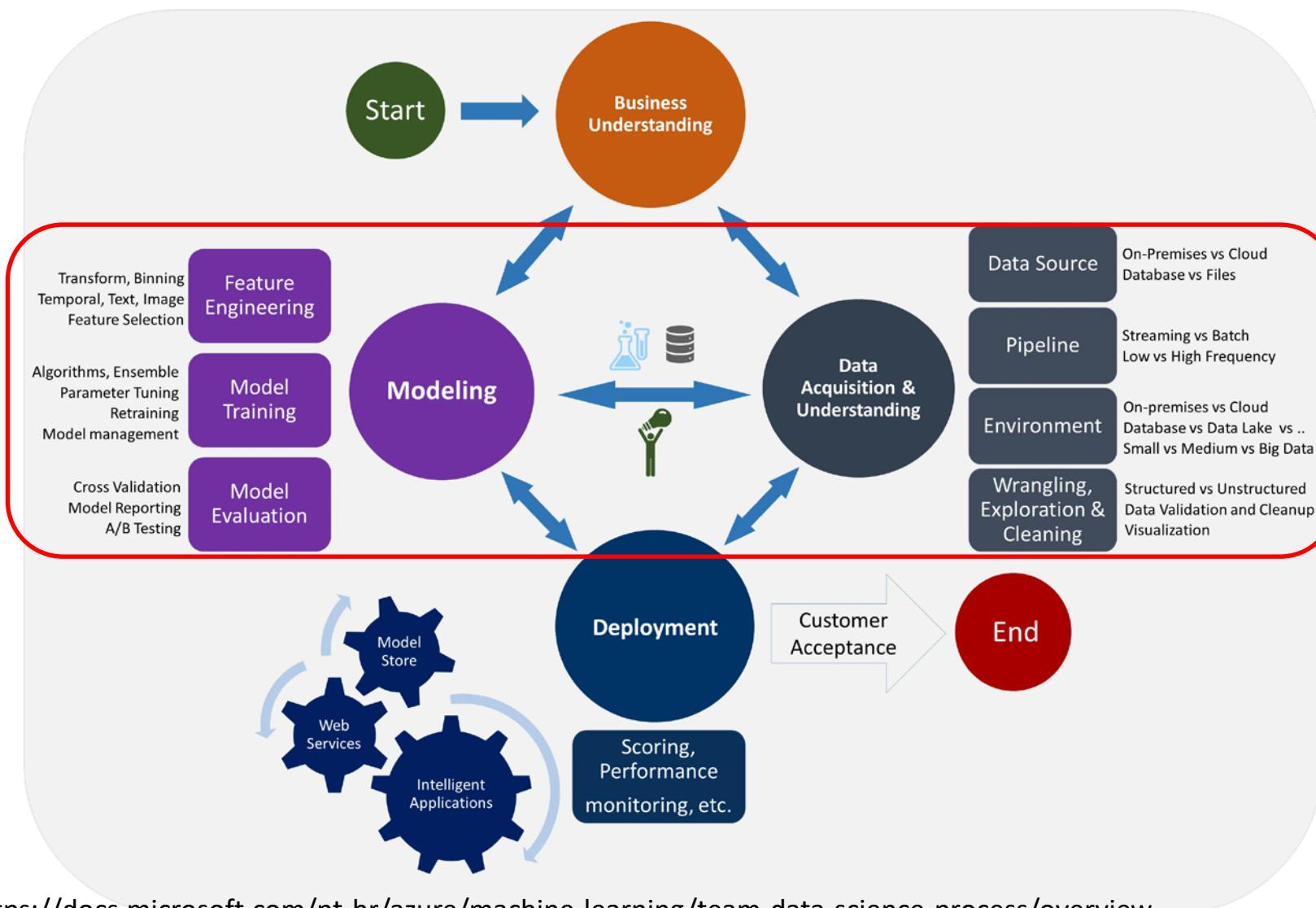
Relevância:

- incapacidade humana;
- melhoria na tomada de decisão;
- apresentação do conhecimento sob ângulos diversos;
- descobrir relacionamentos não-triviais.

extração



Mineração de Dados





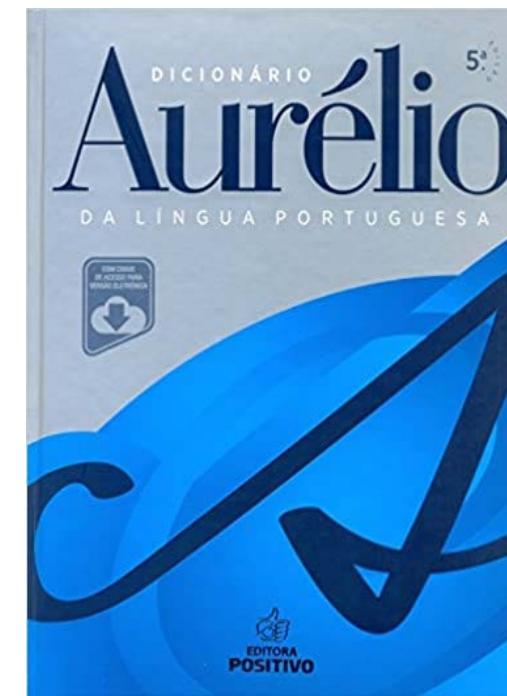
RPE
2025

O aprendizado: tipos e tarefas

Uma definição de aprendizado de máquina

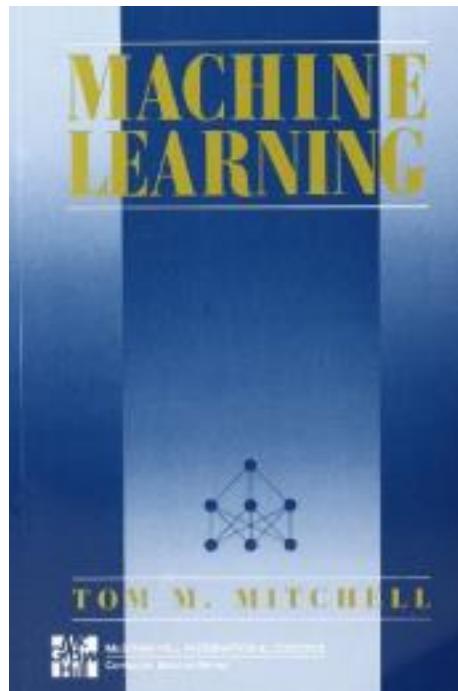
É a área da Inteligência Artificial que pesquisa o desenvolvimento de programas capazes de aprender como realizar uma dada tarefa a partir de suas experiências com os dados.

1. Tomar conhecimento;
2. Reter na memória, mediante o estudo, a observação ou a experiência;
3. Tornar-se apto ou capaz de alguma coisa, em consequência de estudo, observação, experiência, advertência, etc.



“Dada uma medida de desempenho D , diz-se que um programa P aprende uma classe de tarefas T a partir de uma experiência E , se seu desempenho nas tarefas de T melhoram (conforme D) com a experiência E .

Tom Mitchell
Machine Learning, 1997



Direto x indireto

Supervisionado, não-supervisionado, semi-supervisionado, por reforço

Uma visão compartmentalizada

APRENDIZADO = REPRESENTAÇÃO + AVALIAÇÃO + OTIMIZAÇÃO

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

DOI:10.1145/2347736.2347755

Tapping into the “folk knowledge” needed to advance machine learning applications.

BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning



As principais atividades



Atividades descritivas

Mineração de dados



Atividades preditivas

Agrupamento

Associação

Classificação

Regressão

Aprendizado não supervisionado

Aprendizado supervisionado



RPE
2025

O aprendizado: tipos e tarefas

Aprendizado não supervisionado: agrupamento

O objetivo é o desenvolvimento de um modelo capaz de segmentar os dados em grupos de objetos de características semelhantes.



antes

depois

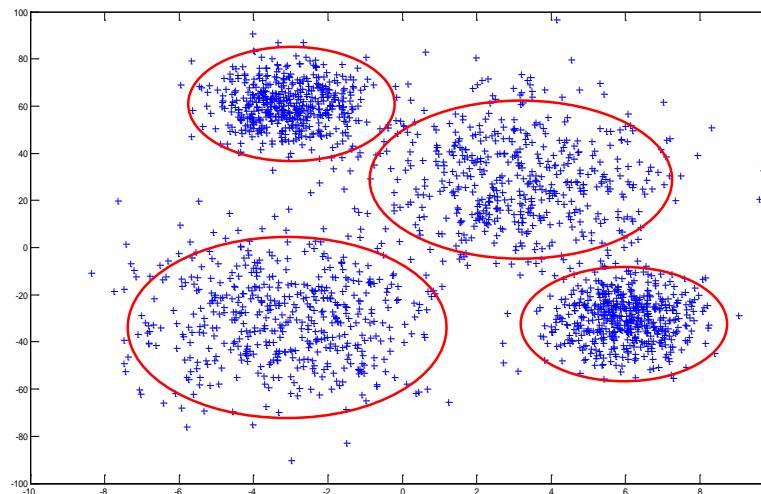


RPE
2025

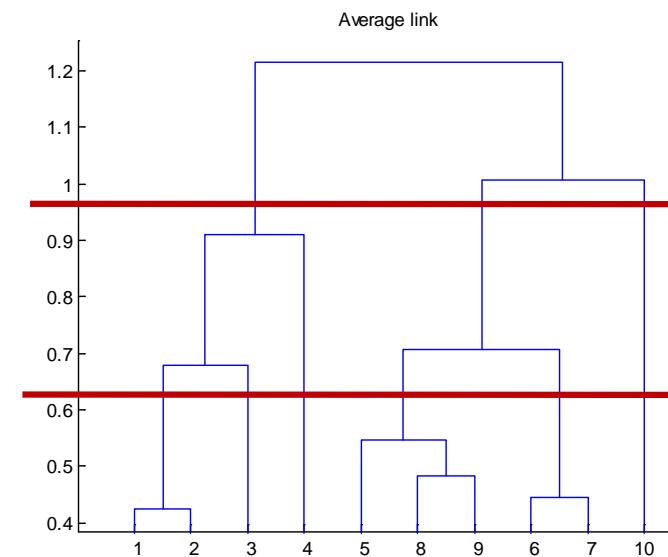
O aprendizado: tipos e tarefas

Aprendizado não supervisionado: agrupamento

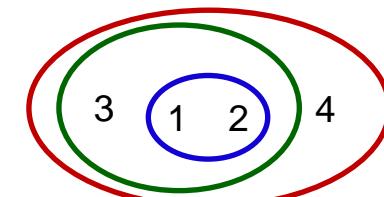
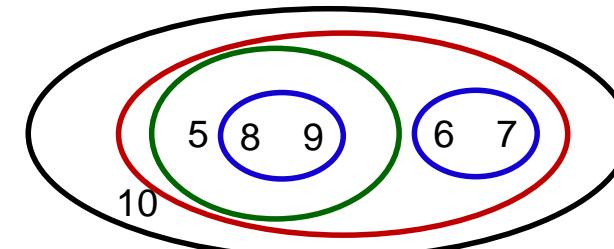
Em geral, não há uma representação “correta”, mas a busca é pela melhor possível.



Particional

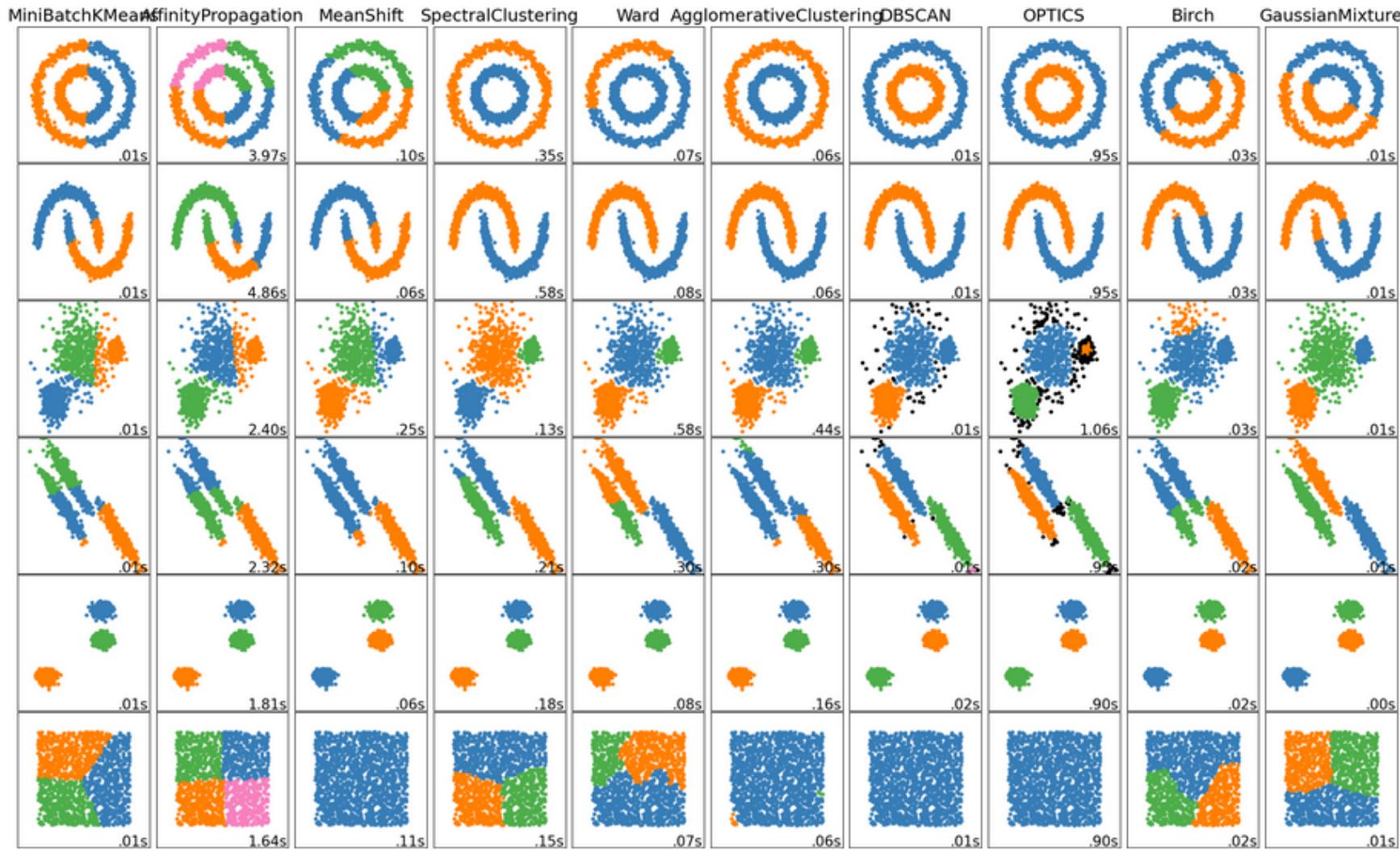


Hierárquico



- Agrupamento de notícias
- Segmentação de clientes
- Localização de documentos
- Agrupamentos genéticos
- Filtros sonoros
- Análise de redes sociais

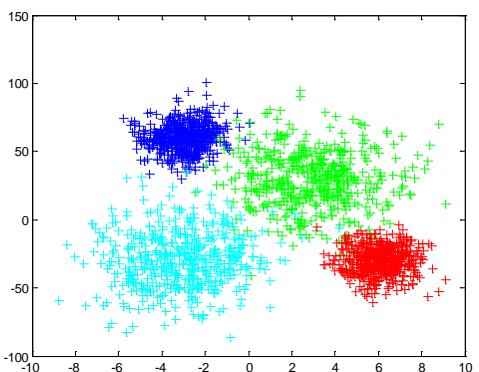
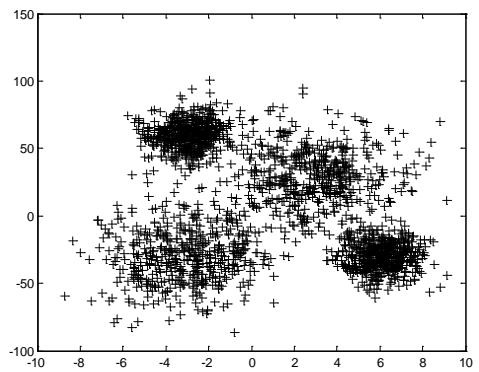
Aprendizado não supervisionado: agrupamento



A comparison of the clustering algorithms in scikit-learn

Espera-se:

homogeneidade interna
heterogeneidade externa

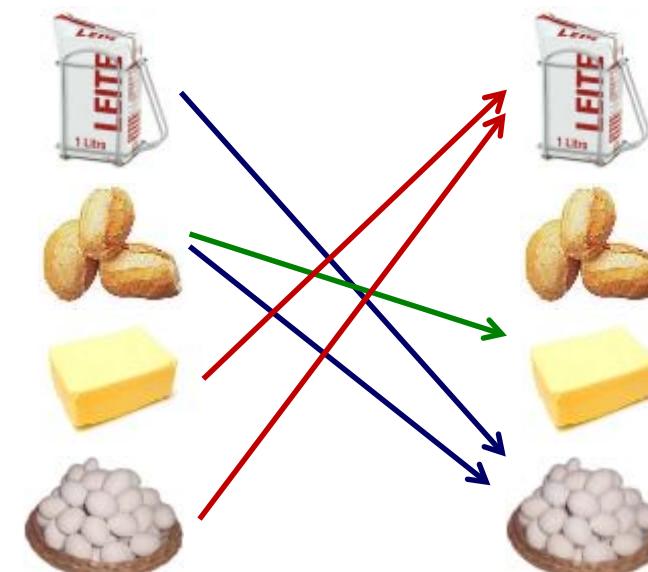




Aprendizado não supervisionado: associação

O objetivo é encontrar relações que indicam ocorrência concomitante e relevante entre valores frequentes das variáveis estudadas.

TID	Produtos
1	Leite, pães, ovos
2	Açúcar, pães
3	Manteiga, pães
4	Açúcar, leite, pães
5	Leite, manteiga
6	Manteiga, pães
7	Leite, manteiga, tomates
8	Leite, manteiga, pães, ovos
9	Leite, manteiga, pães



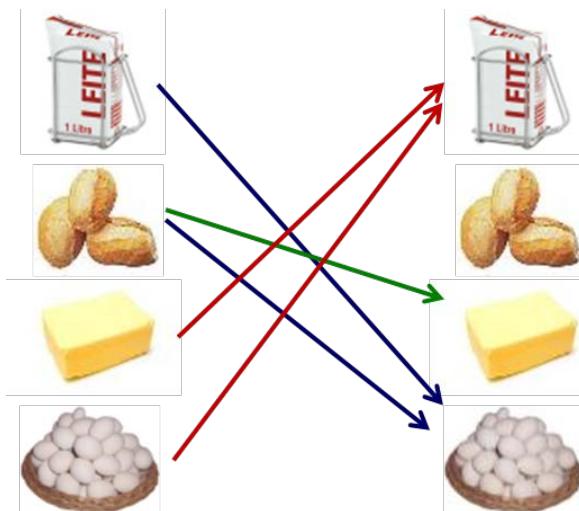
- Identificação de sintomas relacionados;
- Oferta de produtos; sistemas de recomendação;
- Mecanismos de busca na Internet;
- Análise de compras.



Aprendizado não supervisionado: associação

O resultado da associação é uma base de regras que estabelecem relações aproximadas de causa e consequência entre os objetos estudados.

antecedente
SE objeto₁ E ... E objeto_N, ENTÃO objeto_j em que $j \notin \{1, \dots, N\}$
consequente



Regra A: Se compra LEITE e PÃES, então compra OVOS.

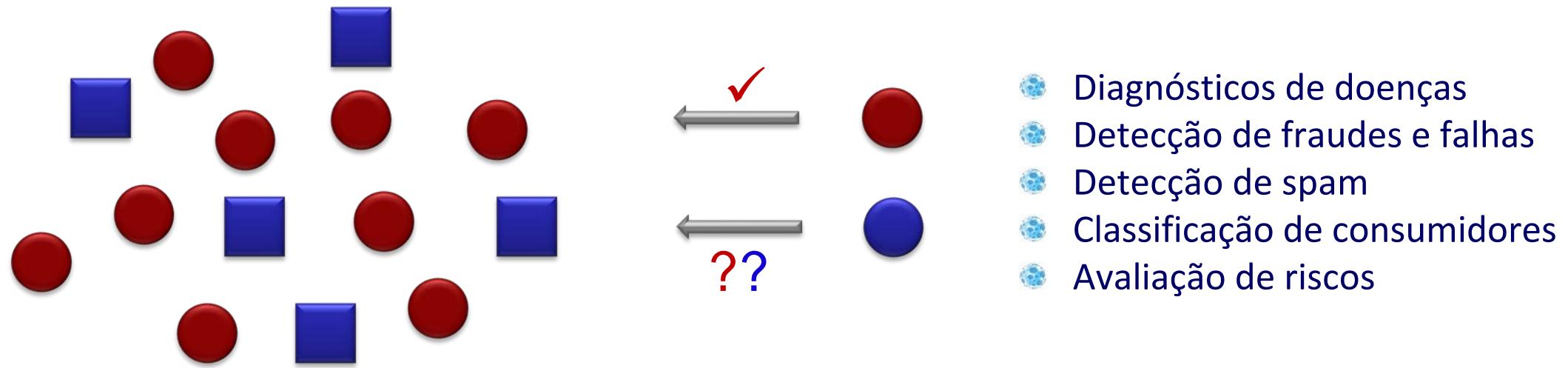
Regra B: Se compra PÃES, então compra MANTEIGA.

Regra C: Se compra MANTEIGA e OVOS, então compra LEITE.



Aprendizado supervisionado: classificação

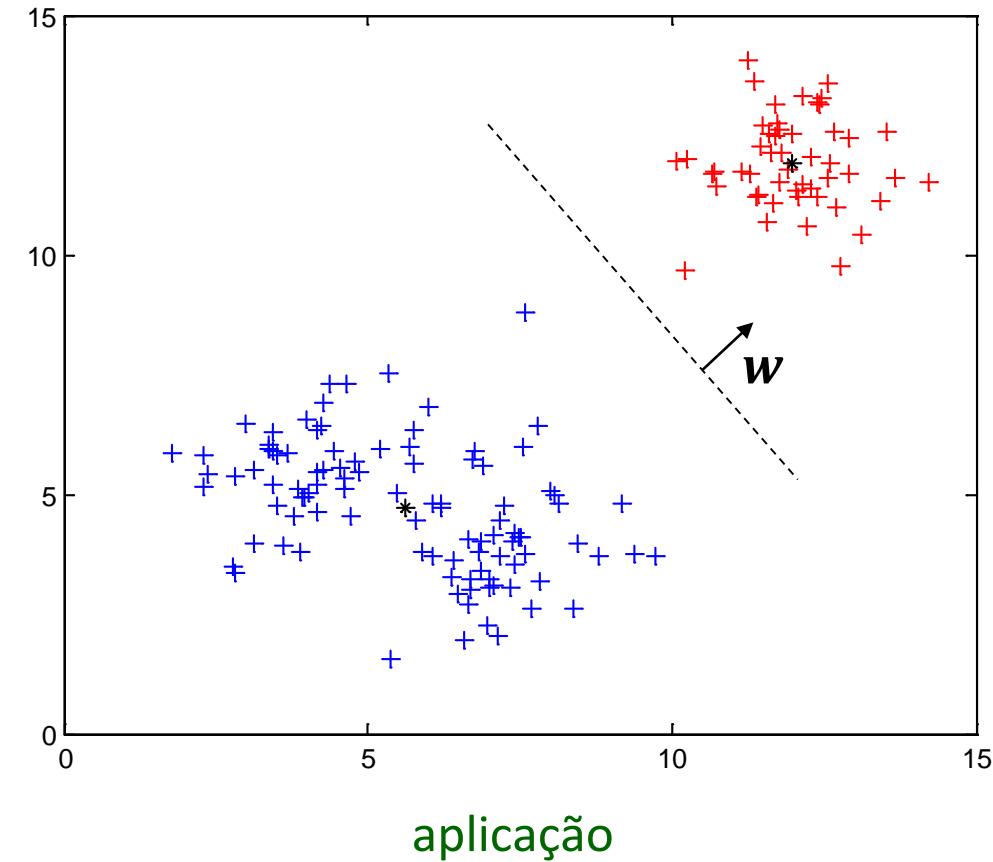
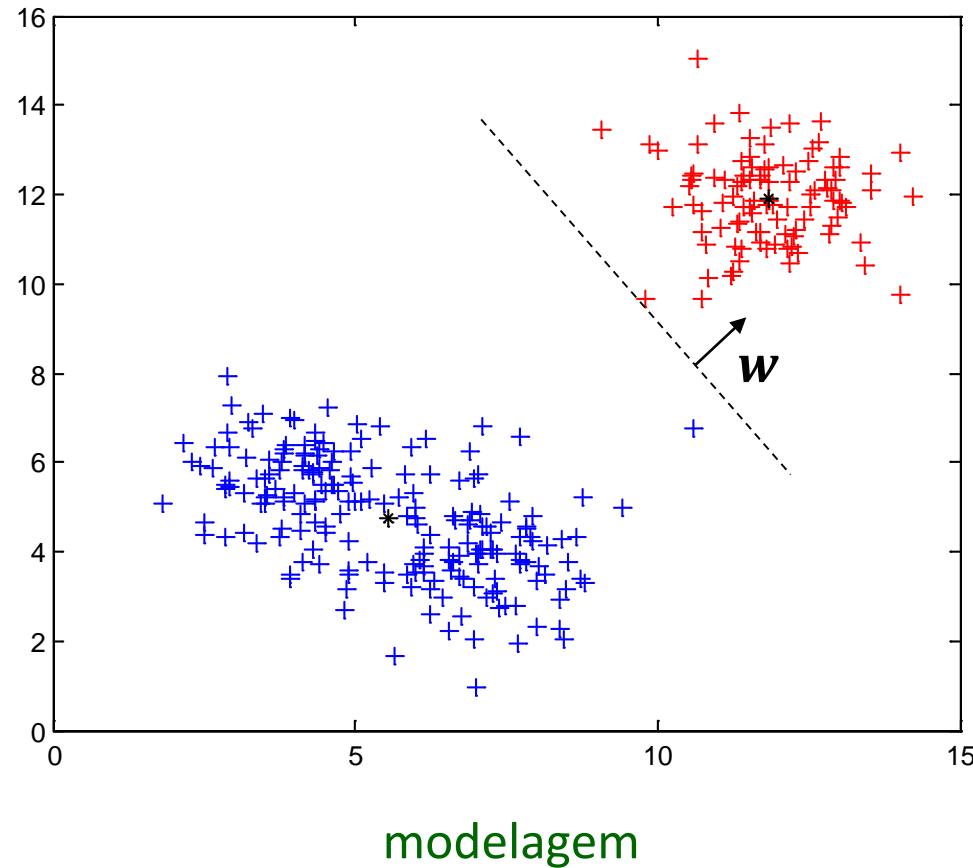
O objetivo é o desenvolvimento de um modelo capaz de identificar a que categoria pertence um novo objeto.



Supõe-se que os objetos de uma categoria ou classe compartilham características comuns, semelhantes, e que essas sejam suficientes para decidir se novos objetos podem ou não ser alocados a ela.



Aprendizado supervisionado: classificação

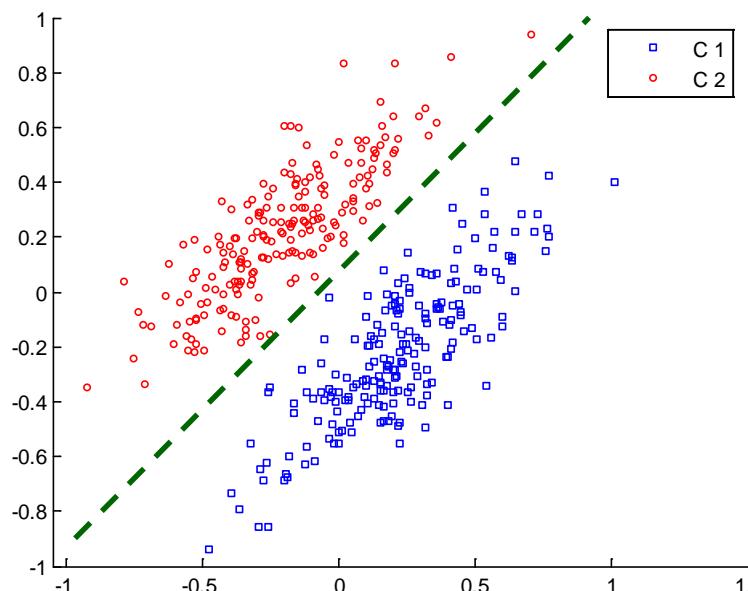




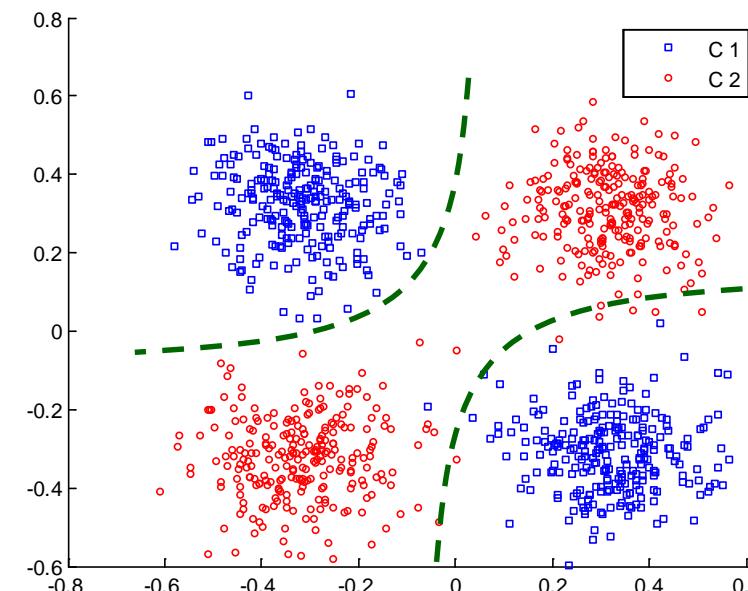
Aprendizado supervisionado: classificação

Um problema é **separável** se existem uma ou mais superfícies de decisão capazes de separar completamente suas classes em suas respectivas regiões de decisão. Se existe uma superfície única e linear (hiperplano), o problema é **linearmente separável**.

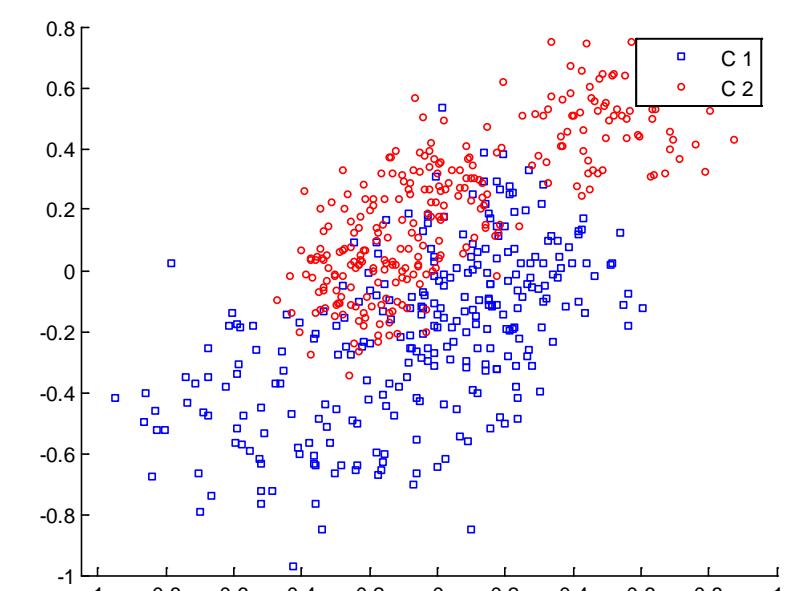
Quanto maior a separabilidade das classes, menor a complexidade do problema. Em geral, não será possível separar perfeitamente as classes de um problema, linearmente ou não.



Linearmente separável



Não linearmente separável



Não separável

Aprendizado supervisionado: classificação

Tipos de aprendizado:

Passivo: o especialista informa como classificar; basta implementar o que foi relatado.

Ativo: o especialista apenas informa as classes de cada registro e é preciso descobrir como classificar.

Tipos de problemas:

Binária: duas classes.

Múltipla: três ou mais classes
(incluso aqui a multiclassificação)

Tipos de tomada de decisão

Individual: apenas um classificador decide a classe de um registro.

Coletiva: vários classificadores atuam e suas respostas são combinadas por votação:

- maioria;
- combinação linear ponderada.

Three Type of Classification Tasks

YAHOO!
JAPAN

Binary
Classification



- Spam
- Not spam

Multiclass
Classification



- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label
Classification

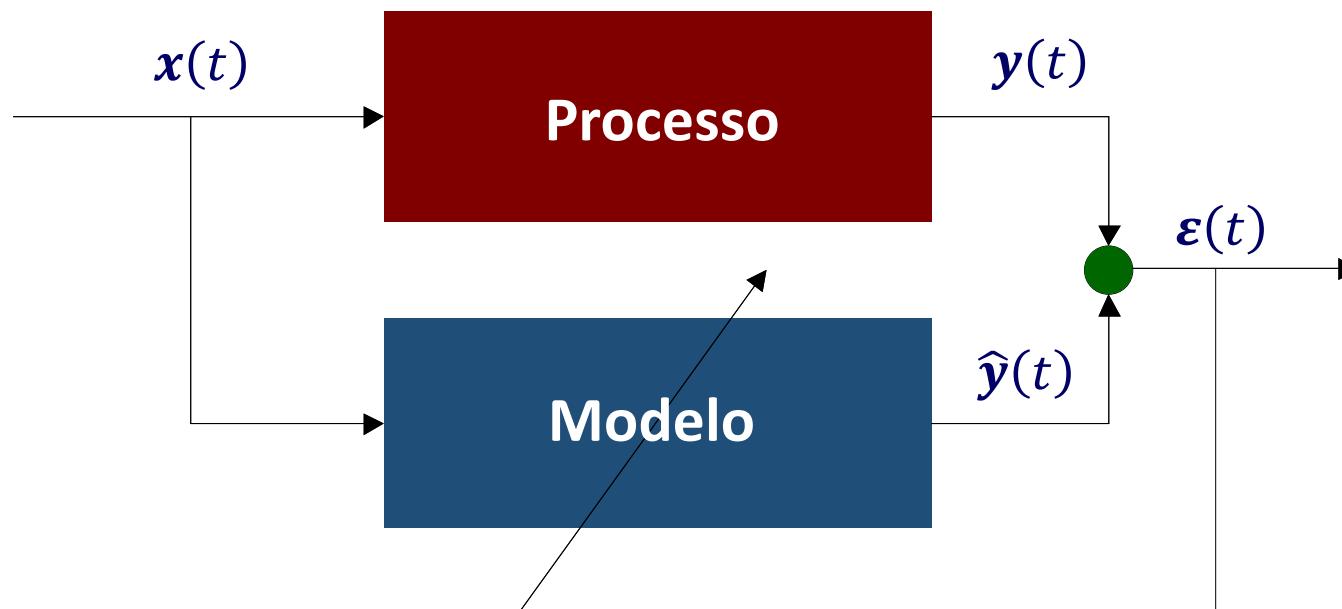


- Dog
- Cat
- Horse
- Fish
- Bird
- ...



Aprendizado supervisionado: regressão

O objetivo é o desenvolvimento de um modelo capaz de identificar o valor mais correto possível para um novo conjunto de entradas.



Qual é a função que produz as saídas?

- Dosagem de medicamentos
- Precificação
- Controle de equipamentos
- Estimativa de emissão de poluentes
- Avaliação de riscos

$y = f(x)$ função real, desconhecida

$\hat{y} = \hat{f}(x)$ função estimada

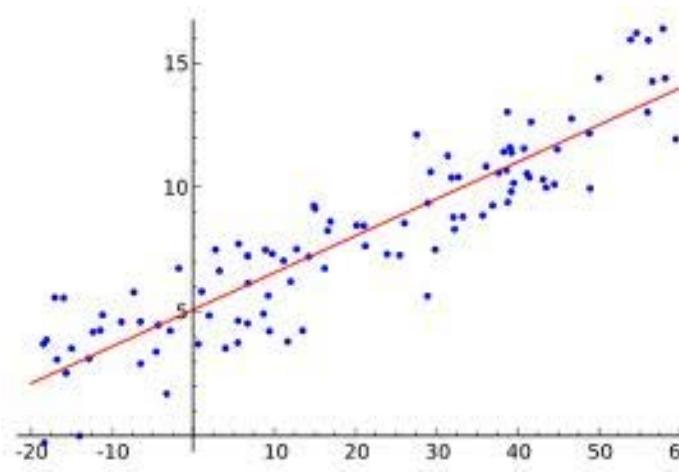
$\epsilon = y - \hat{y}$ erro ou resíduo

RPE
2025

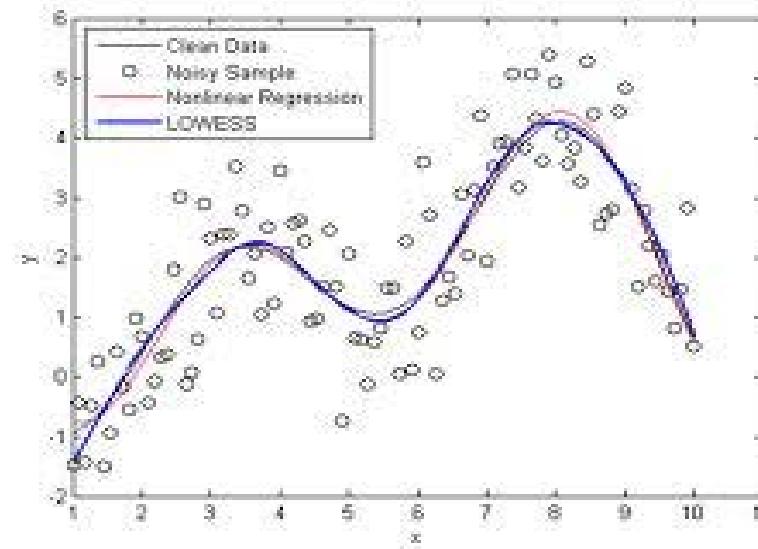
O aprendizado: tipos e tarefas

Aprendizado supervisionado: regressão

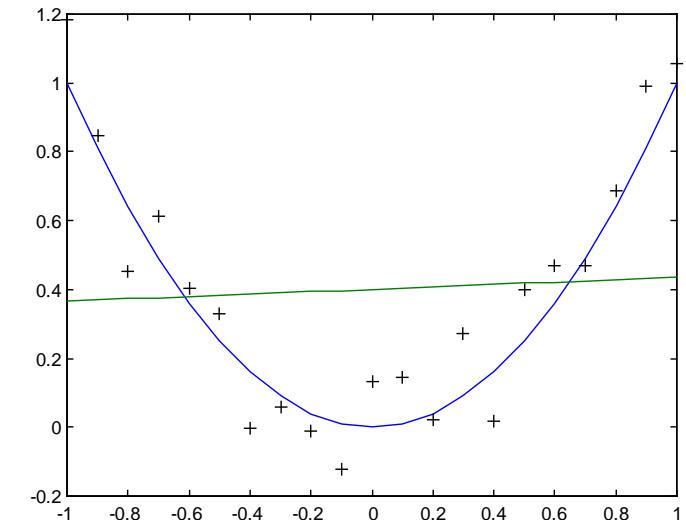
Em geral, os algoritmos para geração de modelos de classificação podem ser utilizados para a regressão também, com pequenas adaptações.



Linear



Não linear



Polinomial



Aprendizado supervisionado: regressão

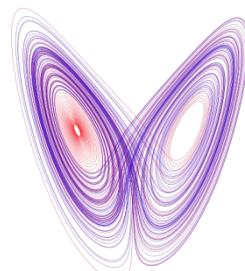
Natureza da informação



Determinístico



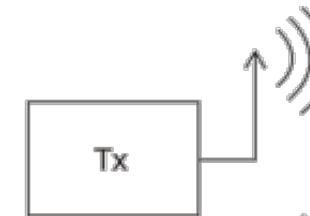
Estocástico



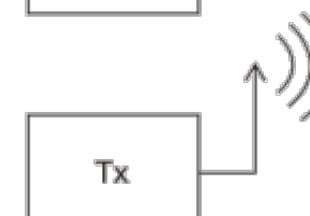
Caótico

Entradas e saídas

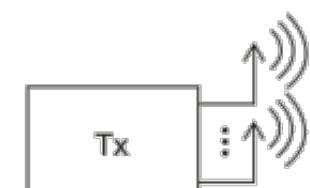
SISO



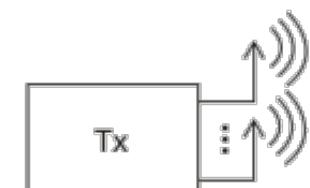
SIMO



MISO



MIMO



Representação do tempo



Estático

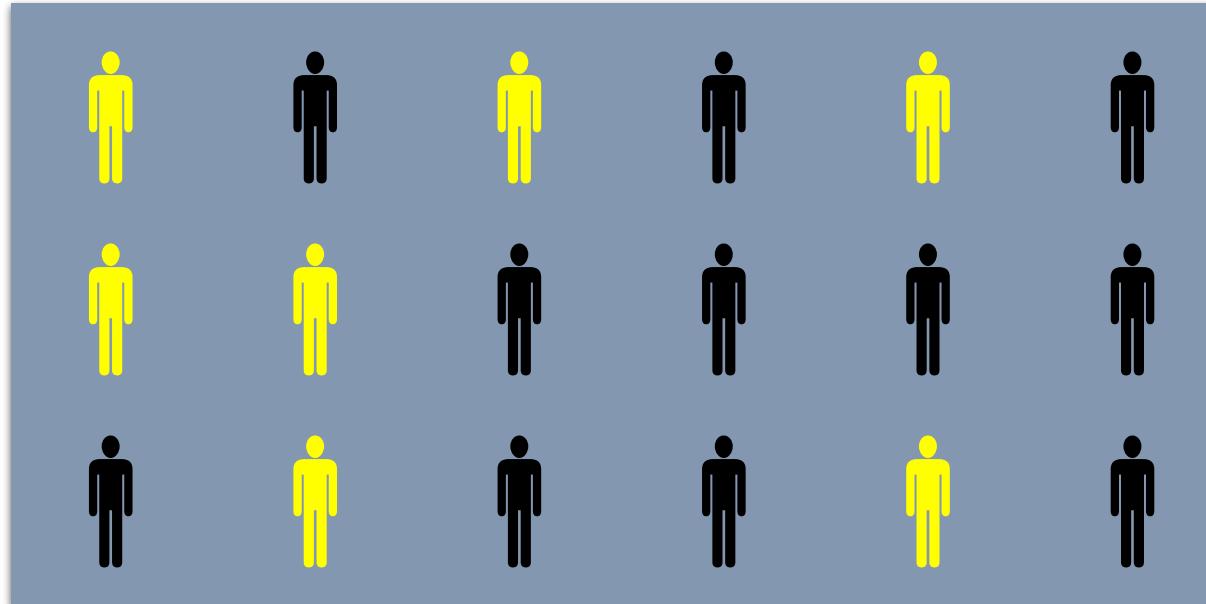


Dinâmico



Seleção de atributos

O objetivo é selecionar os atributos mais relevantes para a construção dos modelos de aprendizado.



Objetivos específicos:

- simplificação de modelos para facilitar a interpretação;
- reduzir tempo de aprendizado;
- evitar a “maldição da dimensionalidade”;
- reduzir o risco de superajuste (*overfitting*).

Suposição:

Os dados possuem atributos irrelevantes ou redundantes



Seleção de atributos

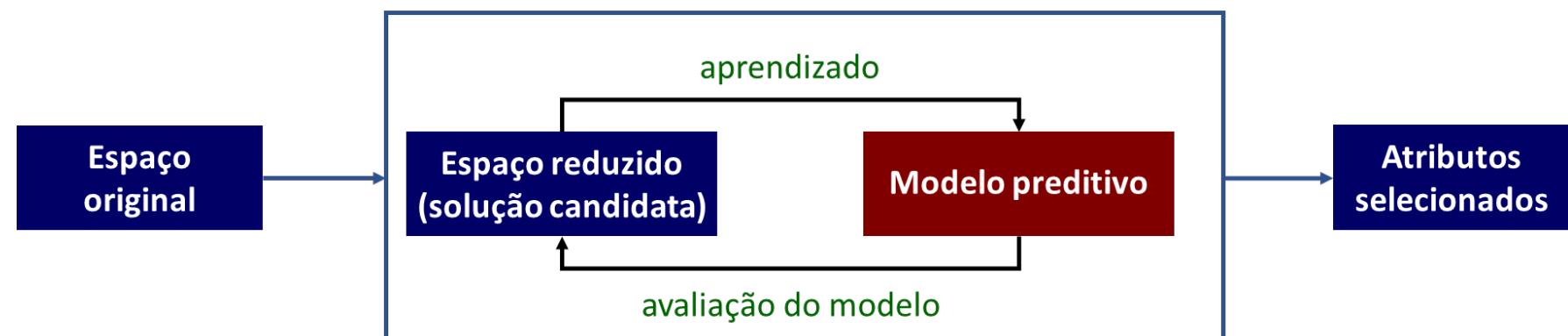
Para selecionar atributos relevantes, utilizar um **algoritmo de busca** por subconjuntos de atributos, que utiliza uma **função de avaliação** para quantificar cada solução candidata.

Tipos:

Filtros



Envoltórios,
Invólucros,
Caixa-preta



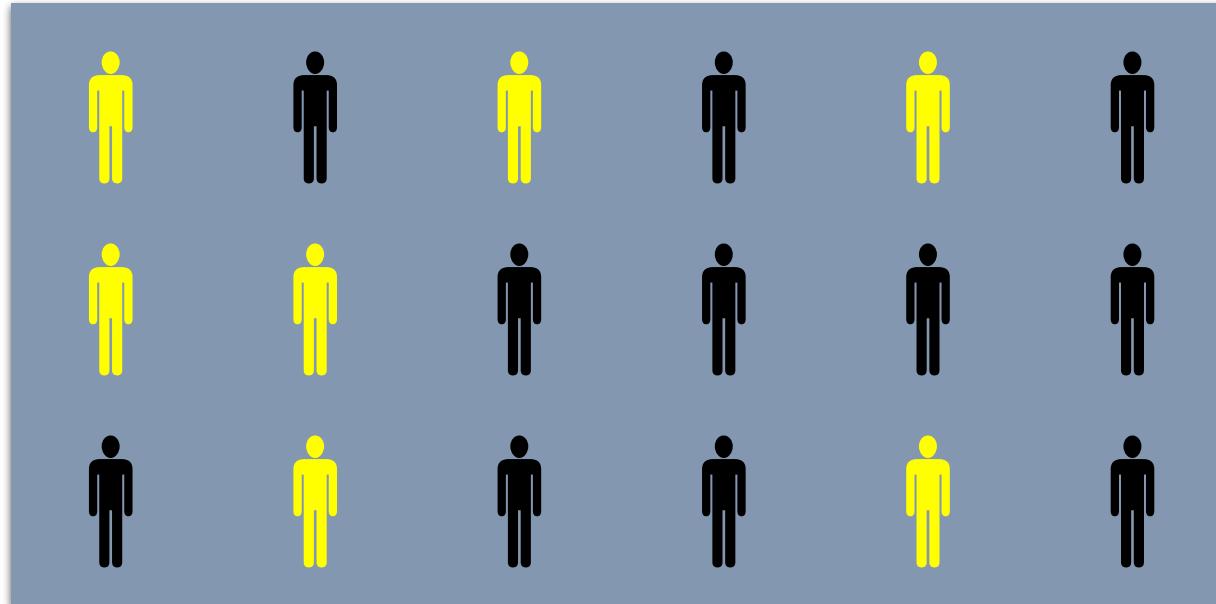
Embutidos





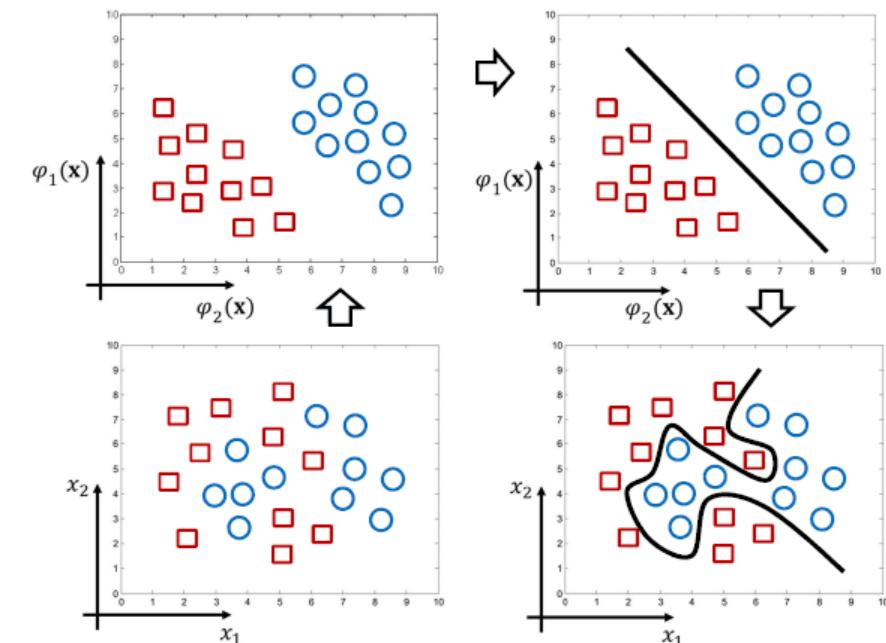
Extração de atributos

O objetivo é produzir atributos relevantes para o aprendizado quando os originais não são suficientes para tal.



Objetivos específicos:

- Produzir atributos informativos e não redundantes para facilitar o aprendizado;
- Reduzir a dimensionalidade preservando as características do conjunto de dados original;
- Facilitar a interpretação de resultados.



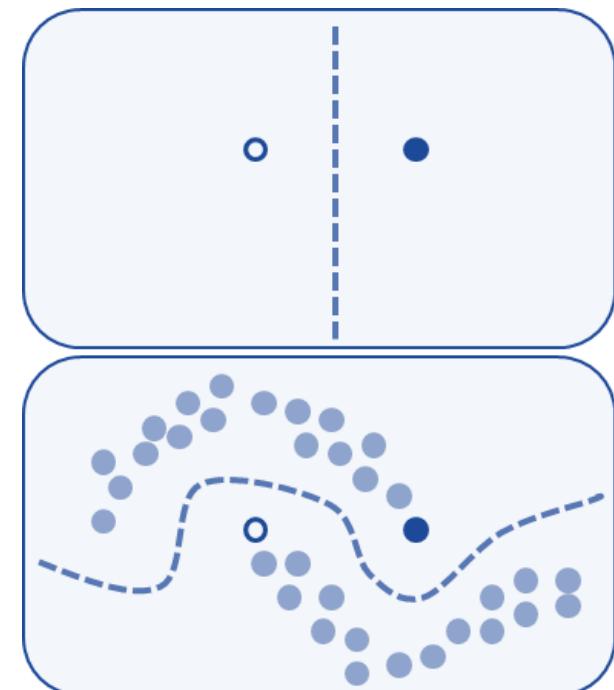
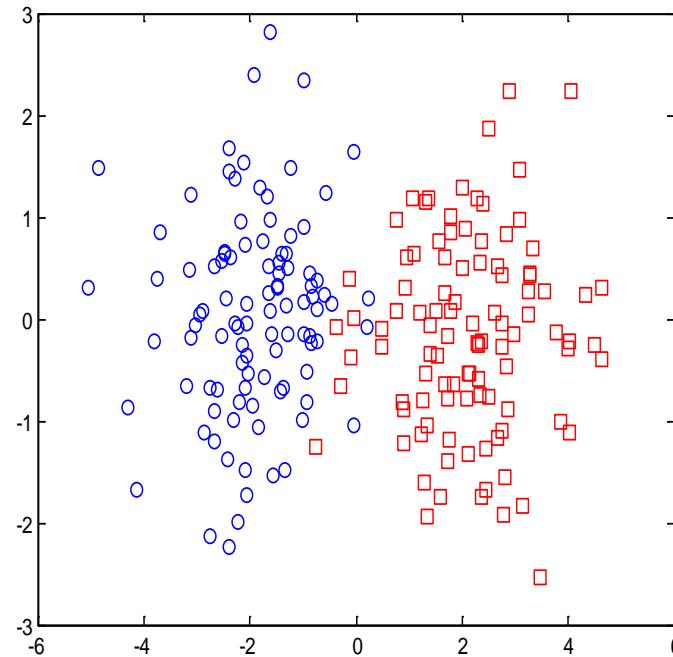
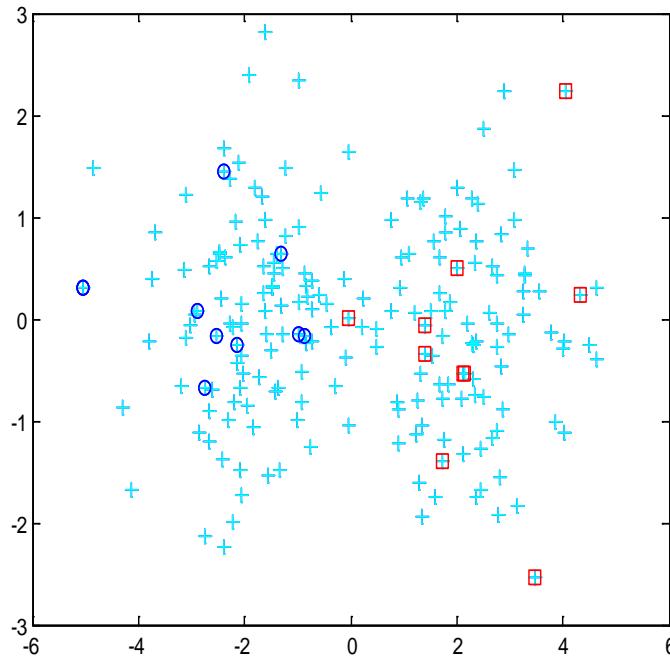


RPE
2025

O aprendizado: tipos e tarefas

Aprendizado semi-supervisionado

O objetivo é melhorar o aprendizado supervisionado com o uso de uma grande quantidade de dados não rotulados, aproveitando uma maior amostragem das variáveis preditivas. Em outras palavras, utiliza (alguns) registros com os valores da variável de resposta e (muitos) registros sem essa variável.

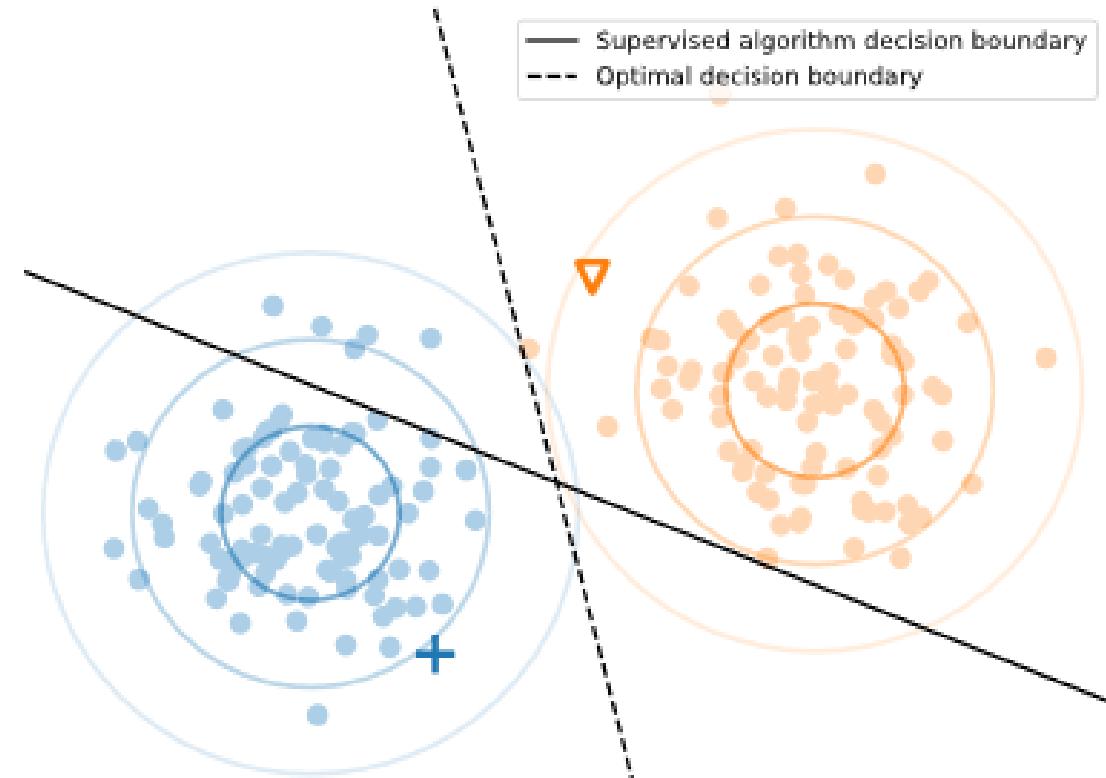




Aprendizado semi-supervisionado

Suposições:

- continuidade: registros próximos têm saídas próximas; continuidade de entradas e saídas;
- baixa densidade: a superfície de decisão não deve passar por regiões de alta densidade de pontos; tendência de formação de grupos homogêneos;
- adequação de baixa dimensionalidade;
- princípio de Vapnik: não se deve resolver um problema mais difícil como um passo intermediário do problema principal.



Open Access | Published: 15 November 2019

A survey on semi-supervised learning

Jesper E. van Engelen & Holger H. Hoos

Machine Learning 109, 373–440(2020) | Cite this article

35k Accesses | 63 Citations | 2 Altmetric | Metrics

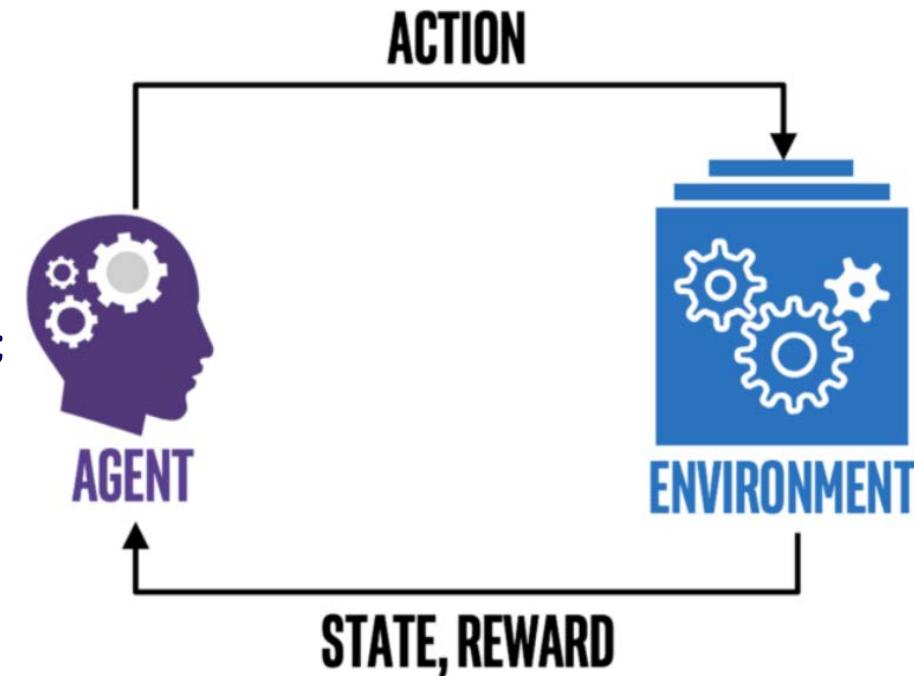


Aprendizado por reforço

O objetivo é o desenvolvimento de um modelo capaz de orientar as ações de agentes inteligentes em uma tomada de decisão sob incertezas, de modo a maximizar uma função de recompensas acumuladas de acordo com as tentativas realizadas.

A cada iteração, o agente:

- Observa o estado atual;
- Recebe a recompensa atual;
- Se ajusta e executa uma nova ação.

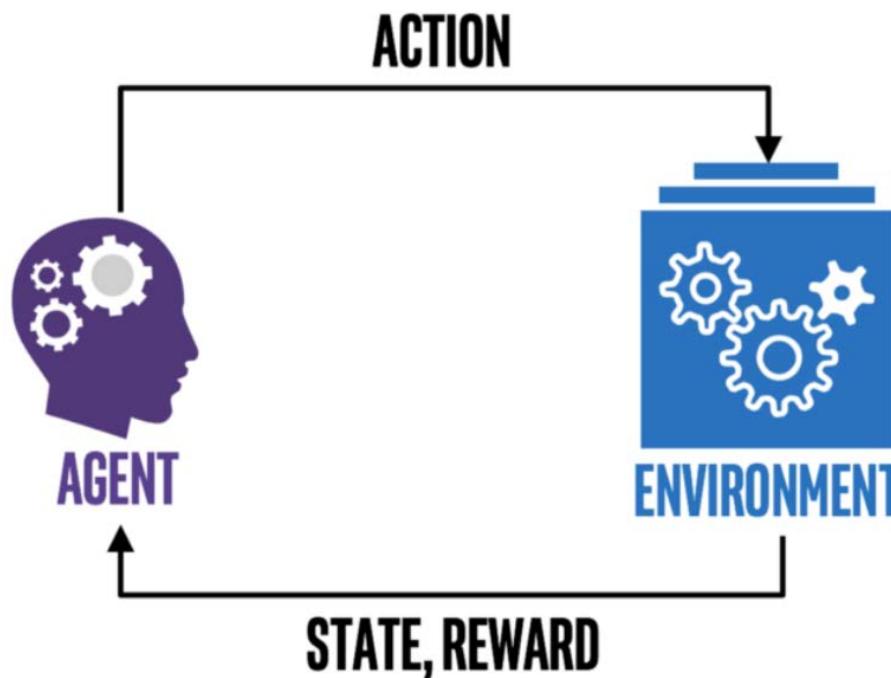


E o ambiente:

- Recebe a nova ação;
 - Produz um novo estado;
 - Calcula uma nova recompensa.
-
- Robótica
 - Jogos
 - Direção autônoma
 - Negociações em mercados



Aprendizado por reforço



A próxima ação depende do estado atual.

O esquema do aprendizado por reforço (observação >> ação >> recompensa >> novo estado) pode ser visto como um processo de Markov:

$$\langle E, A, R, T, \gamma \rangle$$

E : é o espaço de estados

A : é o espaço de ações

R : é a função de recompensas

$$R(e, a): E \times A \rightarrow \mathbb{R}$$

T : é a função de transição de estado

$$T(e, a, e'): E^2 \times A \rightarrow [0, 1]$$

$$T(e, a, e') = p(e'|e, a)$$

γ : é um fator de ajuste entre recompensas de curto e longo prazo

O objetivo é maximizar a recompensa obtida a partir de uma função de escolha $\pi(e): E \rightarrow A$



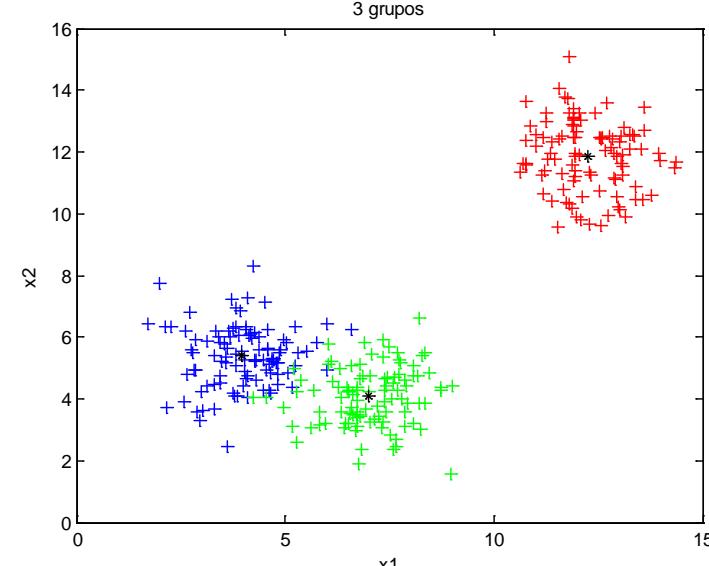
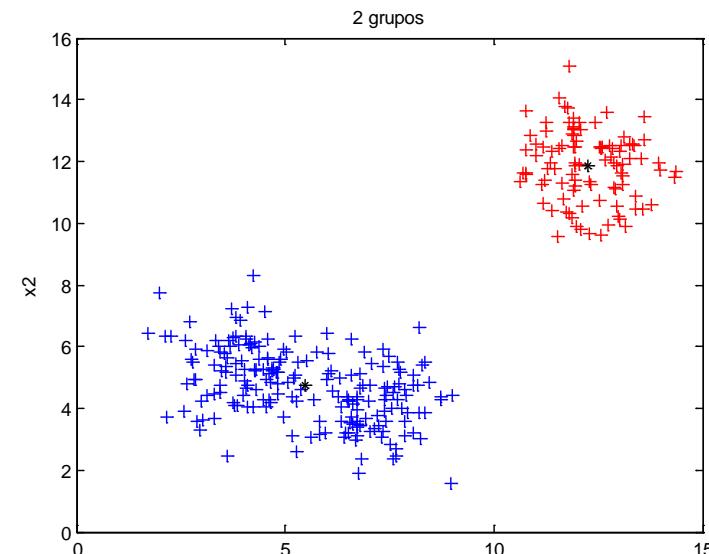
Estratégias de avaliação de desempenho: agrupamento

Um índice de validação de grupos é um valor que permite avaliar um agrupamento encontrado por um algoritmo quanto:

- ✓ à quantidade de grupos; e
- ✓ à qualidade da partição.

A maioria dos índices é baseada na estrutura geométrica da partição, informando o quanto os grupos encontrados estão afastados entre si (heterogeneidade externa, separação) e a proximidade dos elementos em cada grupo (homogeneidade interna, coesão).

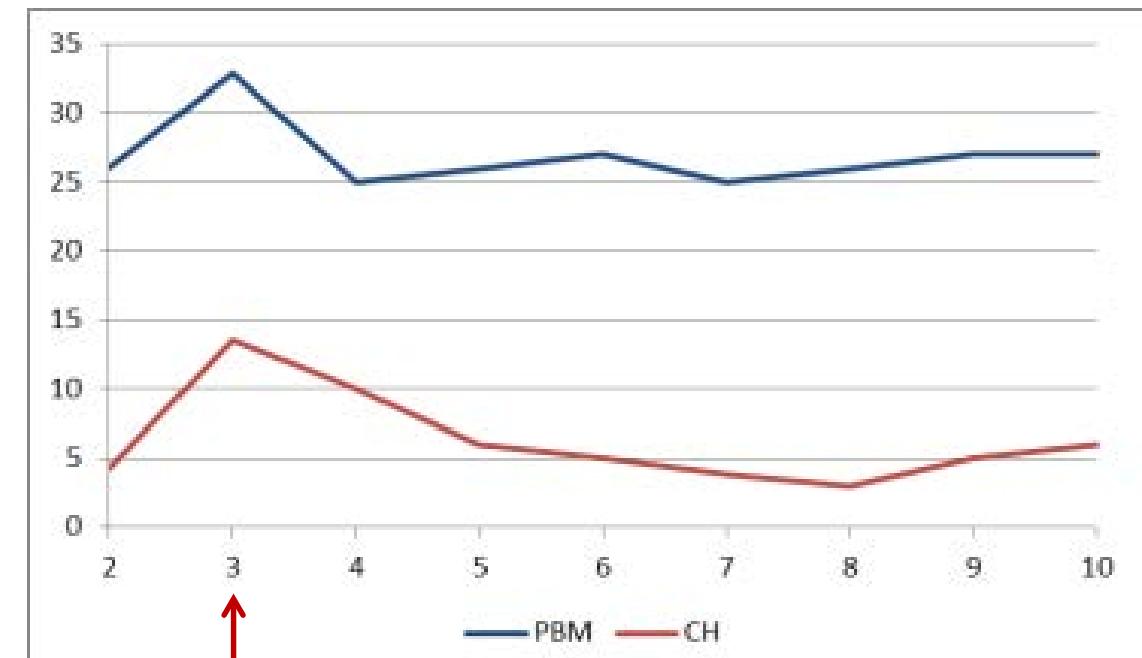
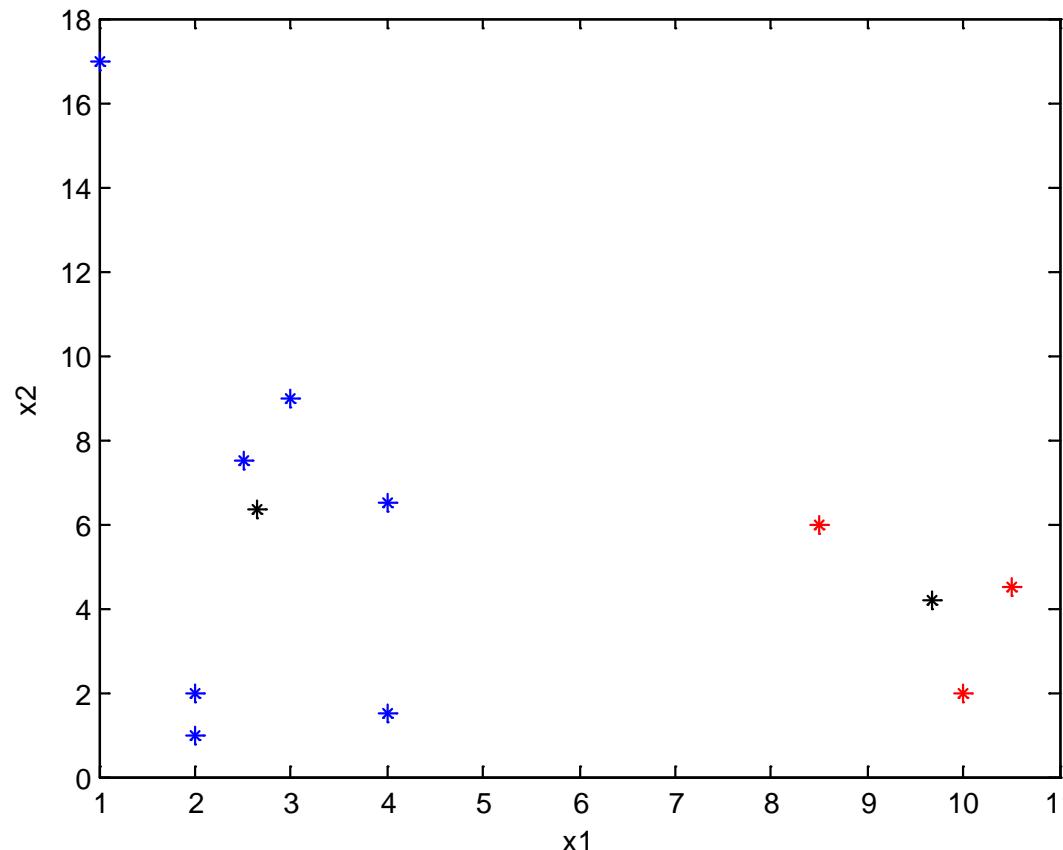
É possível também avaliar a qualidade do particionamento a partir da compatibilidade dos grupos produzidos com as classes conhecidas dos registros, mas essa é uma situação específica. O comum é não haver prévia informação sobre como os registros estão organizados.





Estratégias de avaliação de desempenho: agrupamento

Para se determinar o melhor particionamento, os dados são segmentados diversas vezes, cada vez com uma quantidade diferente de grupos. O melhor agrupamento é aquele que otimizar algum índice de validação de grupos considerado.



Neste exemplo, 3 grupos.

Estratégias de avaliação de desempenho: agrupamento

O índice PBM, proposto por Pakhira, Bandyopadhyay e Maulik, é definido como o quadrado do produto de três fatores:

em que:

$$PBM(K) = \left(\frac{1}{K} \cdot \frac{E_1}{E_K} \cdot D_K \right)^2$$



- K é a quantidade de grupos;
- $E_1 = \sum_{t=1}^N d(\mathbf{x}(t), \mathbf{w}_0)$ é a dispersão global em relação ao centro de todos os dados \mathbf{w}_0 ;
- $E_K = \sum_{i=1}^K \sum_{t=1}^N \mu_i(t) d(\mathbf{x}(t), \mathbf{w}_i)$ é a dispersão dos registros de cada grupo aos seus centros correspondentes;
- $\mu_i(t)$ é a pertinência do registro $\mathbf{x}(t)$ no grupo i ;
- $D_K = \max_{i,j=1..K} (d(\mathbf{w}_i, \mathbf{w}_j))$ é a máxima separação entre dois grupos da partição.



Estratégias de avaliação de desempenho: agrupamento

O índice Calinski-Harabasz é representado pela função:

em que:

$$CH(K) = \frac{B \cdot (N - K)}{W \cdot (K - 1)}$$



- K é a quantidade de grupos e N é a quantidade de registros;
- B representa a soma de distâncias entre os centros dos grupos e o centro global dos dados (média dos registros) ponderadas pelas quantidades N_i de elementos nos grupos:

$$B = \sum_{i=1}^K N_i d(\mathbf{w}_i, \mathbf{w}_0)^2$$

- W representa a soma de distâncias intragrupos ponderadas pelas pertinências $\mu_i(t)$ dos registros aos grupos:

$$W = \sum_{i=1}^K \sum_{t=1}^N \mu_i(t) d(\mathbf{x}(t), \mathbf{w}_i)^2$$



Estratégias de avaliação de desempenho: agrupamento

O índice Silhouette ou silhueta é representado pela função:

$$\bar{s}(k) = \frac{1}{N} \sum_{t=1}^N s(t)$$

↑ [-1,1]

em que:

- K é a quantidade de grupos, N é a quantidade de registros e $s(t)$ representa o valor da função silhueta para o registro t calculada como:

$$s(t) = \frac{b(t) - a(t)}{\max(a(t), b(t))}$$

$$a(t) = \frac{1}{|C_k| - 1} \sum_{\substack{j \in C_k \\ t \neq j}} d(t, j) \quad , \quad b(t) = \min \frac{1}{|C_k|} \sum_{\substack{j \in C_k \\ t \notin C_k}} d(t, j)$$

$a(t)$ é a distância média do registro t para todos os elementos j do seu grupo C_k

$b(t)$ é a menor distância média do registro t para todos os elementos j dos outros grupos C_k dos quais ele não faz parte



Estratégias de avaliação de desempenho: associação

Cobertura: Informa a frequência , absoluta ou relativa, do antecedente nos dados

SE sexo=feminino E estado.civil=casada, ENTÃO crédito=sim cobertura=18

Suporte: Informa a frequência, absoluta ou relativa, da ocorrência simultânea dos objetos

SE sexo=feminino E estado.civil=casada, ENTÃO crédito=sim suporte=12

Confiança: Informa a proporção de registros relacionados ao antecedente em que o objeto do consequente foi observado.

$$Confiança = \frac{Suporte}{SuporteAntecedente} = \frac{Suporte}{Cobertura}$$

SE $\boxed{\text{objeto}_1 \text{ E } \dots \text{ E } \text{objeto}_N}$, ENTÃO $\boxed{\text{objeto}_j}$
em que $j \notin \{1, \dots, N\}$



Estratégias de avaliação de desempenho: associação

Lift: $[0, +\infty[$: mede o aumento da ocorrência do consequente dado o antecedente.

$$Lift = \frac{Confiança}{Suporte_{Consequente}}$$

>1

Novidade: $[-0,25; 0,25]$: mede a diferença entre os suportes real e esperado.

$$Novidade = Suporte - Suporte_{Esperado}$$

>0

Convicção: $[0; +\infty[$: mede o grau de implicação da regra.

$$Convicção = \frac{1 - Suporte_{Consequente}}{1 - Confiança}$$

>1

antecedente
SE $objeto_1 \text{ E } \dots \text{ E } objeto_N$, ENTÃO $objeto_j$
em que $j \notin \{1, \dots, N\}$ consequente

RPE
2025

O aprendizado: tipos e tarefas

Estratégias de avaliação de desempenho: classificação

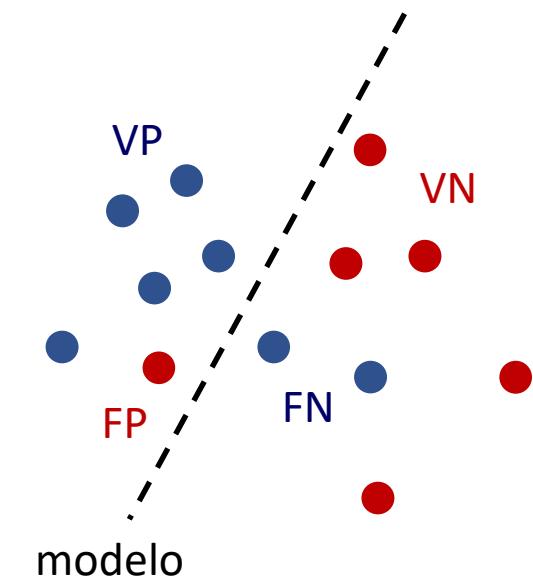
Matriz de confusão		Classe real		VP: verdadeiros positivos
		Positiva	Negativa	VN: verdadeiros negativos
Classe predita	Positiva	VP	FP	FP: falsos positivos
	Negativa	FN	VN	FN: falsos negativos

Qual a proporção de acertos globais? $Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$

Dentre os classificados como Positivos, qual a proporção de acertos? $Precisão = \frac{VP}{VP + FP}$

Dentre os Positivos, qual a proporção de acertos? $Sensibilidade = \frac{VP}{VP + FN}$
(recall)

Dentre os Negativos, qual a proporção de acertos? $Especificidade = \frac{VN}{VN + FP}$





Estratégias de avaliação de desempenho: classificação

Matriz de confusão		Classe real		VP: verdadeiros positivos
		Positiva	Negativa	
Classe predita	Positiva	VP	FP	VN: verdadeiros negativos
	Negativa	FN	VN	FP: falsos positivos

FN: falsos negativos

Qual o desempenho geral na classe positiva?

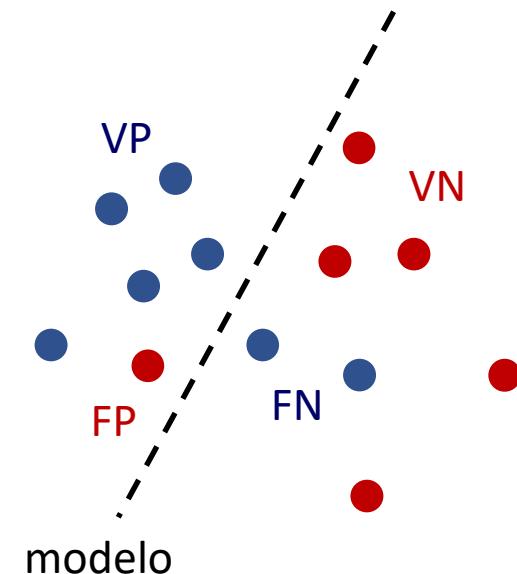
$$\text{Medida } F = 2 \cdot \frac{\text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \quad (\text{F_1 score})$$

$$\text{Média } G = \sqrt{\text{Precisão} \cdot \text{Sensibilidade}} \quad (\text{Fowlkes-Mallows Index})$$

média harmônica

média geométrica

$$M = \sqrt{(VP + FP) \cdot (VP + FN) \cdot (VN + FN) \cdot (VN + FP)}$$



Se a acurácia não é adequada em problemas desbalanceados, qual usar?

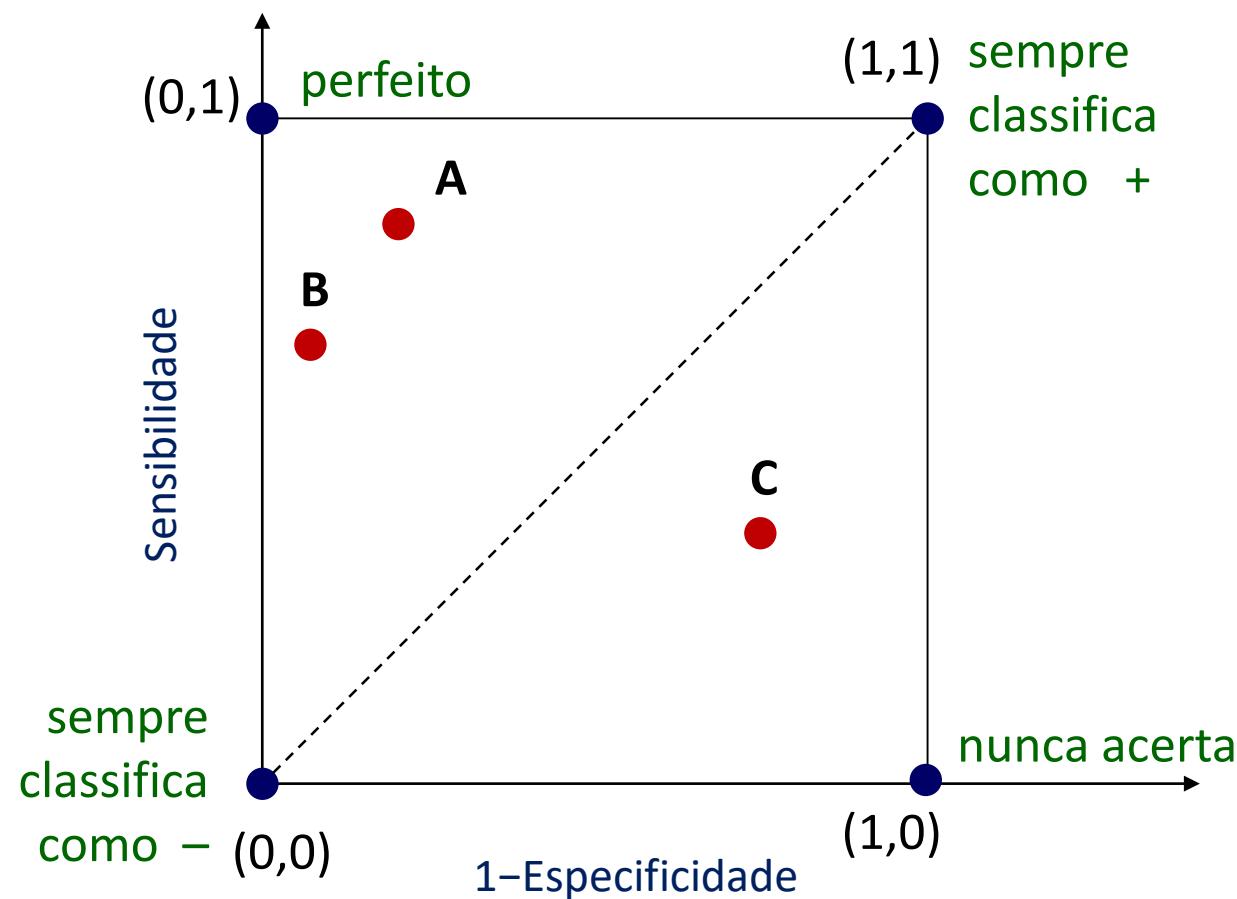
$$\text{Coef. corr. Matthews} =$$

$$\frac{VP \cdot VN - FP \cdot FN}{M}$$

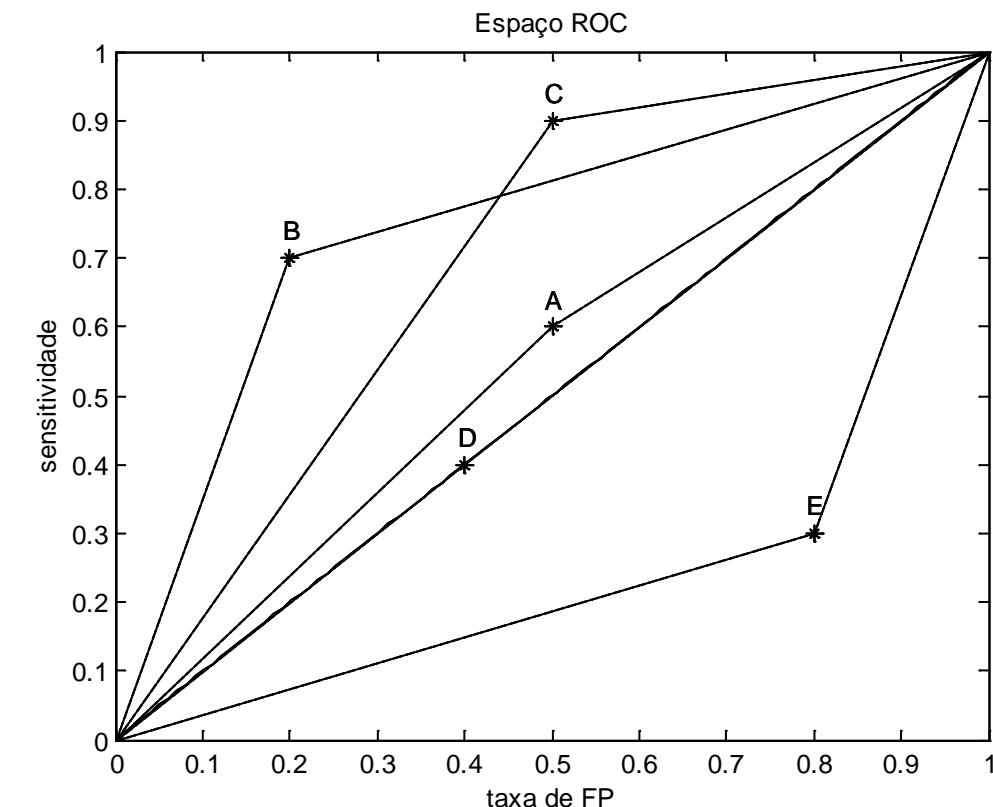


Estratégias de avaliação de desempenho: classificação

Espaço ROC (*Receiver Operating Characteristic*)



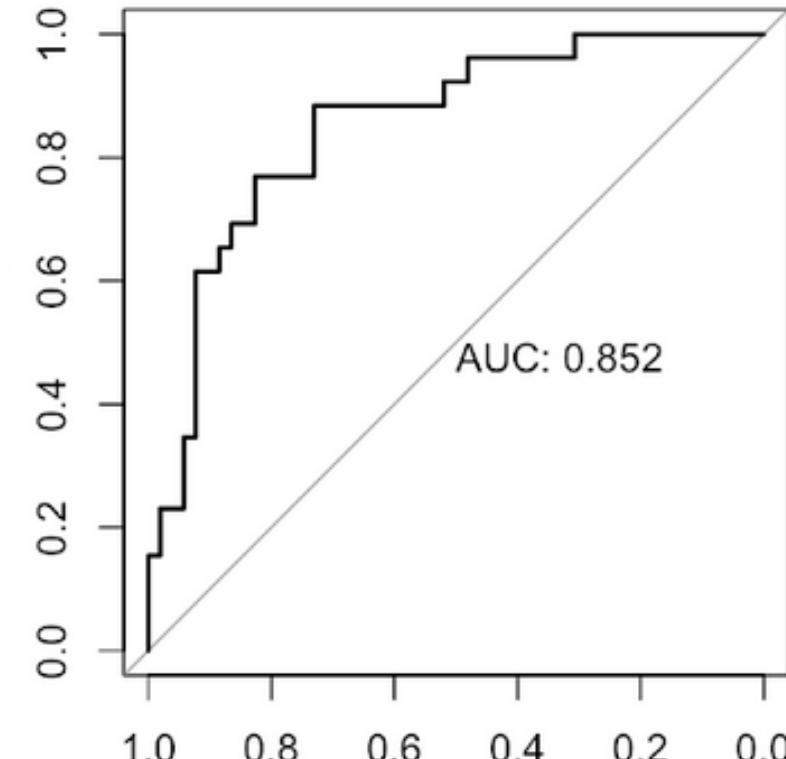
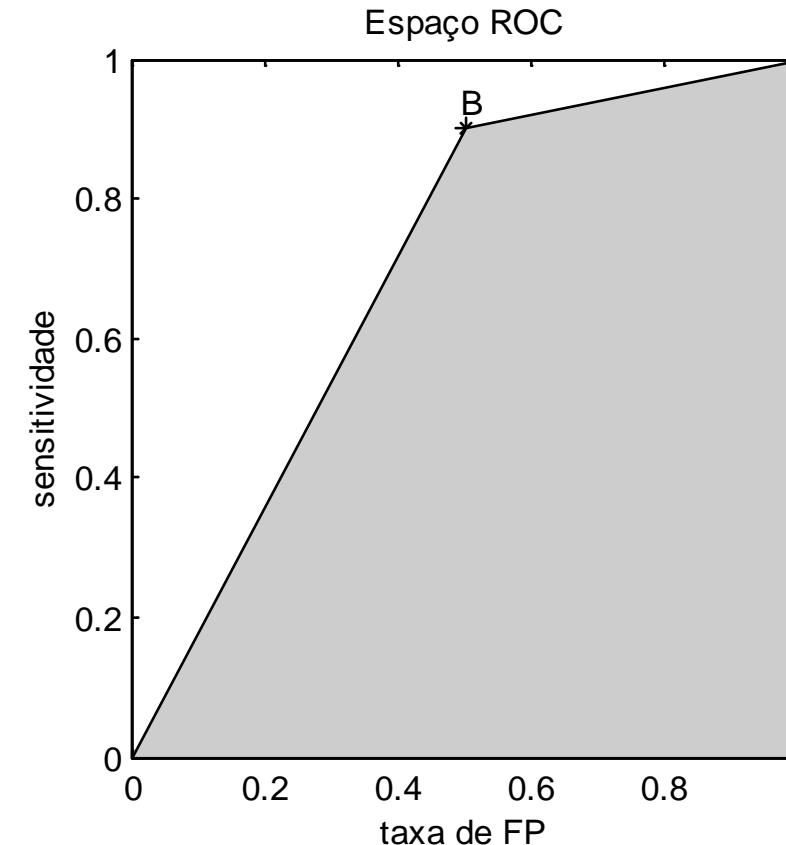
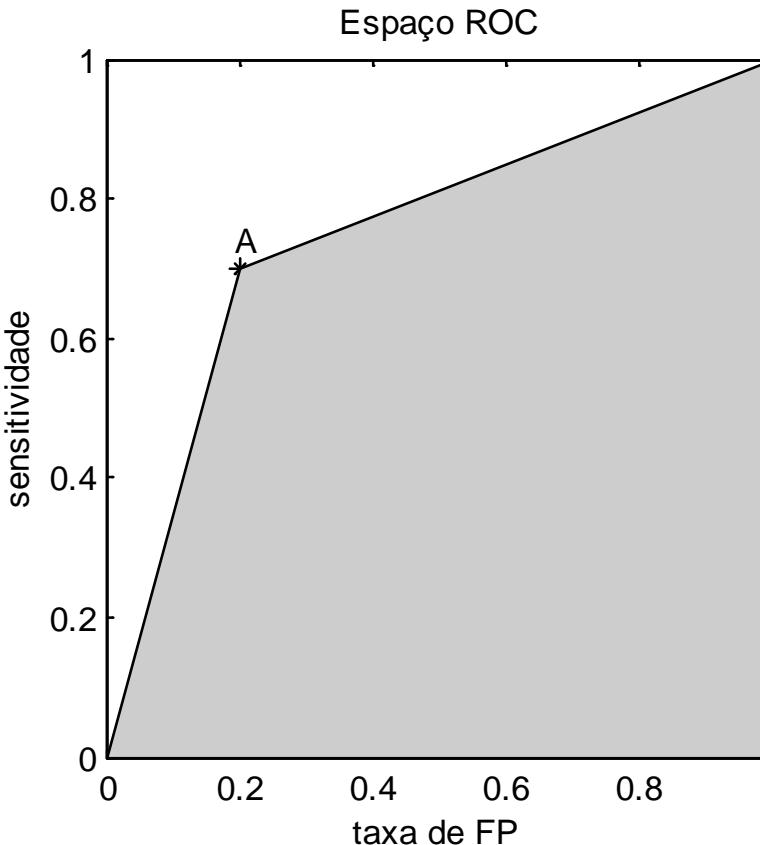
Curvas ROC





Estratégias de avaliação de desempenho: classificação

Área sob a curva ROC (AUC)



O melhor classificador é o que possuir a maior área sob a curva ROC.

RPE
2025

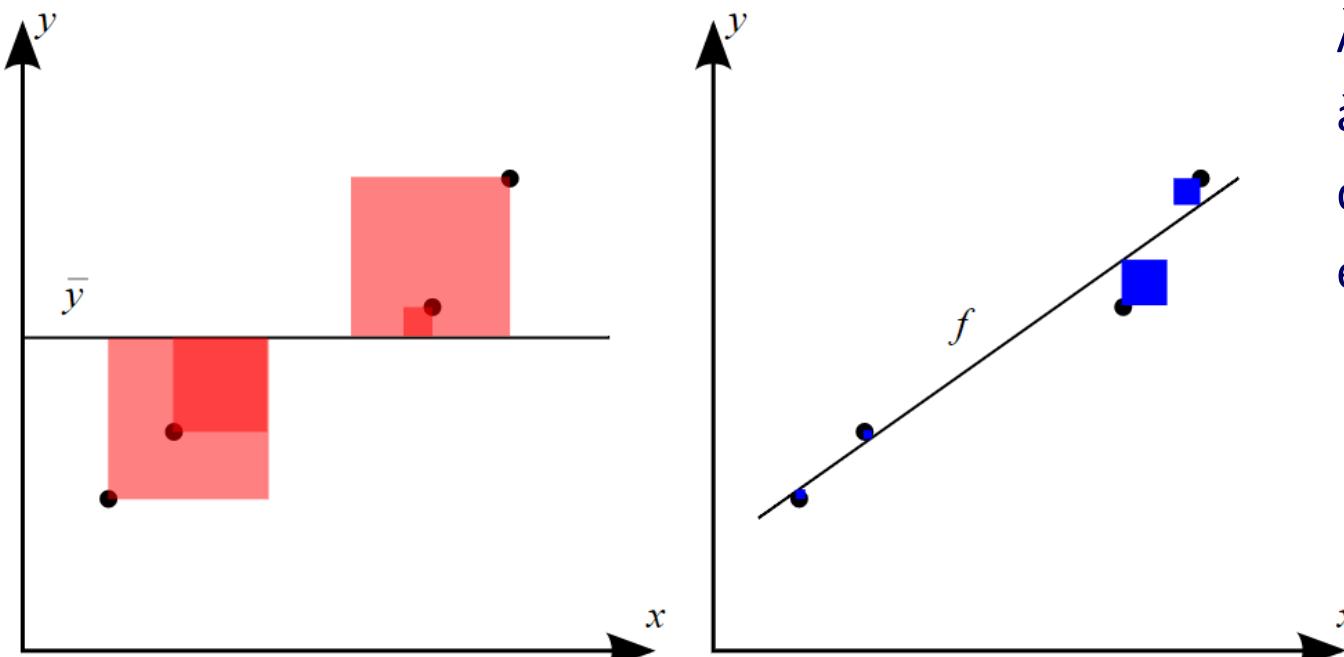
O aprendizado: tipos e tarefas

Estratégias de avaliação de desempenho: regressão

Coeficiente de determinação (R^2) :

$$R^2 = \frac{\sum_{t=1}^N (\hat{y}(t) - \bar{y})^2}{\sum_{t=1}^N (y(t) - \bar{y})^2} = 1 - \frac{\sum_{t=1}^N (\hat{y}(t) - y(t))^2}{\sum_{t=1}^N (y(t) - \bar{y})^2}$$

O R^2 é a proporção da variância da variável resposta que é “capturada” pela variabilidade das estimativas.



À medida que variáveis explicativas são adicionadas à modelagem, a tendência do R^2 é aumentar devido ao superajuste. Assim, para reduzir esse efeito, adota-se o R^2 ajustado:

$$R_a^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

N registros, p variáveis

$$R_a^2 \leq R^2$$



Estratégias de avaliação de desempenho: regressão

Raiz do erro médio quadrático (RMSE) :
(root-mean-square error)

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N e^2(t)} = \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{y}(t) - y(t))^2}$$

Erro médio absoluto (MAE) :
(mean absolute error)

$$MAE = \frac{1}{N} \sum_{t=1}^N |e(t)| = \frac{1}{N} \sum_{t=1}^N |\hat{y}(t) - y(t)|$$

Erro percentual médio absoluto (MAPE) :
(mean absolute percentage error)

$$MAPE = \frac{1}{N} \sum_{t=1}^N \frac{|e(t)|}{y(t)} = \frac{1}{N} \sum_{t=1}^N \frac{|\hat{y}(t) - y(t)|}{y(t)}$$

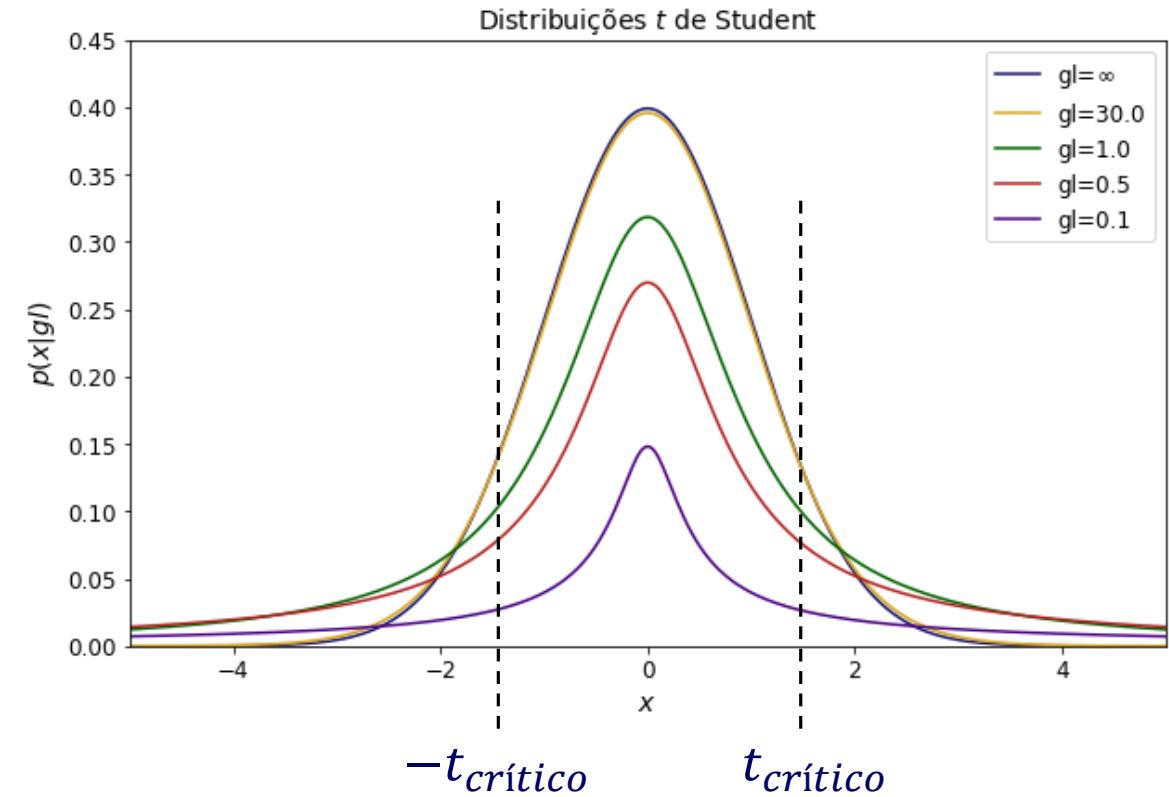


Estratégias de avaliação de desempenho: qualificando o resultado

É comum que os modelos obtidos sejam afetados aleatoriamente, seja na parametrização dos próprios, seja na amostragem de registros realizada.

Assim, a execução do mesmo experimento algumas dezenas de vezes se faz necessária para a obtenção das médias (amostrais) dos desempenhos. Nestes casos, a determinação de intervalos de confiança para as médias populacionais agrega valor à análise dos desempenhos.

O intervalo de confiança para a média populacional de uma distribuição com variância populacional desconhecida é determinado com uma distribuição t de Student (William Gosset), de onde são obtidos os valores críticos a partir dos graus de liberdade dos dados e do grau de confiança exigido.



$$P\left(\bar{x} - t_{\text{crítico}}(n-1) \cdot \frac{s}{\sqrt{n}} \leq x \leq \bar{x} + t_{\text{crítico}}(n-1) \cdot \frac{s}{\sqrt{n}}\right) =$$

grau de confiança = 1 - nível de significância

Estratégias de avaliação de desempenho: qualificando o resultado

Além da determinação do intervalo de confiança, outro importante aspecto está na comparação entre as médias de desempenhos dos diferentes modelos estudados.

É importante fazer um teste de hipóteses para amostras independentes com variâncias populacionais desconhecidas e diferentes:

- a) Definir a hipótese nula: H_0 (comparação entre as médias considerando a igualdade: $=, \geq$ ou \leq);
Ex.: Dados dois modelos, deseja-se verificar se é possível afirmar que seus desempenhos próximos na amostragem realizada significam que estes são iguais na população. Então,

$$H_0: \mu_1 = \mu_2 \Leftrightarrow H_0: \mu_1 - \mu_2 = 0.$$

- b) Calcular a estatística (valor observado) t de comparação do teste e seu grau de liberdade:

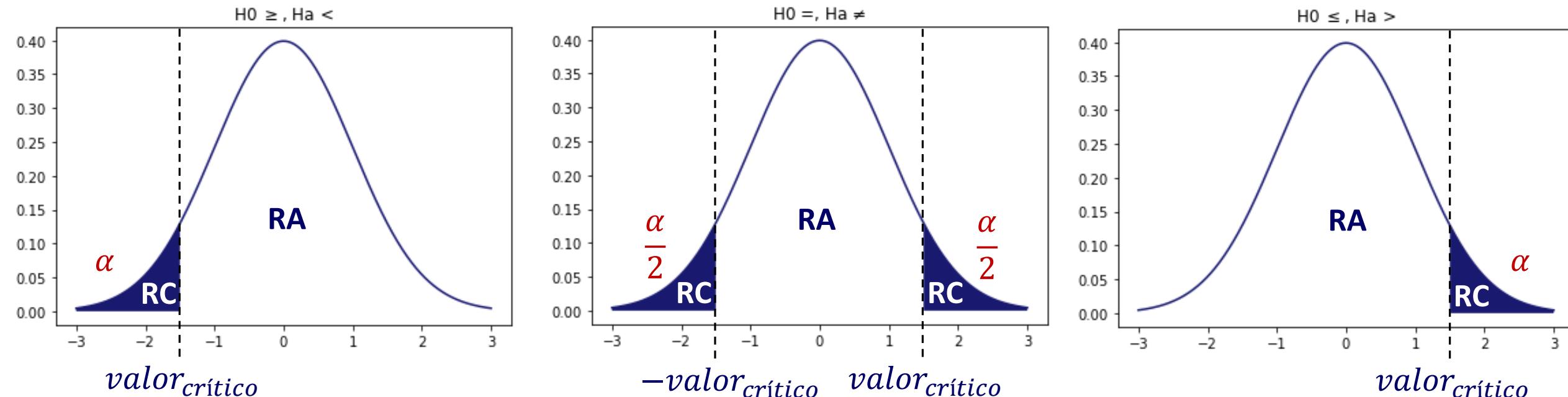
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

grau de liberdade = $\min(n_1 - 1, n_2 - 1)$



Estratégias de avaliação de desempenho: qualificando o resultado

- c) De acordo com a comparação do item (a), identificar o tipo de teste: unilateral (\leq ou \geq) ou bilateral ($=$) e sua região crítica (RC);



- d) Obter o valor crítico do teste de acordo com o seu tipo usando o seu nível de significância (α) na tabela t de Student; nível de significância é a probabilidade de se rejeitar H_0 sendo ela verdadeira (erro tipo I); o valor crítico define a fronteira entre a região crítica e a região de aceitação (RA);



O aprendizado: tipos e tarefas

Estratégias de avaliação de desempenho: qualificando o resultado

		Realidade	
		H_0 é V	H_0 é F
Decisão	Não rejeitar H_0	Acerto	Erro tipo II (β)
	Rejeitar H_0	Erro tipo I (α)	Acerto

- e) Se $t \in RC$, então rejeita-se H_0 .
Se $t \notin RC$, então não se rejeita H_0 .

Retomando o exemplo, se o teste falhar (rejeição de H_0), não se pode afirmar que seus desempenhos sejam iguais (o que sugere uma real diferença de desempenho).

Se o teste funcionar (não se rejeita H_0), não se pode afirmar que seus desempenhos sejam diferentes (o que sugere que os modelos são iguais em desempenho).

Teste de médias amostrais com parâmetros populacionais desconhecidos

- f) Alternativamente, pode-se rejeitar ou não H_0 a partir do seu **nível descritivo** ou **p-valor (p)**. O p-valor é um valor definido em $[0,1]$, encontrado a partir de uma distribuição normal padrão para a estatística t calculada, associado à quantidade de evidências de não rejeição de H_0 .

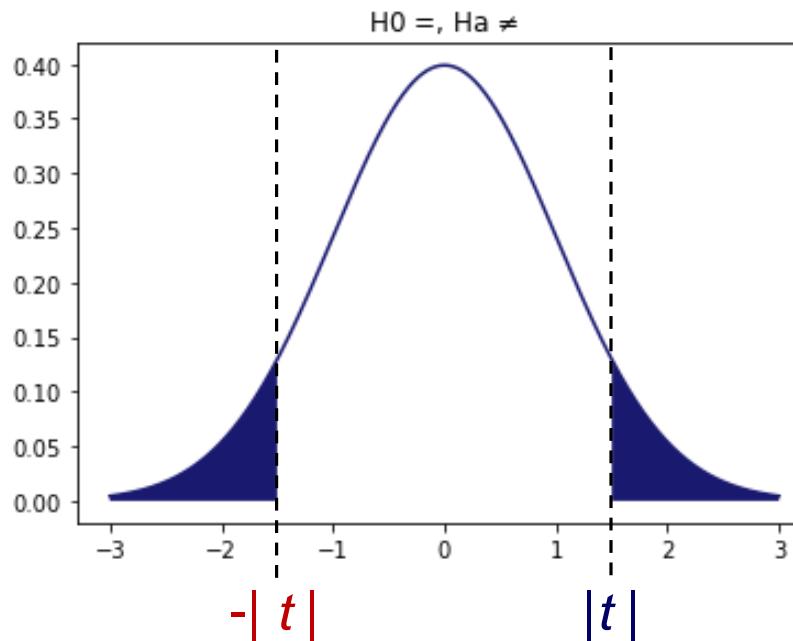


Tabela - Normal Padrão de 0 a z

Segunda casa decimal de z

z	0	1	2	3	4	5	6	7	8	9
	0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767

Segunda casa decimal de z

P($0 \leq Z \leq z$)

Teste de médias amostrais com parâmetros populacionais desconhecidos

Quanto menor for o p-valor, menos provável é a estatística amostral usada para inferir sobre a população (associada à igualdade de características ou à ausência de diferenças observáveis).

Em outras palavras, mais raro é o evento que se quer testar. Assim, se o p-valor for menor ou igual ao nível de significância do teste, H_0 será rejeitada. Caso contrário, ela não é rejeitada.

Retomando o exemplo, considerando que H_0 afirma que os desempenhos são iguais, quanto menor o p-valor, mais evidências se tem de que os modelos têm desempenhos diferentes ou mais rara será a hipótese de que os desempenhos são iguais.

Se o teste for realizado com nível de significância de 1% e o p-valor obtido for de 0,004, por exemplo, então rejeita-se H_0 , isto é, não se pode dizer que os desempenhos são iguais, o que **sugere** que os desempenhos sejam diferentes. Mas se o p-valor fosse de 0,1001, então H_0 não poderia ser rejeitada.

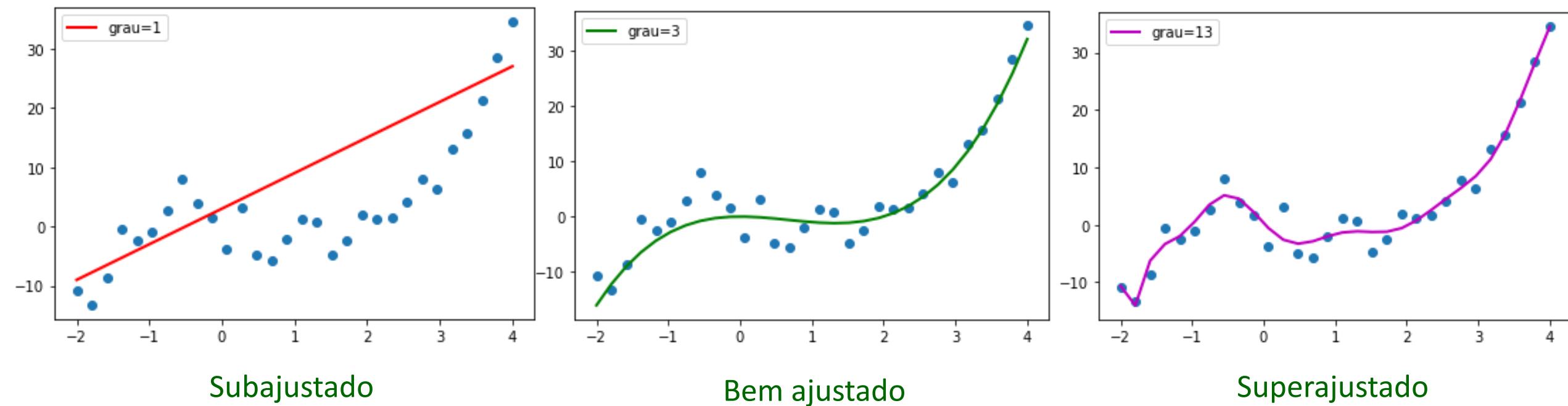
O valor do nível de significância informa que há 1% de probabilidade do teste falhar, ou seja, H_0 ser verdadeira mesmo com $p < \alpha$.



Os riscos de uma modelagem: superajuste e subajuste

O **superajuste** (*overfitting*) ocorre quando o modelo é mais complexo que o problema, estando **muito** ajustado aos dados de **treinamento** e não representando adequadamente outras amostras de dados.

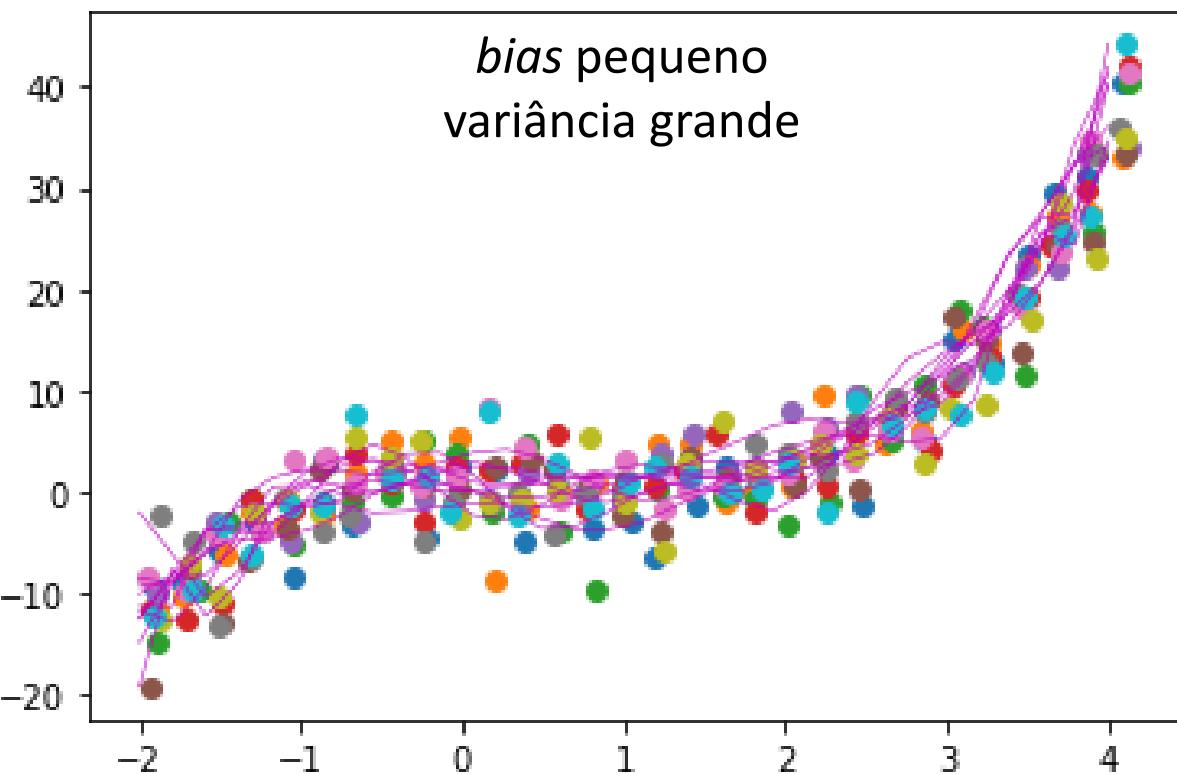
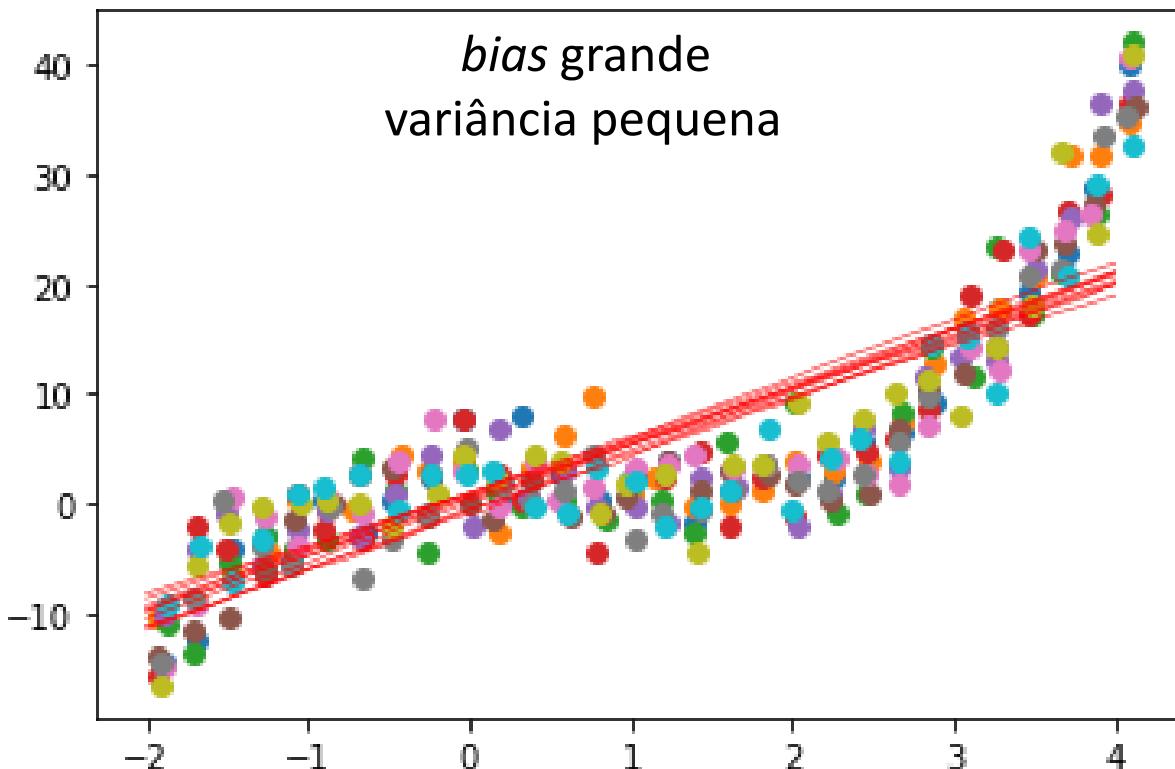
O **subajuste** (*underfitting*) ocorre quando o modelo é mais simples que o problema, estando **pouco** ajustado aos dados de **treinamento** e não representando adequadamente outras amostras de dados.





O dilema *bias* x variância

O viés ou ***bias*** de um modelo é a diferença entre o valor esperado do estimador e o valor real do parâmetro que está sendo estimado. A **variância** é o valor esperado dos desvios quadráticos em relação à media.



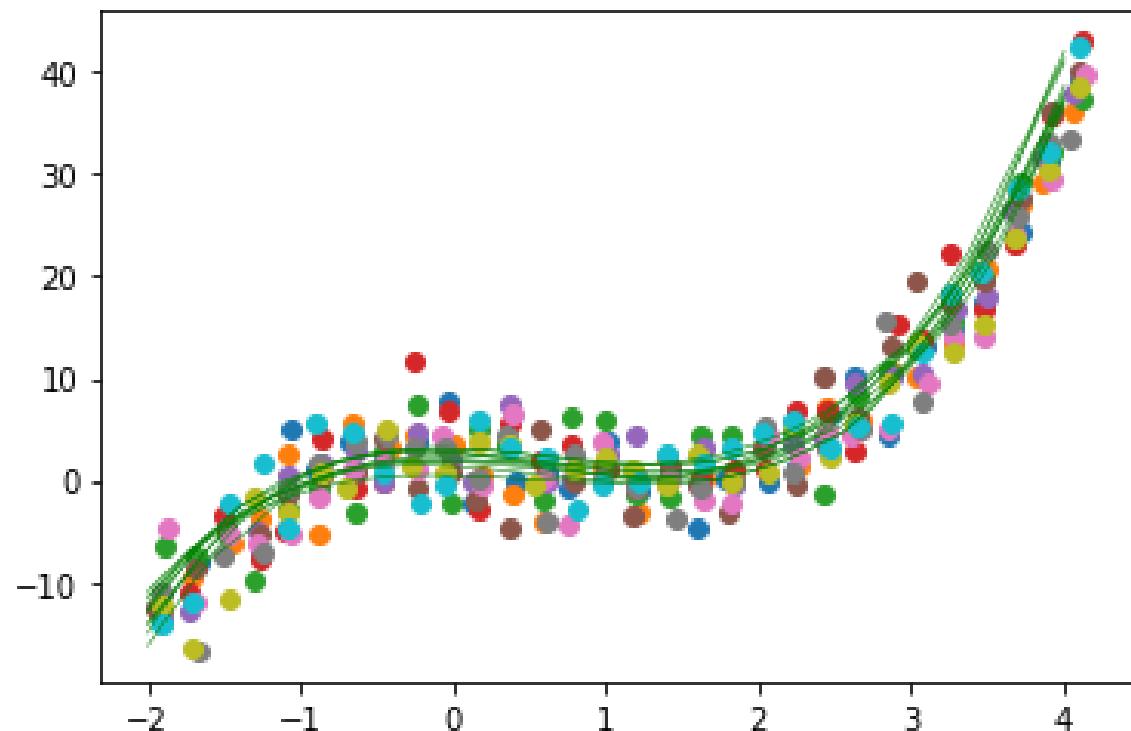
Modelos com ***bias*** pequeno tendem a possuir uma variância maior das estimativas de parâmetros entre as amostras. De modo análogo, quanto maior o ***bias***, menor a variância.



O dilema *bias x variância*

O **viés** ou ***bias*** de um modelo é a diferença entre o valor esperado do estimador e o valor real do parâmetro que está sendo estimado. A **variância** é o valor esperado dos desvios quadráticos em relação à media.

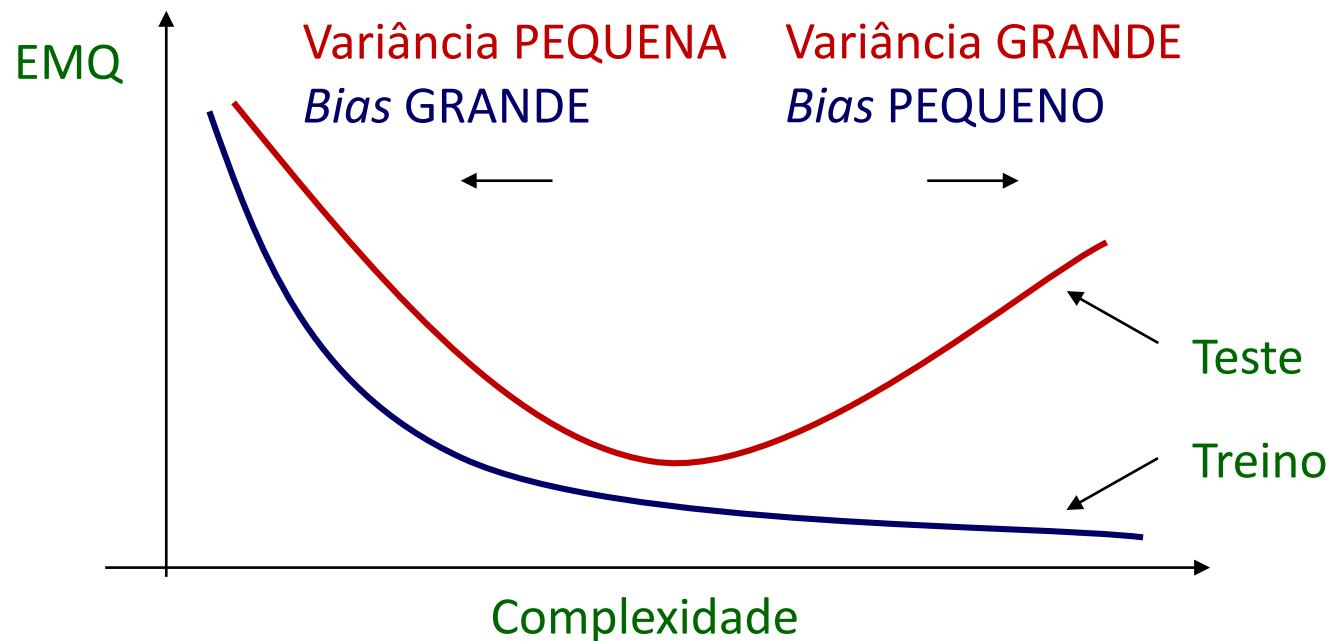
- *Bias* grande pode fazer com que um algoritmo de aprendizado perca as relações relevantes entre os atributos e as variáveis de decisão (subajuste).
- Variância grande pode fazer com que um algoritmo de aprendizado modele o ruído aleatório dos dados de treinamento (superajuste).



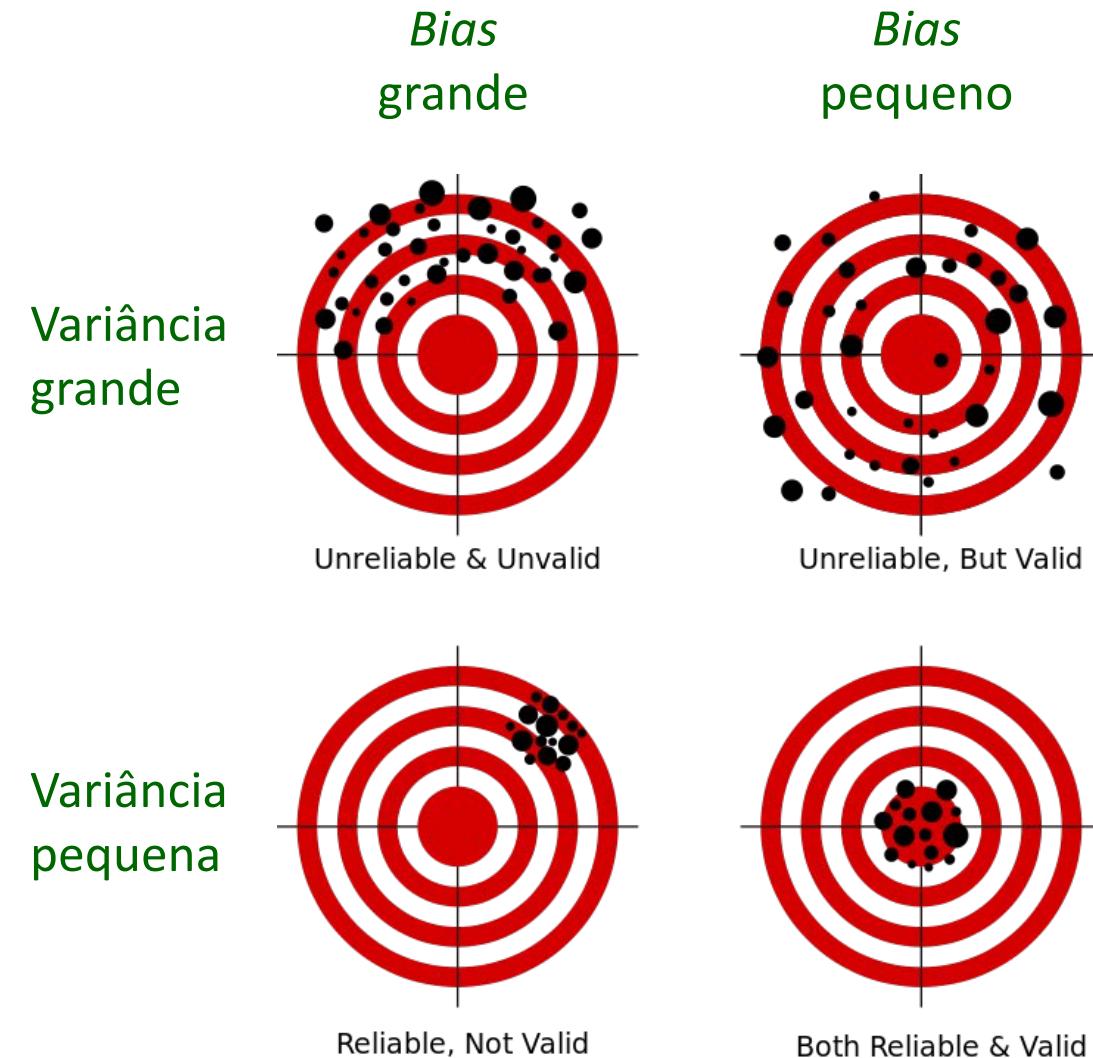
A estrutura correta reduz simultaneamente o bias e a variância.



O dilema *bias x variância*



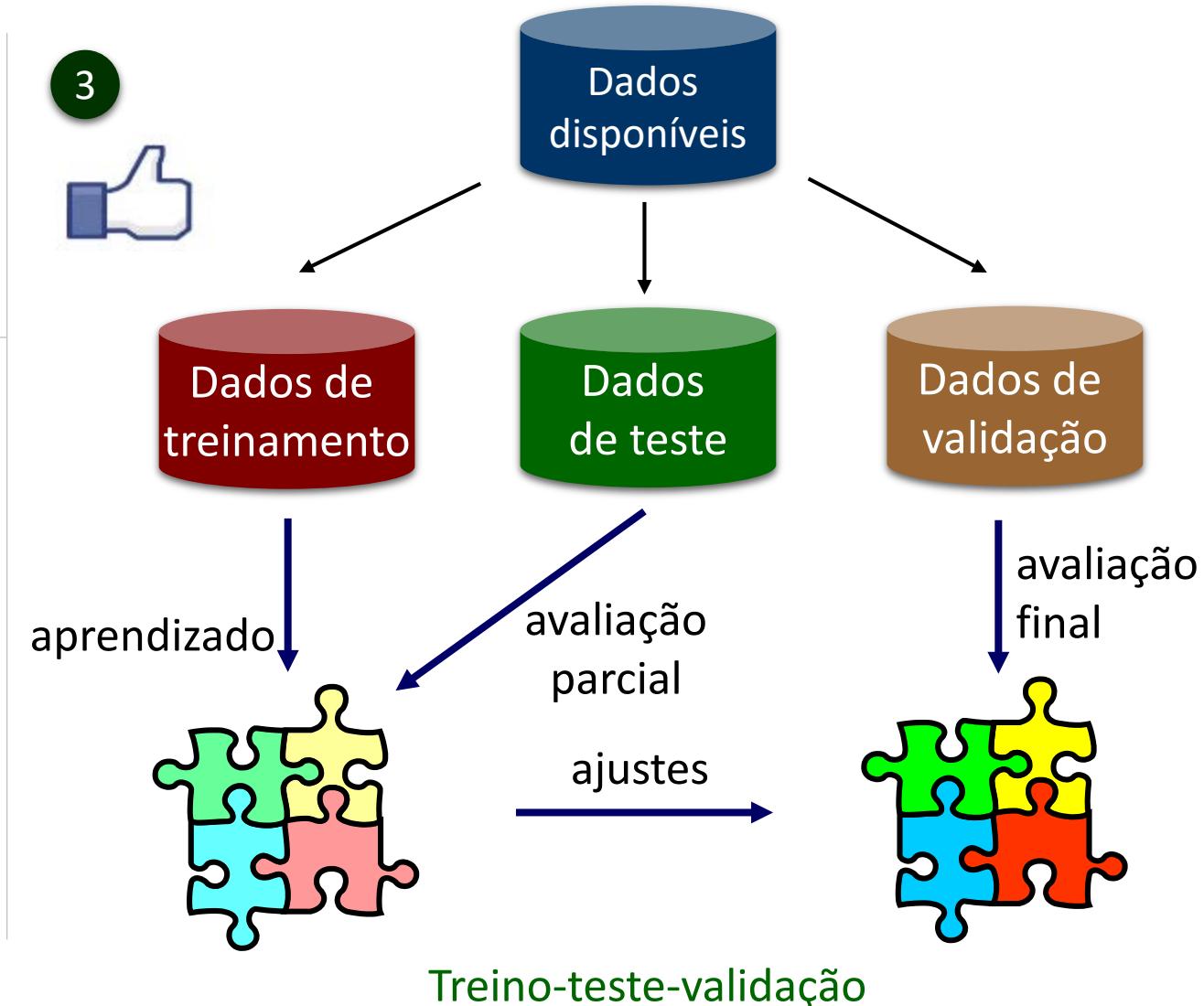
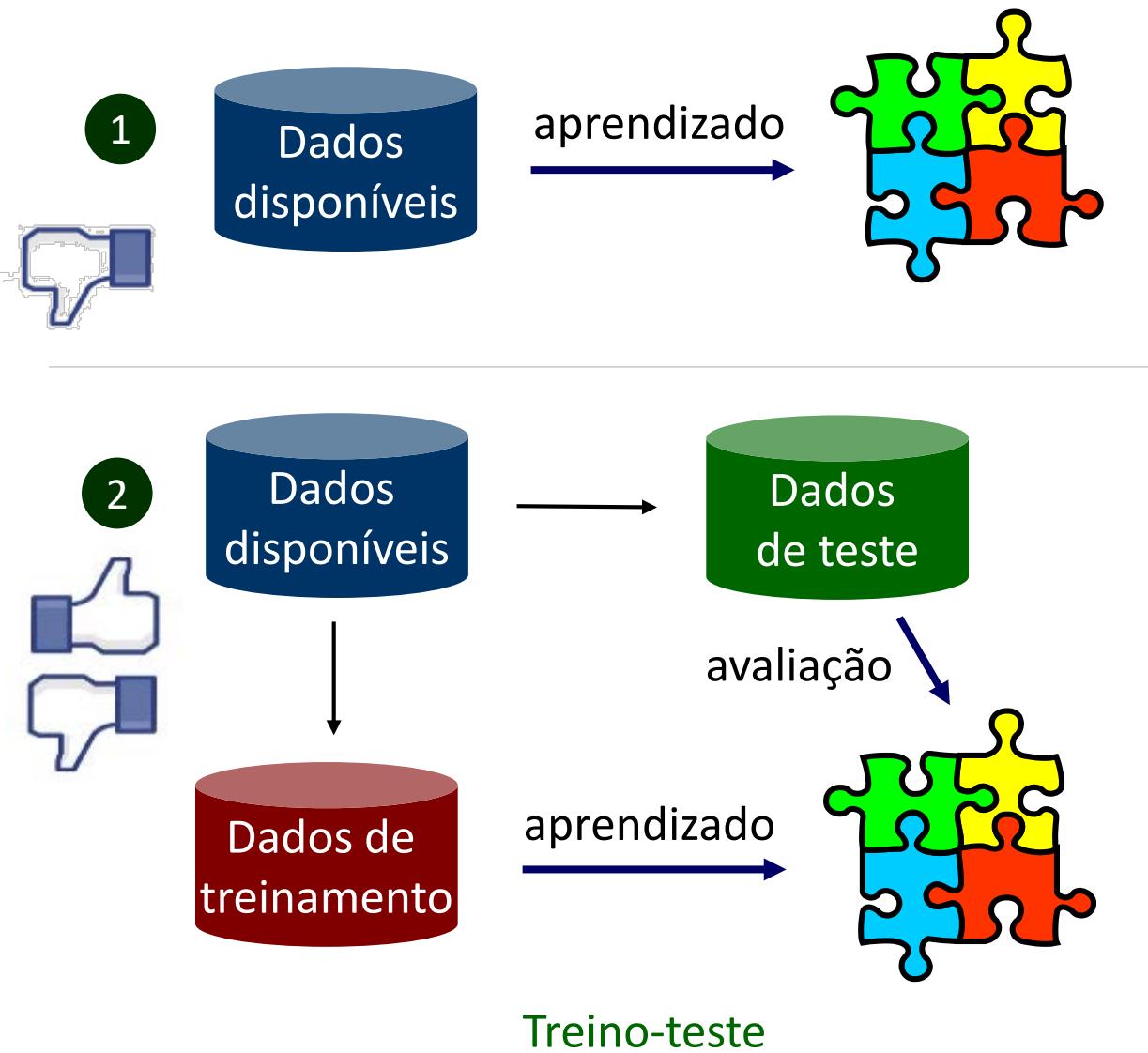
A estrutura correta tem melhor capacidade de predição nos registros de teste.





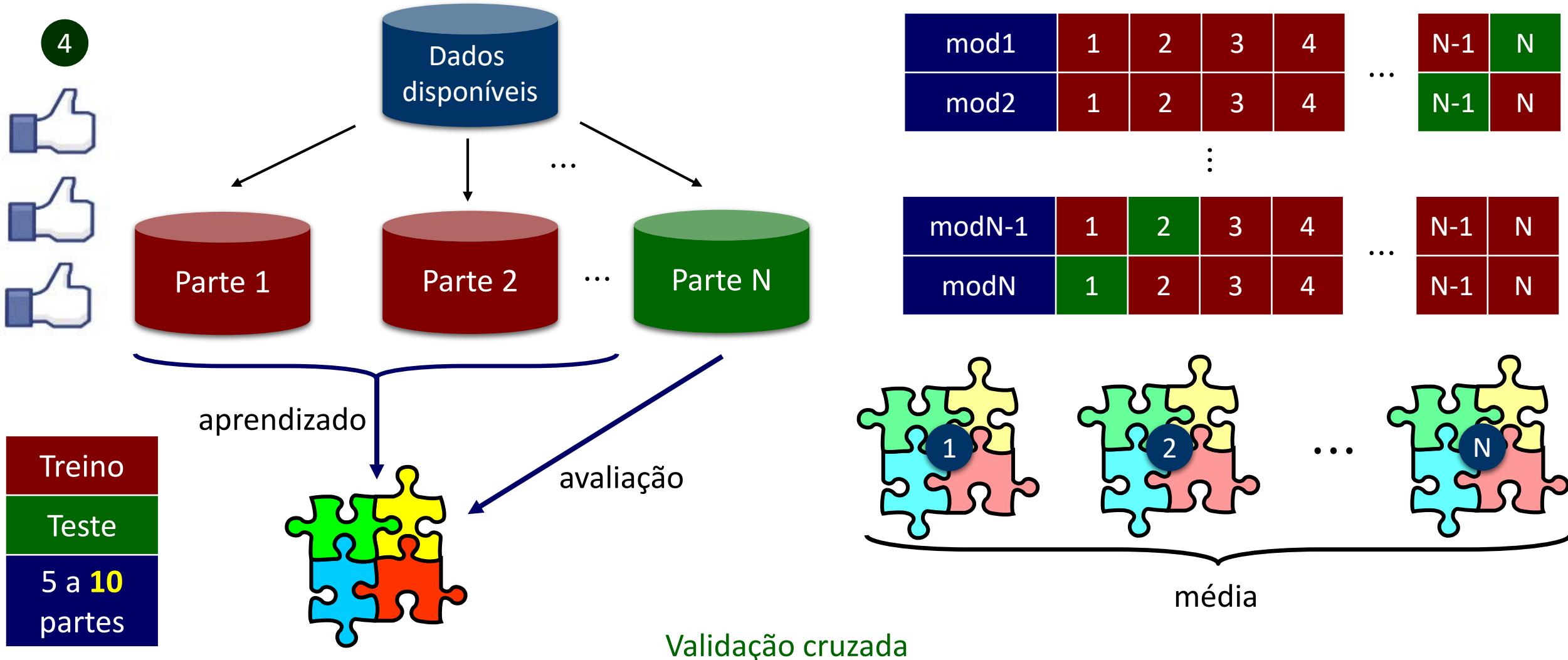
Estratégias de avaliação de desempenho

Avaliando a capacidade de generalização: técnicas de amostragem

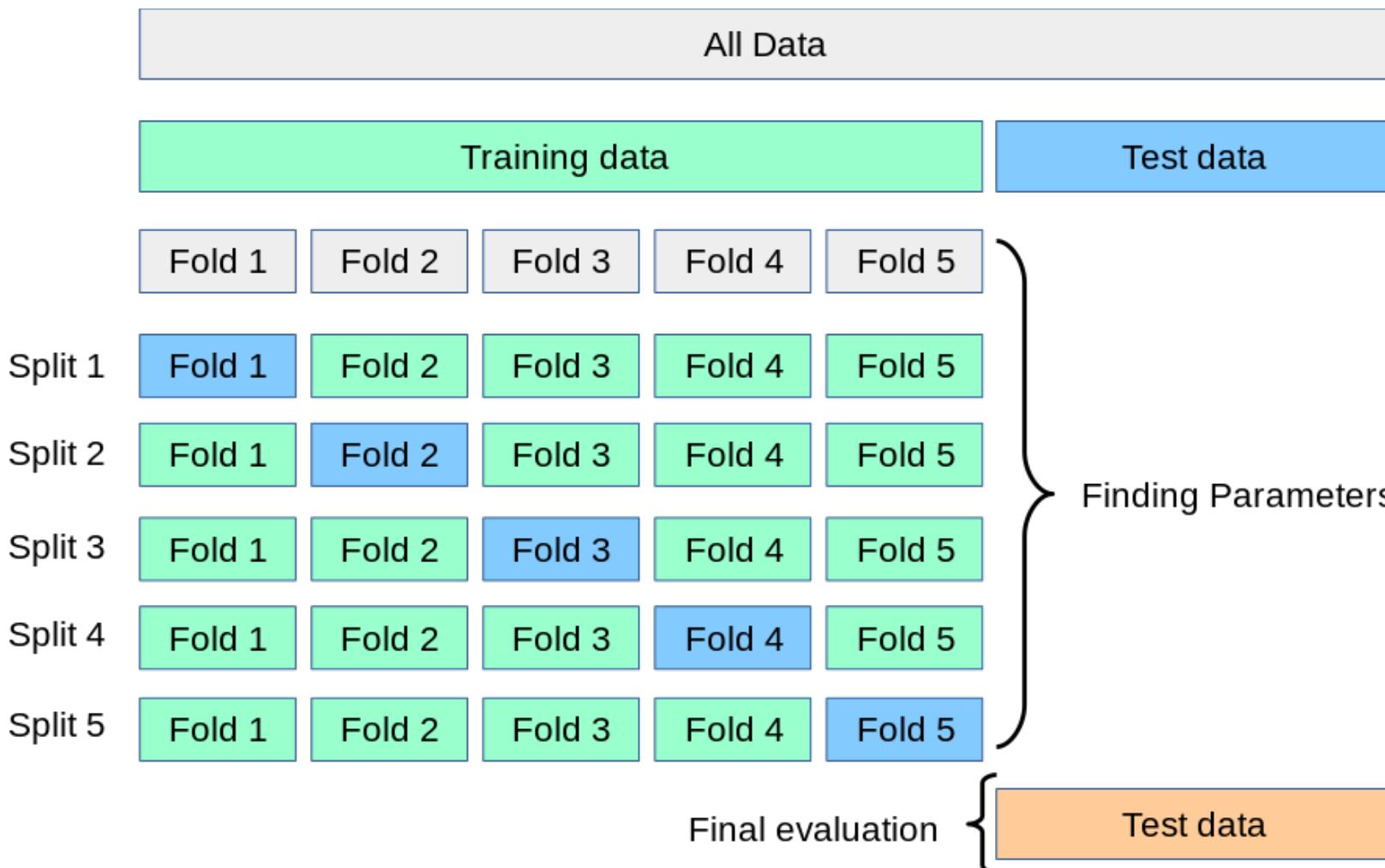




Avaliando a capacidade de generalização: técnicas de amostragem



Avaliando a capacidade de generalização: técnicas de amostragem



O esquema ao lado é admissível para **seleção** de modelos ou definição de parâmetros, mas **não é** o ideal para **avaliação** de modelos.

Qual modelo entregar para a implantação?



Avaliando a capacidade de generalização: técnicas de amostragem

Exemplo: Suponha uma base com 1000 registros.

Estratégia de amostragem	% de acertos			
	Treino	Teste	Validação	Expectativa
Treino (100)	95	-	-	95
Treino-teste (70-30)	90	88	-	88
Treino-teste-validação (70-15-15)	89	80	85	85
Validação cruzada com 4 partes (75-25)	88	83	-	83
Etapa 1	81	77	-	-
Etapa 2	90	87	-	-
Etapa 3	89	82	-	-
Etapa 4	92	86	-	-

Avaliando a capacidade de generalização: técnicas de amostragem

Outras técnicas

Leave-one-out: validação cruzada para bases pequenas de modo que a quantidade de partes é a mesma da de registros. O aprendizado, então, considera 1 registro para o teste e N-1 registros para o treino.

Ex.: 30 registros => 30 modelos diferentes gerados, cada um usando um registro diferente como teste e os demais registros para a partir de 29 registros de treinamento e testados no registro não utilizado. Cada modelo usa um registro diferente como teste.

Hold out com repetição: aplicar treino-teste (*holdout*) por N vezes, normalmente divididos aleatoriamente na proporção de 2/3 e 1/3 ou 70% e 30%.

Ex.: 3000 registros, 4 vezes => 4 modelos gerados a partir de 2100 registros de treinamento e testados nos demais 900 registros.

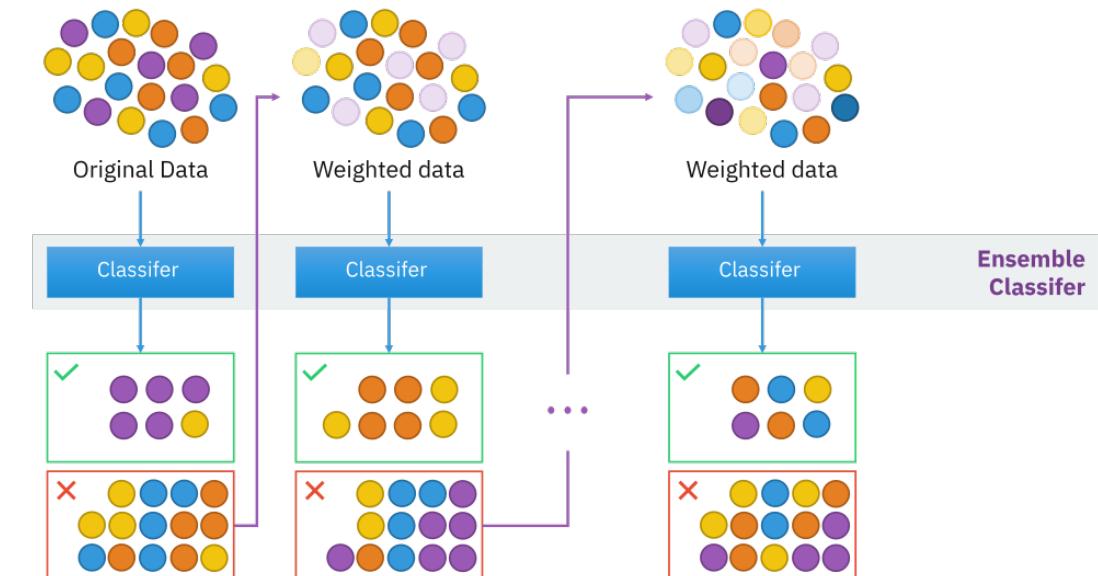
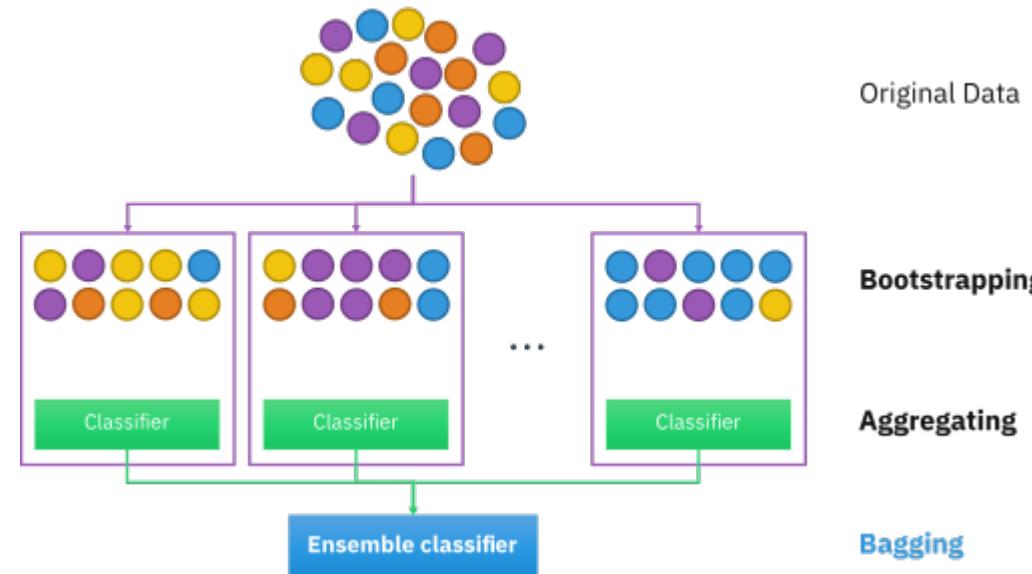
Bootstrap: realizar k amostragens com reposição para treino e gerar modelos a partir deles; os registros não selecionados são usados para teste.

Avaliando a capacidade de generalização: técnicas de amostragem

Outras técnicas

Bagging (bootstrap aggregation): realizar o *bootstrap* várias vezes, manter os modelos produzidos, e decidir a classe de um novo registro por votação desses modelos.

Boosting: iterativamente modelos são construídos para lidar com os erros de classificação da iteração anterior; dessa forma, um novo registro é classificado pela ponderação dos modelos construídos.





Outros itens de avaliação

- Rapidez: de construção e de aplicação de um modelo
- Escalabilidade: capacidade para trabalhar com grandes volumes de dados
- Robustez: capacidade para trabalhar com bases de dados difíceis e diversificadas
- Interpretabilidade: facilidade de entendimento das decisões e/ou do modelo

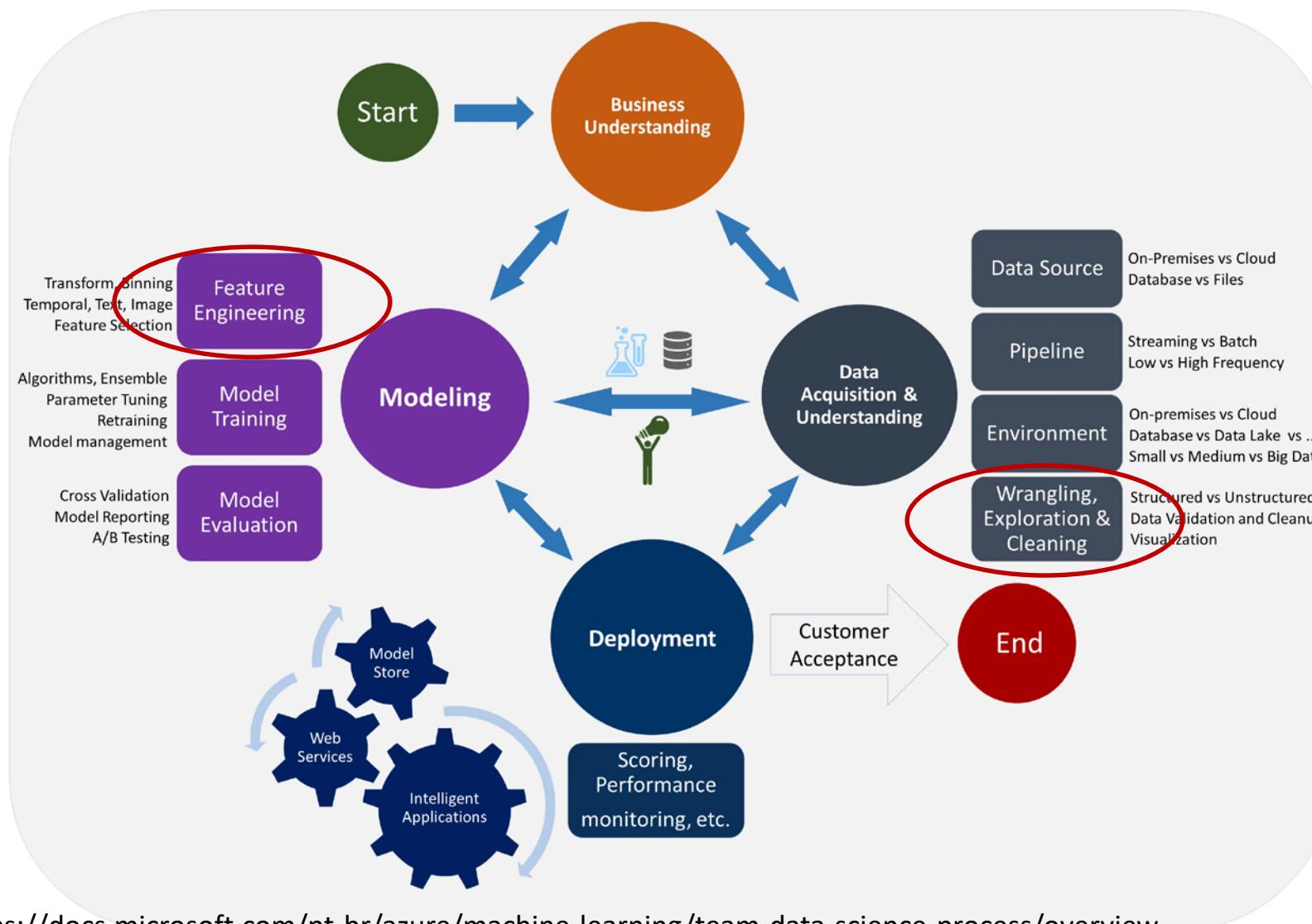


Já vistos:

- Eficácia: estimativa da correção do modelo (tarefas preditivas)
- Qualidade descritiva: estimativa de quão bem um conjunto de dados é representado por um modelo (tarefas descritivas)



Motivação



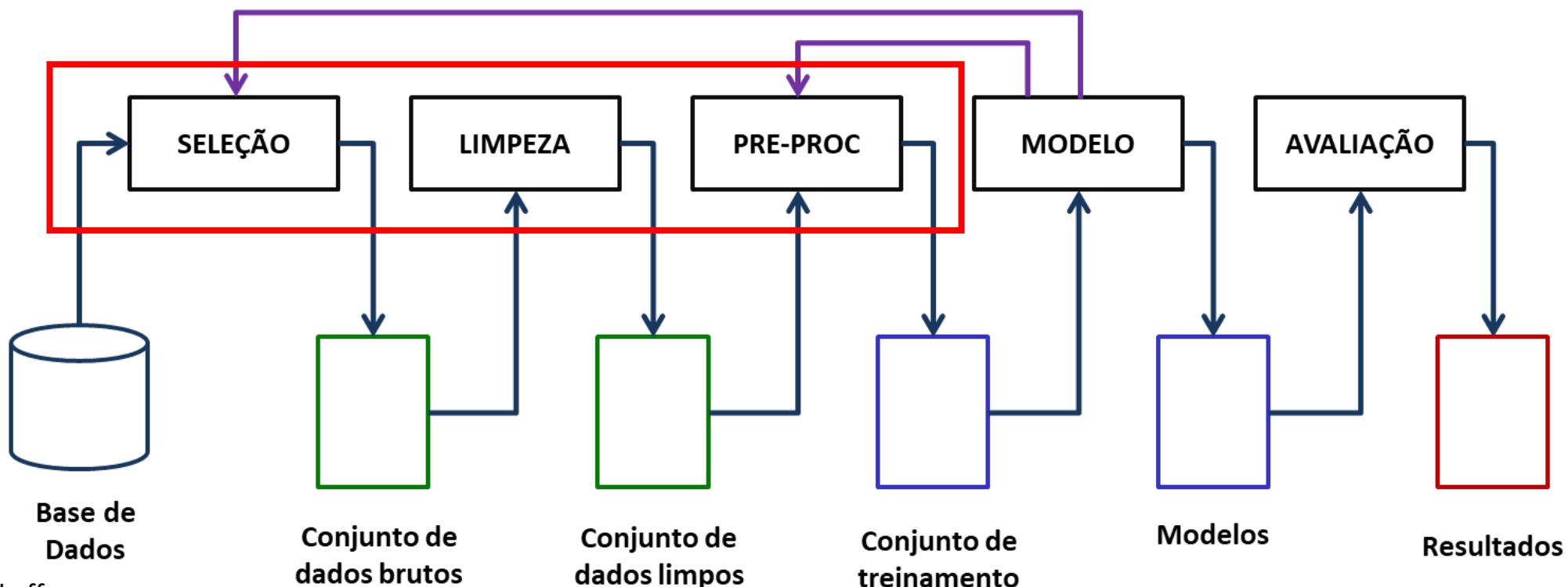
Preparação de dados



- Melhorar a qualidade dos dados disponíveis;
- Compreendê-los mais facilmente;
- Favorecer um aprendizado mais eficaz.

A importância

- os dados do mundo real são impuros (incompletos, ruidosos, inconsistentes);
- aprendizado com alto desempenho requer dados bem preparados;
- dados de qualidade possibilitam encontrar resultados de alta qualidade.





Tipos de dados



Análise Exploratória de Dados : caracterização de variáveis

Resume as características principais de uma base de dados usando estatísticas e visualização de dados.

Tendência central

Médias

Mediana

Moda

Posição

Mediana

Percentil

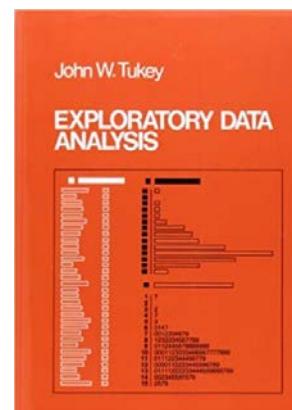
Quartil

Dispersão

Amplitude

Variância

Desvio-padrão

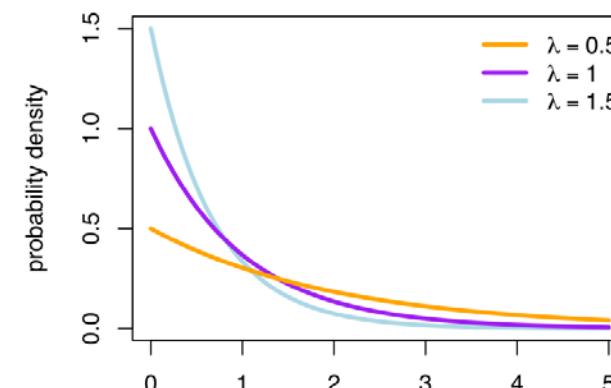
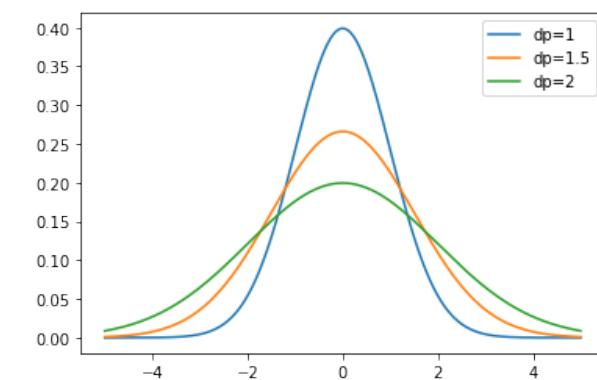
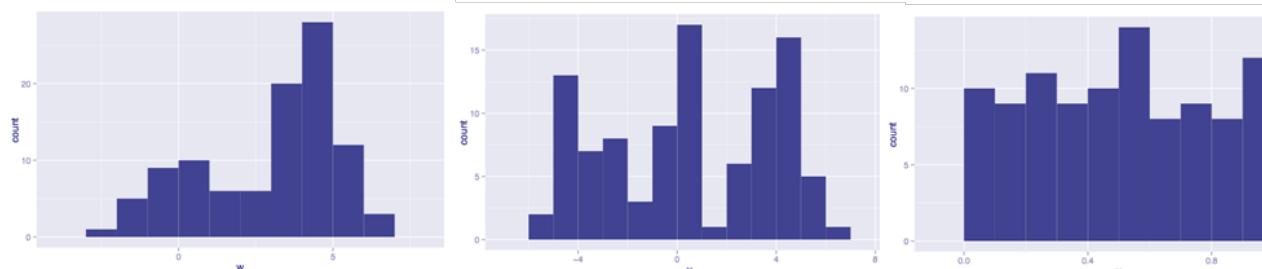
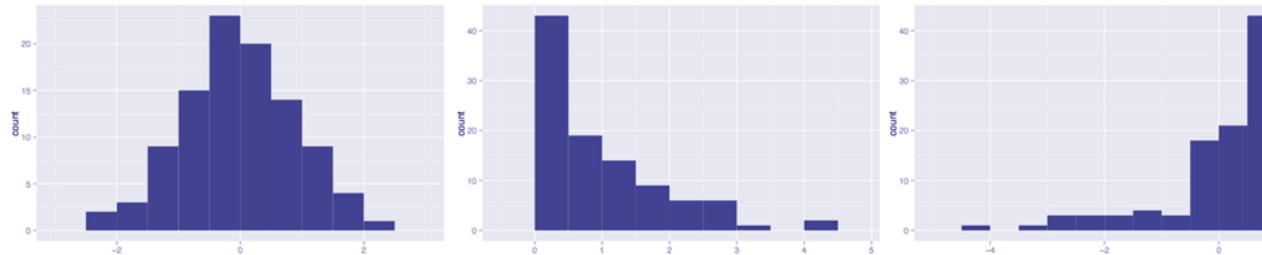




RPE
2025

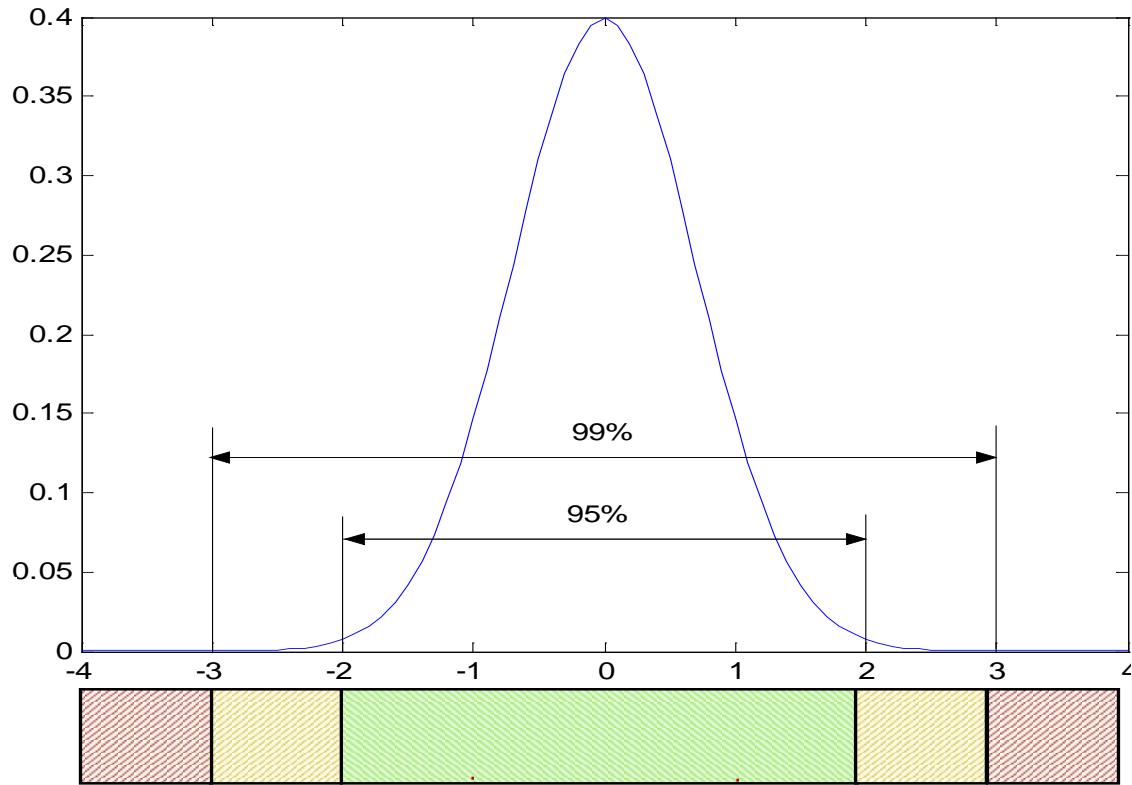
Preparação de dados

Análise Exploratória de Dados : caracterização de variáveis



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$



matplotlib.pyplot.hist
seaborn.distplot

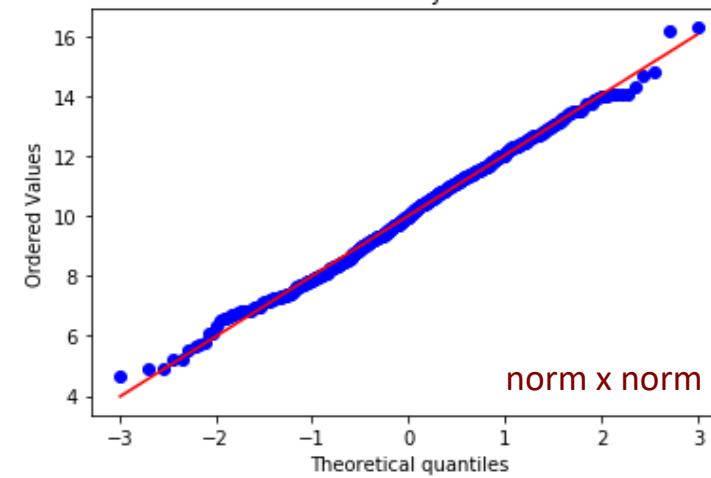


RPE
2025

Preparação de dados

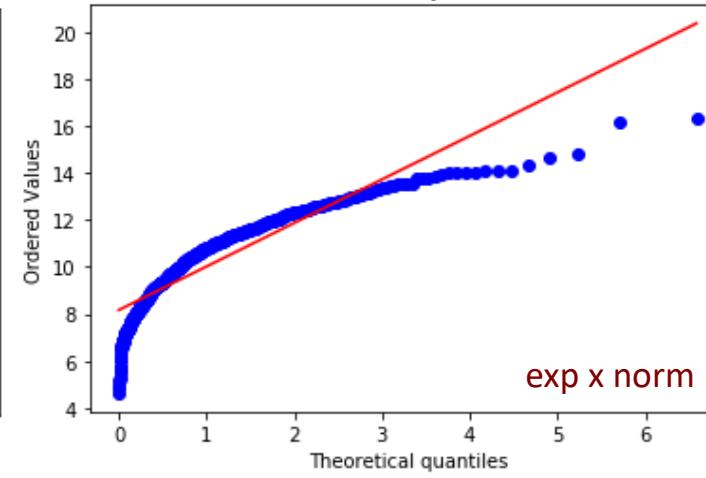
Análise Exploratória de Dados : caracterização de variáveis

Probability Plot



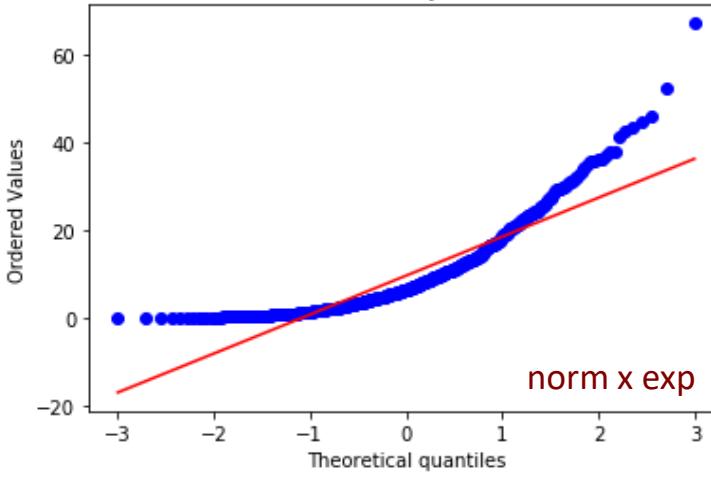
norm x norm

Probability Plot



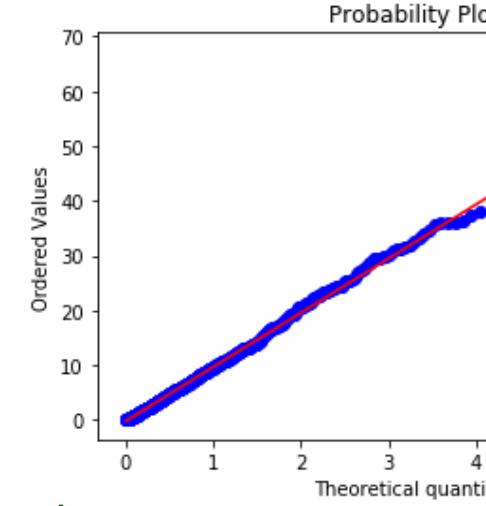
exp x norm

Probability Plot

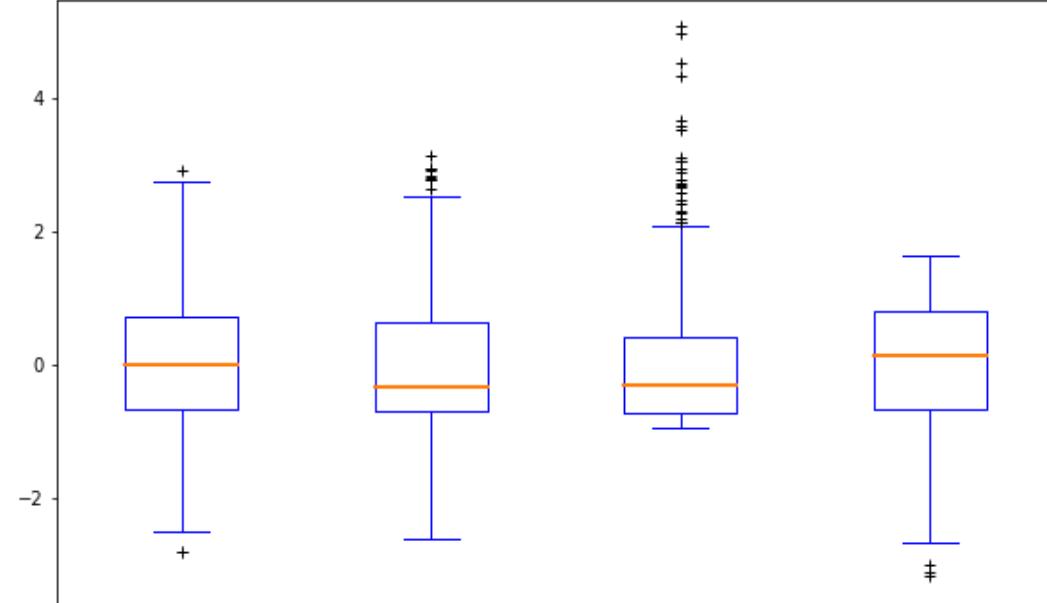


norm x exp

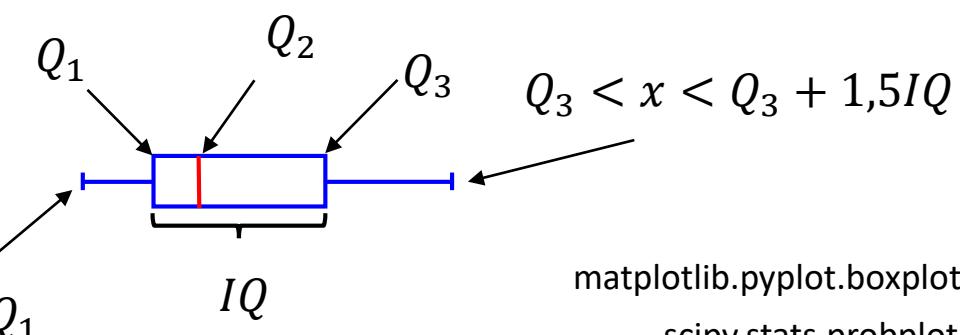
QQ plot



exp x exp



Box-plot



$Q_1 - 1,5IQ < x < Q_1$

IQ

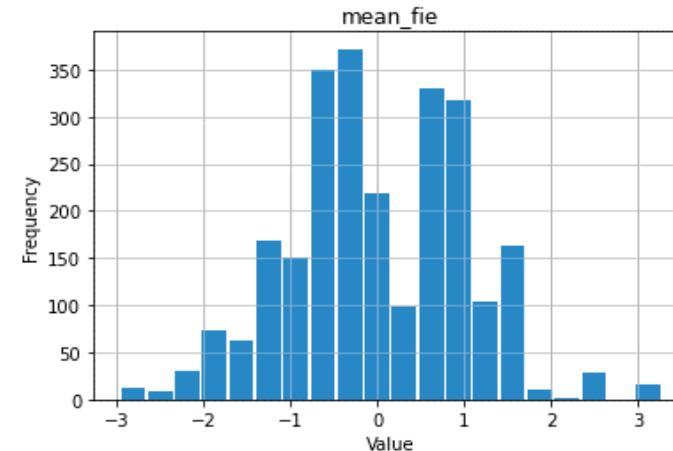
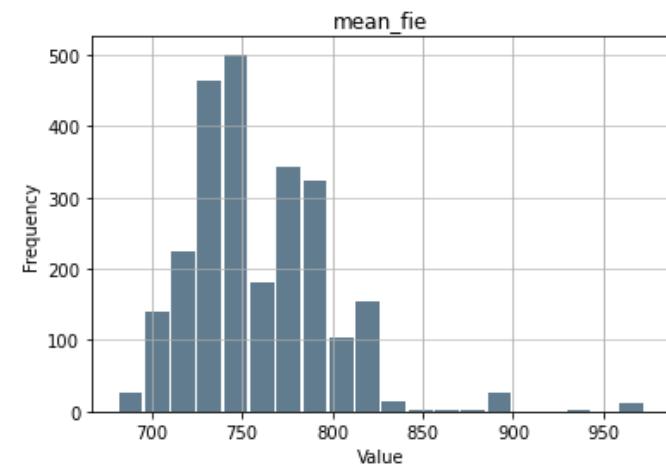
matplotlib.pyplot.boxplot
scipy.stats.probplot

Análise Exploratória de Dados : caracterização de variáveis

Eventualmente pode-se desejar reduzir a assimetria de uma amostra, tornando-a mais próxima de uma distribuição normal e minimizando interferências na variância. Neste caso, uma transformação de potência pode ser útil e uma das mais robustas é a de Yeo-Johnson, que não faz restrição aos valores que a variável pode assumir:

$$\varphi(y, \lambda) = \begin{cases} \frac{((y + 1)^\lambda - 1)}{\lambda} & , \text{se } \lambda \neq 0 \text{ e } y \geq 0 \\ \log(y + 1) & , \text{se } \lambda = 0 \text{ e } y \geq 0 \\ -\frac{((-y + 1)^{2-\lambda} - 1)}{2-\lambda} & , \text{se } \lambda \neq 2 \text{ e } y < 0 \\ -\log(-y + 1) & , \text{se } \lambda = 2 \text{ e } y < 0 \end{cases}$$

$$y(\varphi, \lambda) = \begin{cases} (\varphi \cdot \lambda + 1)^{1/\lambda} - 1 & , \text{se } \lambda \neq 0 \text{ e } y \geq 0 \\ e^\varphi - 1 & , \text{se } \lambda = 0 \text{ e } y \geq 0 \\ 1 - (-(2 - \lambda) \cdot \varphi + 1)^{1/2-\lambda} & , \text{se } \lambda \neq 2 \text{ e } y < 0 \\ 1 - e^{-\varphi} & , \text{se } \lambda = 2 \text{ e } y < 0 \end{cases}$$





RPE
2025

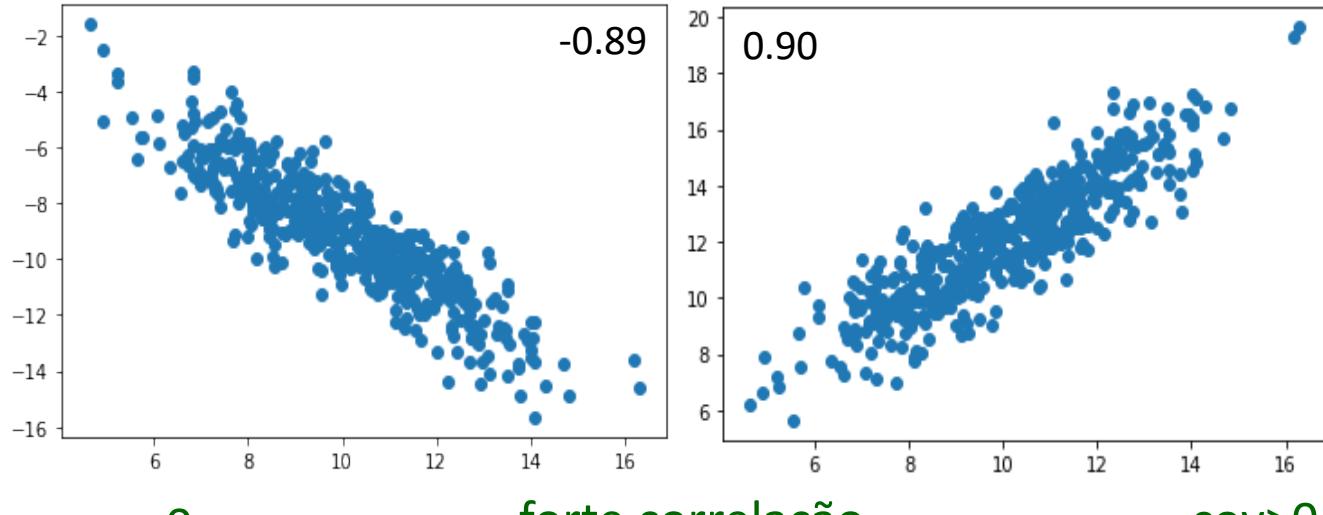
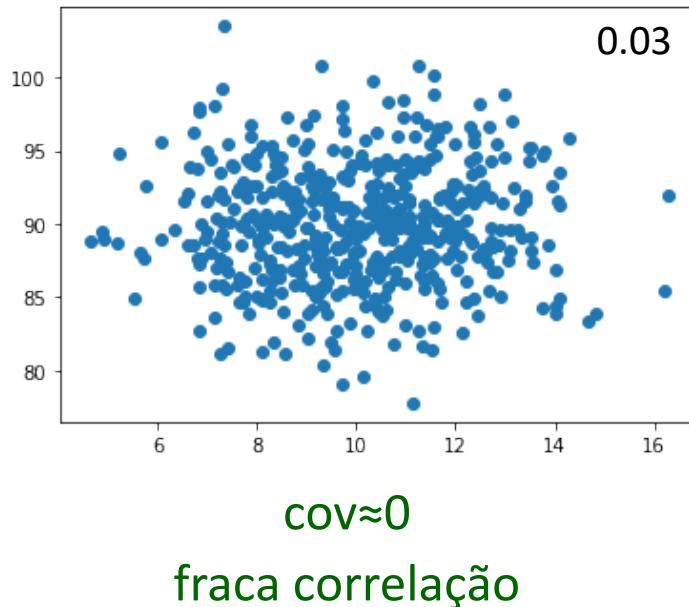
Preparação de dados

Análise Exploratória de Dados: variação conjunta de variáveis

Covariância e correlação de Pearson:

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

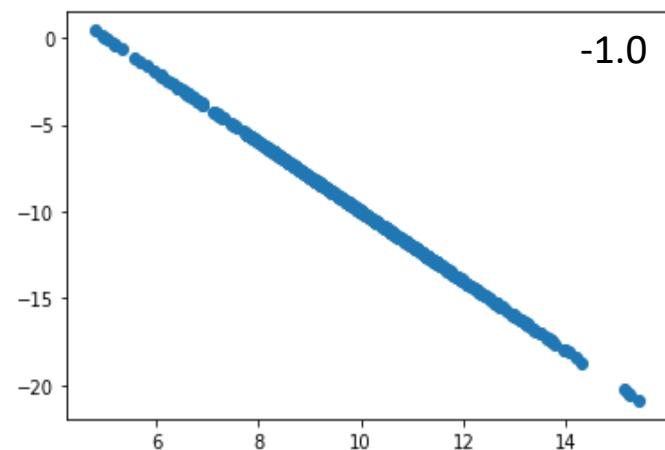
$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$



cov<0

forte correlação

cov>0



numpy.cov
numpy.corrcoef

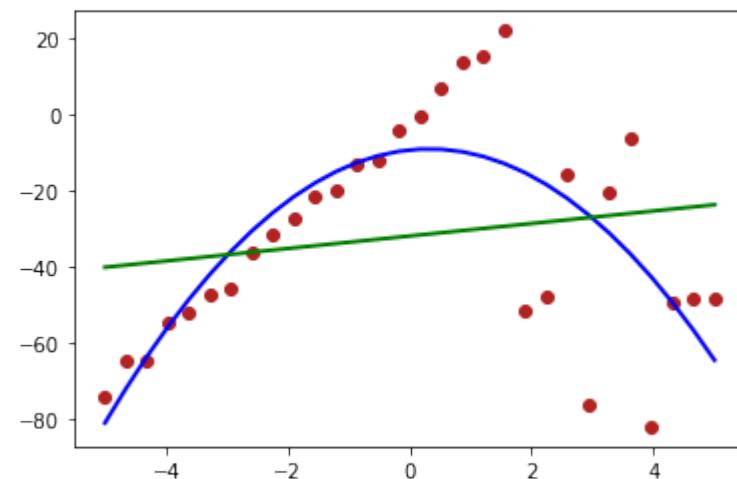
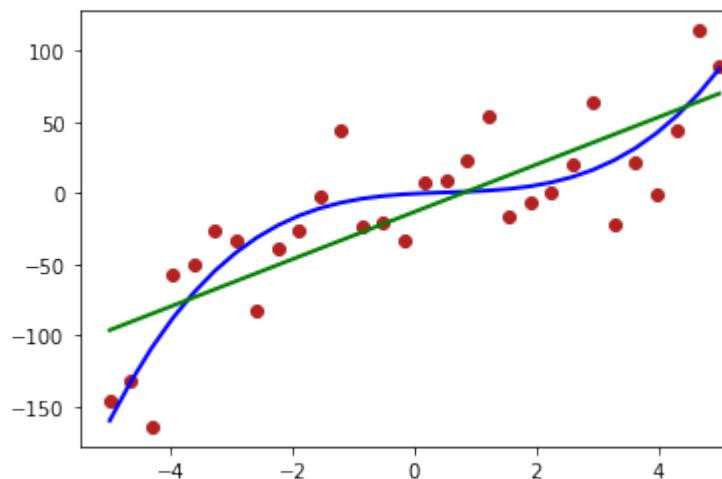
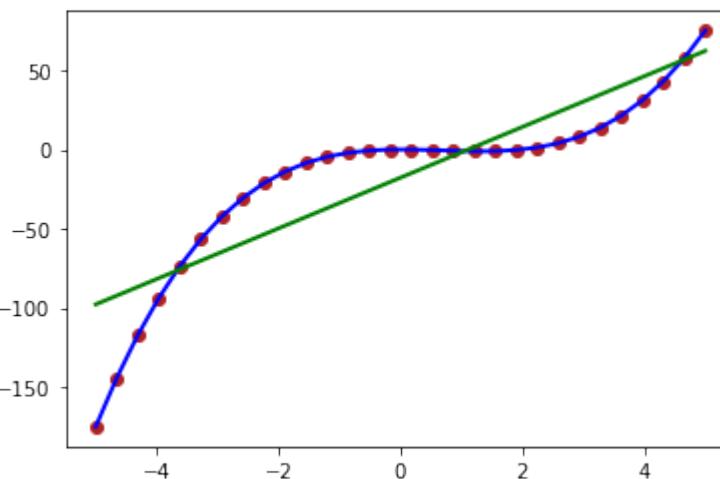


Análise Exploratória de Dados: variação conjunta de variáveis

Correlações de Spearman:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(\text{ordem}_X, \text{ordem}_Y)}{\sigma_{\text{ordem}_X} \cdot \sigma_{\text{ordem}_Y}}$$

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$



Análise Exploratória de Dados: variação conjunta de variáveis

Quarteto de Anscombe

Graphs in Statistical Analysis*

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. Usefulness of graphs

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

- (1) numerical calculations are exact, but graphs are rough;
- (2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
- (3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

Graphs can have various purposes, such as: (i) to help us perceive and appreciate some broad features of the data, (ii) to let us look behind those broad features and see what else is there. Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.

Good statistical analysis is not a purely routine matter, and generally calls for more than one pass

through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

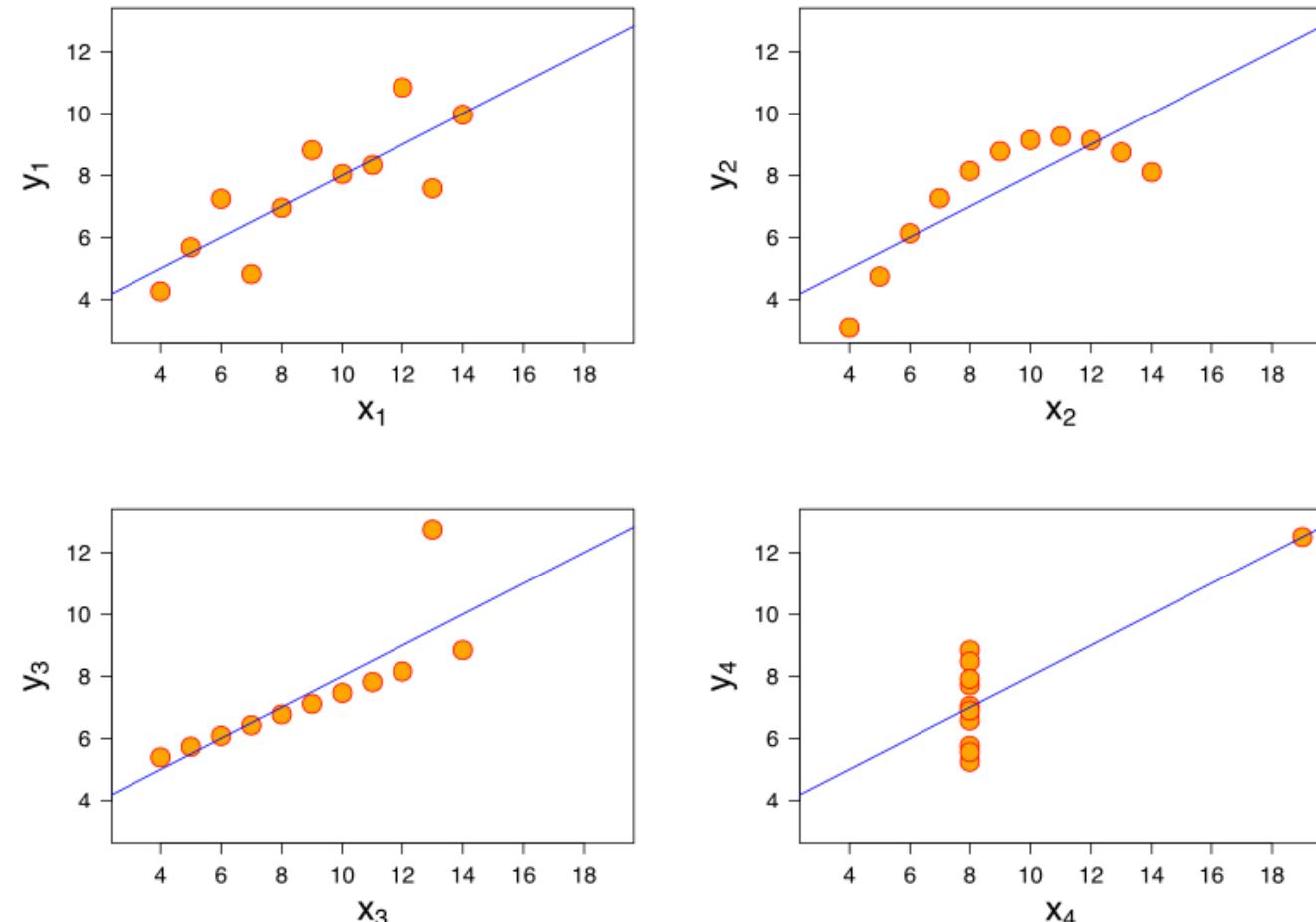
Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

A few simple types of statistical analysis are now considered.

2. Regression analysis—the simplest case

Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor) variable x . Before anything else is done, we should scatterplot the y values against the x values and see what sort of relation there is—if any. Many different kinds of things can happen:

- (1) the (x, y) points lie nearly on a straight line;
- (2) the (x, y) points lie nearly on a smooth curve, not a straight line;
- (3) the y -values are scattered, without relation to the x -values;
- (4) something intermediate between (1) or (2) and (3);
- (5) most of the (x, y) points lie close to a line or smooth curve, but a few are scattered a long way away.



Mesmas médias, desvios e correlações!

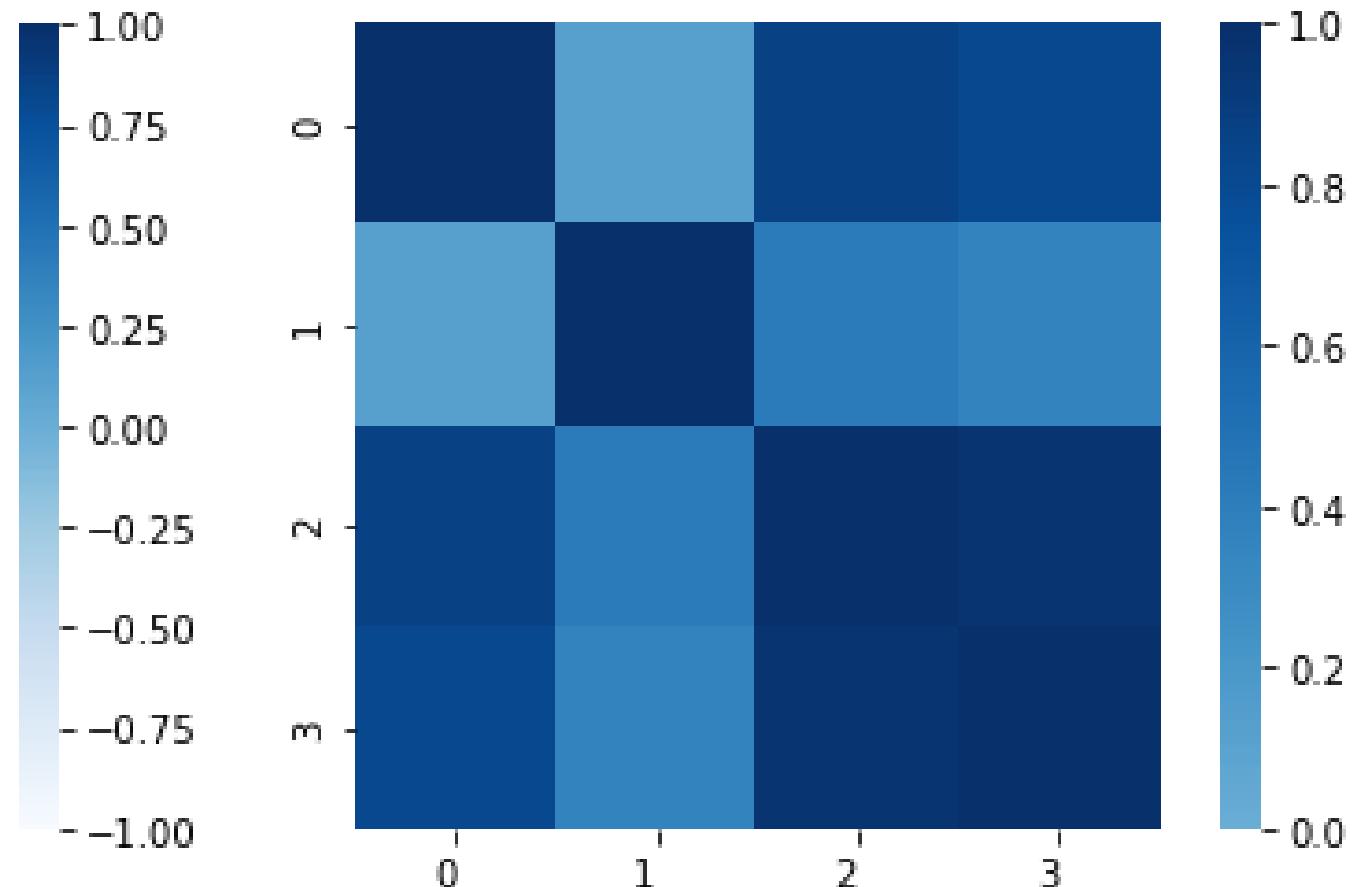
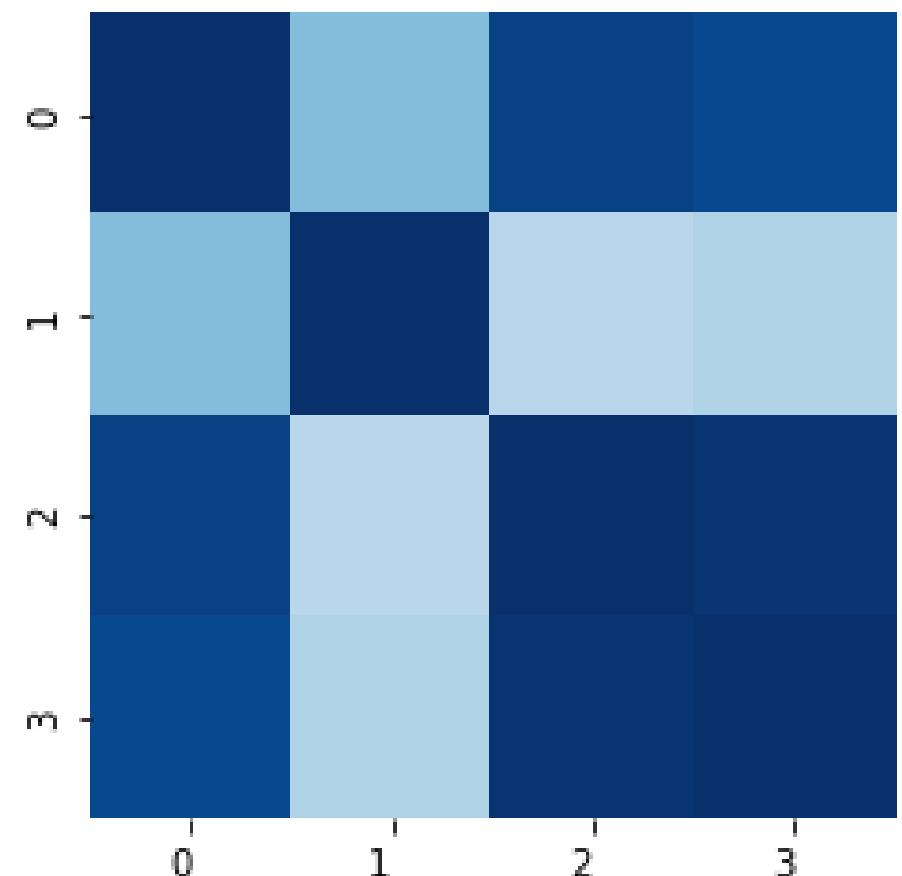


RPE
2025

Preparação de dados

Análise Exploratória de Dados: variação conjunta de variáveis

Mapa de calor da matriz de correlação



seaborn.heatmap

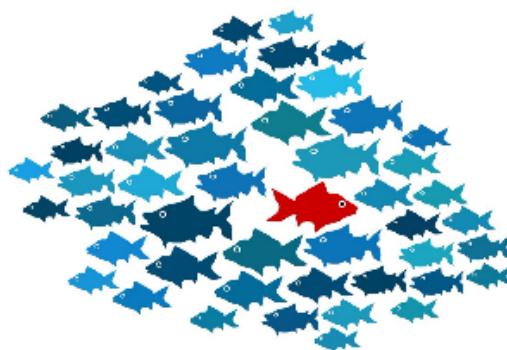


Análise Exploratória de Dados: valores discrepantes

Um valor **discrepante (outlier)** é um valor muito diferente dos demais que sugere ter sido gerado por outro processo. A identificação de *outliers* pode ser usada em detecção de fraudes, falhas e na limpeza de dados.

Tratamentos:

- ➊ ignorá-los;
- ➋ discretizar a variável;
- ➌ eliminar o registro;
- ➍ removê-lo e preencher o vazio produzido.

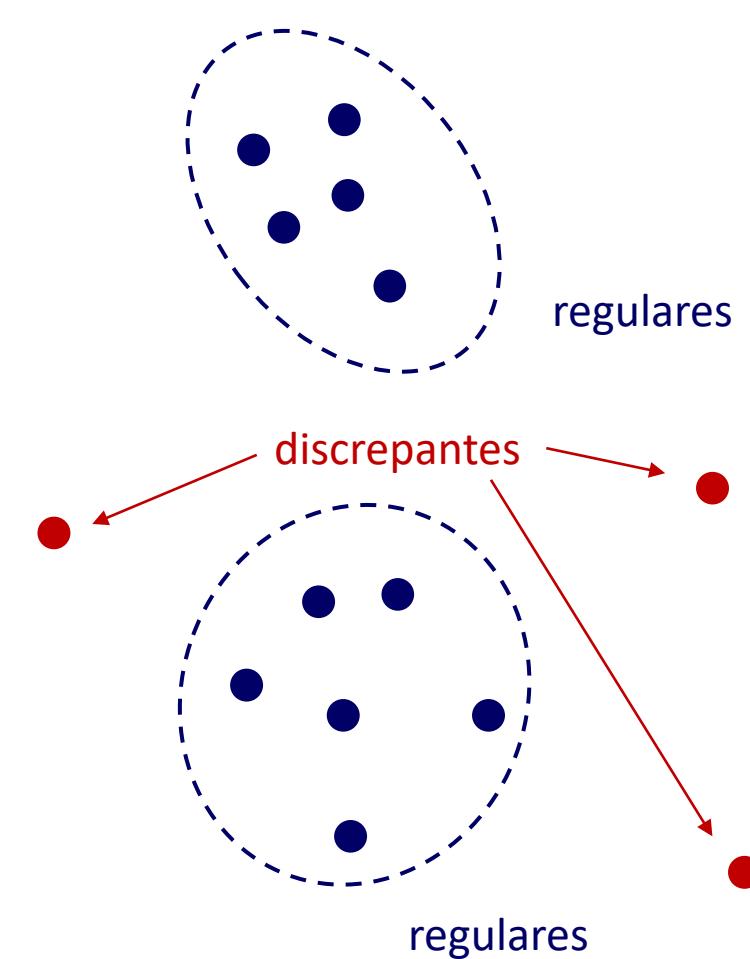
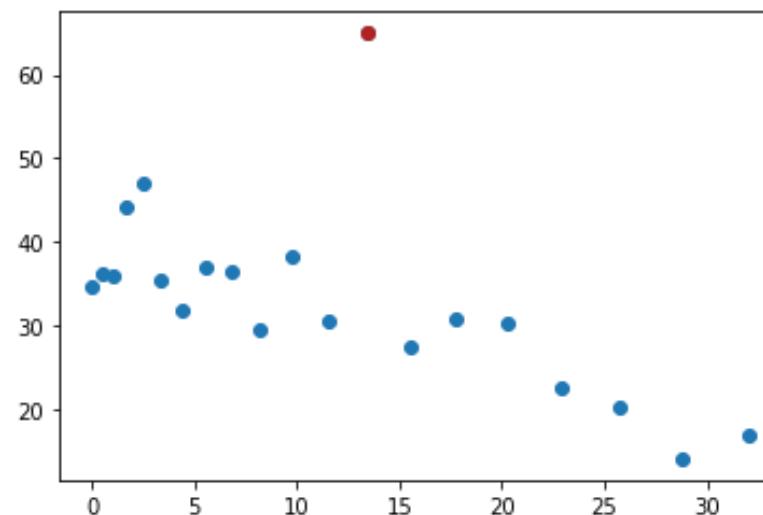
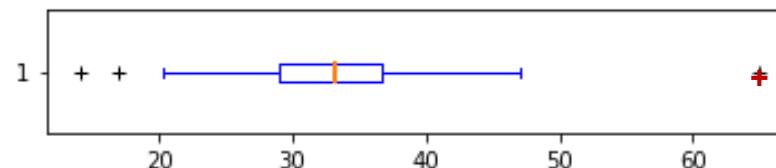
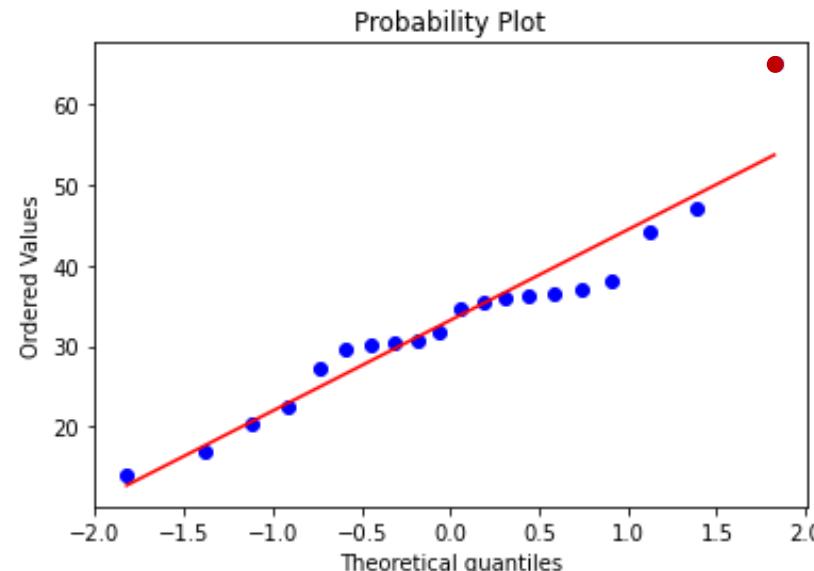
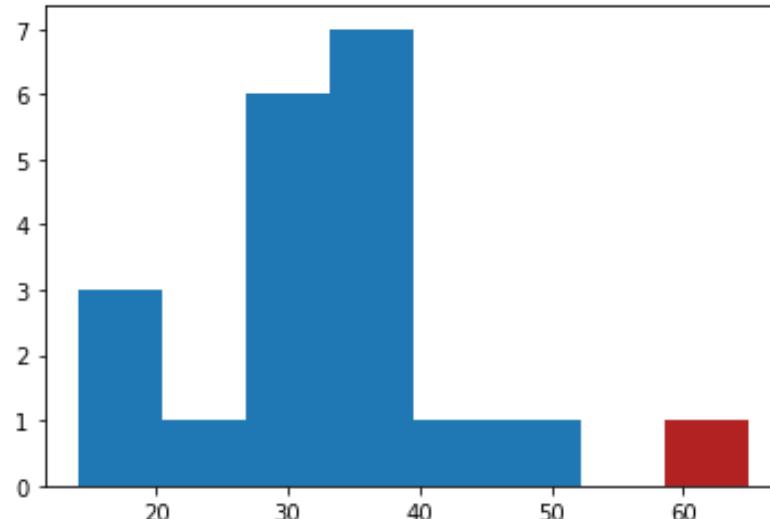


Possíveis origens:

- ➊ Erros na aquisição (instrumentação, digitação);
- ➋ Erros nos experimentos geradores;
- ➌ Erros de manipulação durante os processamentos anteriores (integração, transformação etc.);
- ➍ Resultados de amostragens descontroladas (mal planejadas e verificadas);
- ➎ Valores corretos, mas raros.

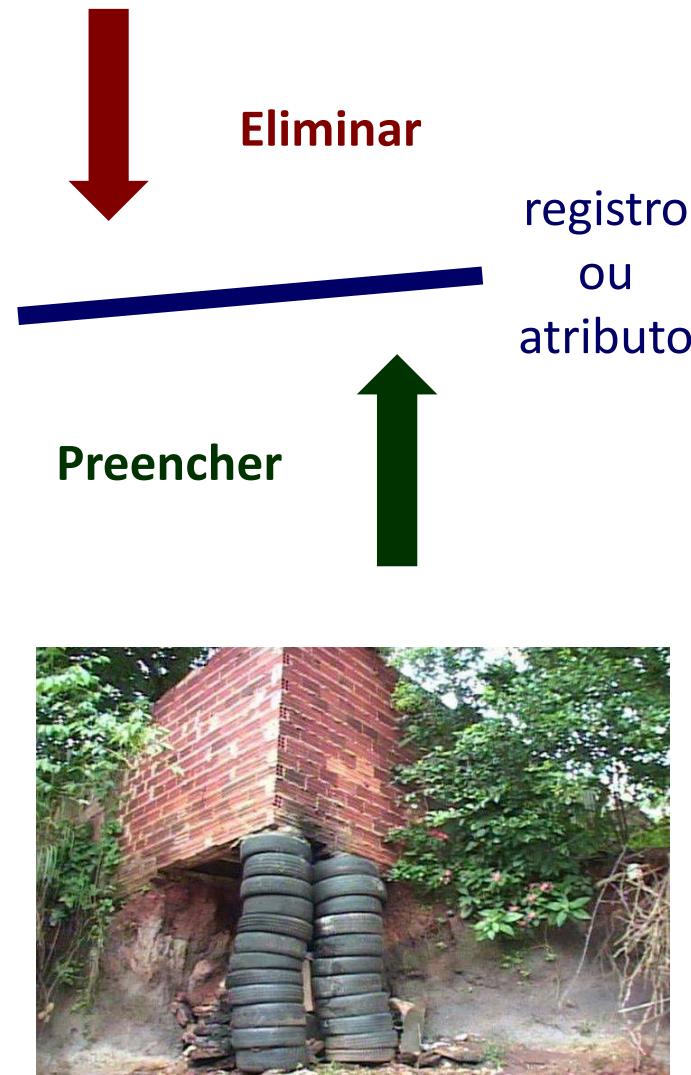


Análise Exploratória de Dados: valores discrepantes



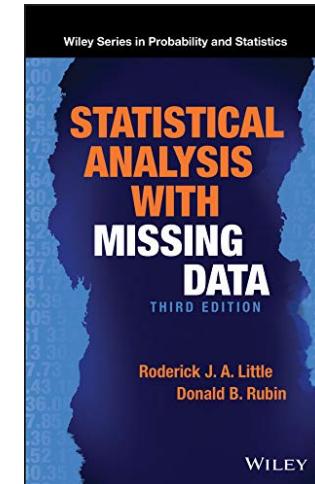
Análise Exploratória de Dados: valores ausentes

	IDADE	SALDO	RENDA	TEMPO
0	94	2000	1500	26
1	39	-5000		3
2	64	8000		14
3	64			
4	62	7000	9000	29
5	54	1000		30
6	56	-2000		15
7	60	5000		24
8	39	3500		19
9	18	500		23
10	46	6000		2
11	22	-24000	6500	6
12	64	5400	2800	14
13	46	-1245	2200	
14	19	4123	3400	24



Além do processamento adicional, a tarefa de substituição dos vazios nos dados pode enviesar os modelos e reduzir a eficiência em geral. Entretanto, em um cenário de poucos dados, o custo-benefício é maior se o preenchimento for matematicamente viável.

- valor constante;
- valor aleatório;
- moda, mediana ou média;
- desvio-padrão constante;
- k-vizinhos mais próximos;
- qualquer modelo preditivo;
- extração e interpolação;
- imputação múltipla.



Análise Exploratória de Dados: transformação de variáveis

É comum que algumas variáveis não sejam numéricas e seu algoritmo de modelagem não saiba como lidar com elas. Para aproveitá-las é necessário convertê-la em uma outra numérica.

O caso mais simples é quando a variável original é **ordinal**. Basta transformá-la em uma **numérica discreta**.

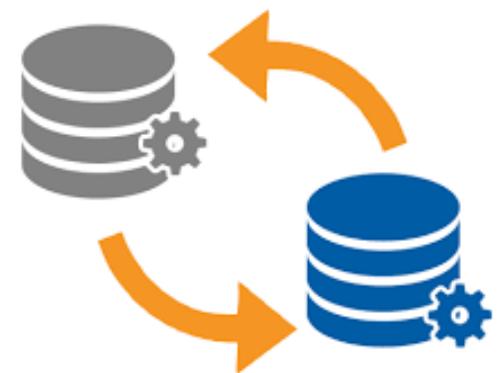
Ex.: escolaridade: { Analfabeto, ..., Doutorado} \Rightarrow escolaridade: { 0, ..., 8}

Quando não há ordem, a relação 1:1 vista acima dificilmente ocorre. Neste caso, é preciso maior cuidado ao fazer o mapeamento para evitar a inclusão de uma ordem artificialmente.

Ex.: UTI: { Vaga, Ocupada} \Rightarrow VagaUTI: {0, 1}

Em outros casos, o algoritmo de modelagem não trabalha com conjuntos numéricos contínuos ou se deseja trabalhar com menos valores. A solução é a **discretização** da variável em blocos ou intervalos. A vantagem é a redução da complexidade e do espaço de busca.

Ex.: O intervalo [-100,100] pode ser discretizado em valores *Positivos* e *Não-positivos*



`preprocessing.OrdinalEncoder`
`preprocessing.OneHotEncoder`

Análise Exploratória de Dados: padronização

A padronização de dados procura uniformizar os domínios das variáveis (independentes, principalmente) de modo que o aprendizado não seja prejudicado em virtude dos diferentes domínios das variáveis (efeito de escala).

$$x'_i = \frac{x_i}{\max. x_i}$$

máximo

$$x'_i = \frac{x_i - \min. x_i}{\max. x_i - \min. x_i}$$

amplitude

$$x'_i = \frac{x_i - \mu(x_i)}{\alpha \cdot \sigma(x_i)}$$

$$x'_i = \frac{x_i - P_{50}}{\alpha \cdot (P_{75} - P_{25})}$$

Gaussianas,
padrão,
z-score

robusta

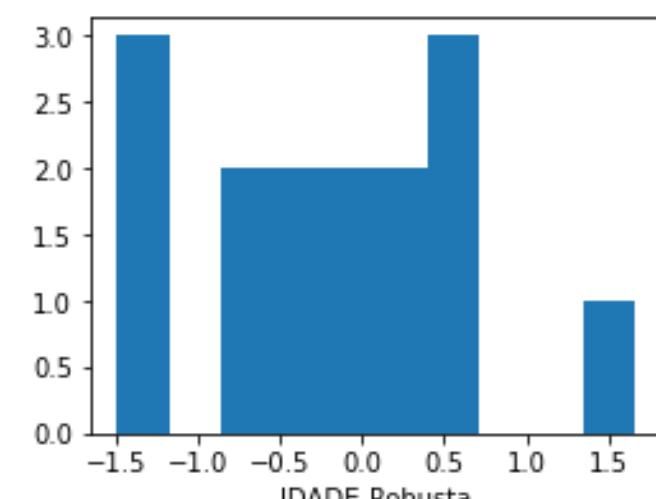
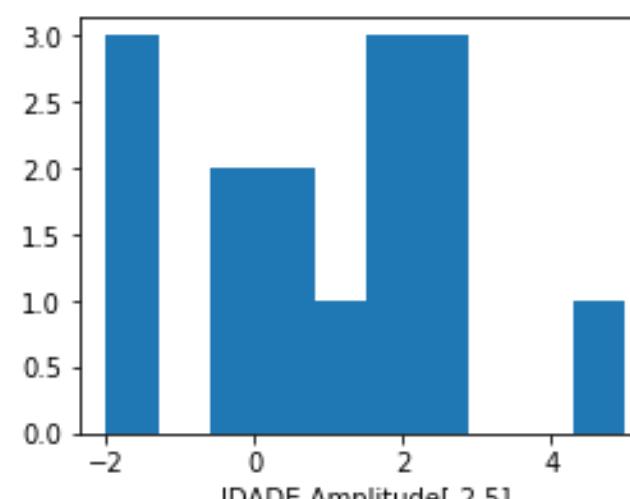
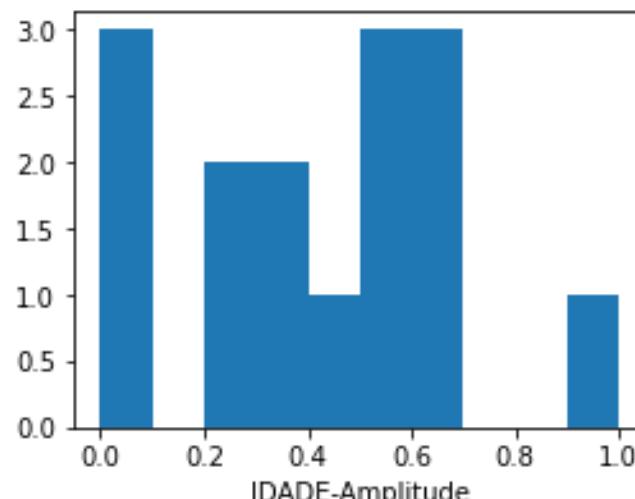
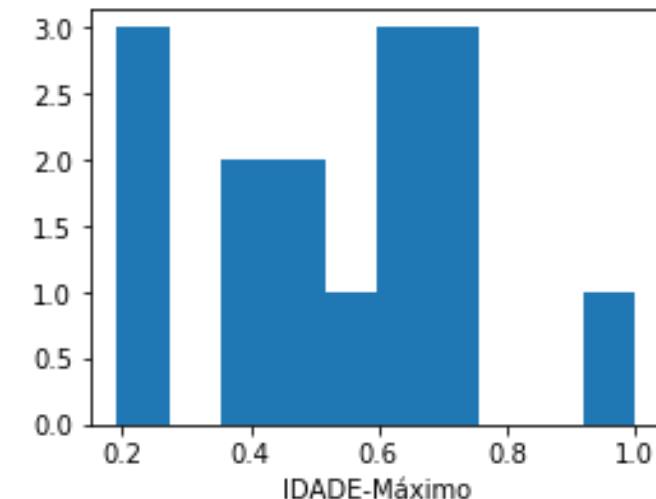
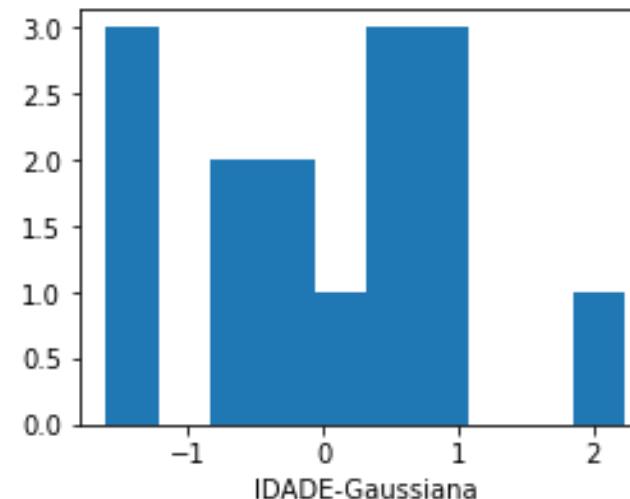
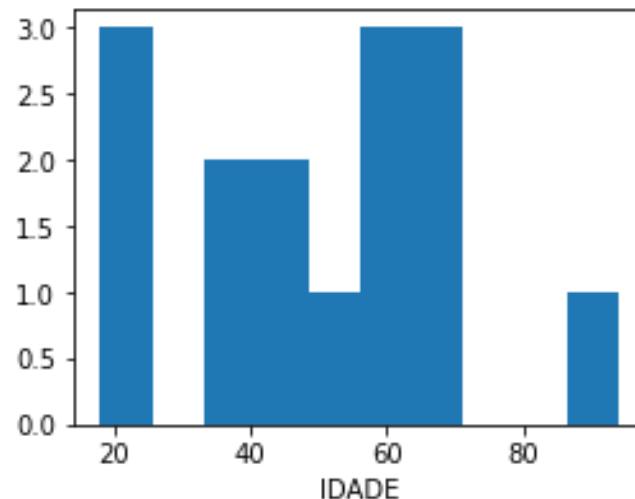
	IDADE	IDADE-Gaussianas	IDADE-Máximo	IDADE-Amplitude	IDADE-Amplitude[-2,5]	IDADE-Robusta
0	94	2.235786	1.000000	1.000000	5.000000	1.666667
1	39	-0.546301	0.414894	0.276316	-0.065789	-0.625000
2	64	0.718284	0.680851	0.605263	2.236842	0.416667
3	64	0.718284	0.680851	0.605263	2.236842	0.416667
4	62	0.617117	0.659574	0.578947	2.052632	0.333333
5	54	0.212450	0.574468	0.473684	1.315789	0.000000
6	56	0.313617	0.595745	0.500000	1.500000	0.083333
7	60	0.515951	0.638298	0.552632	1.868421	0.250000
8	39	-0.546301	0.414894	0.276316	-0.065789	-0.625000
9	18	-1.608552	0.191489	0.000000	-2.000000	-1.500000
10	46	-0.192217	0.489362	0.368421	0.578947	-0.333333
11	22	-1.406218	0.234043	0.052632	-1.631579	-1.333333
12	64	0.718284	0.680851	0.605263	2.236842	0.416667
13	46	-0.192217	0.489362	0.368421	0.578947	-0.333333
14	19	-1.557968	0.202128	0.013158	-1.907895	-1.458333



RPE
2025

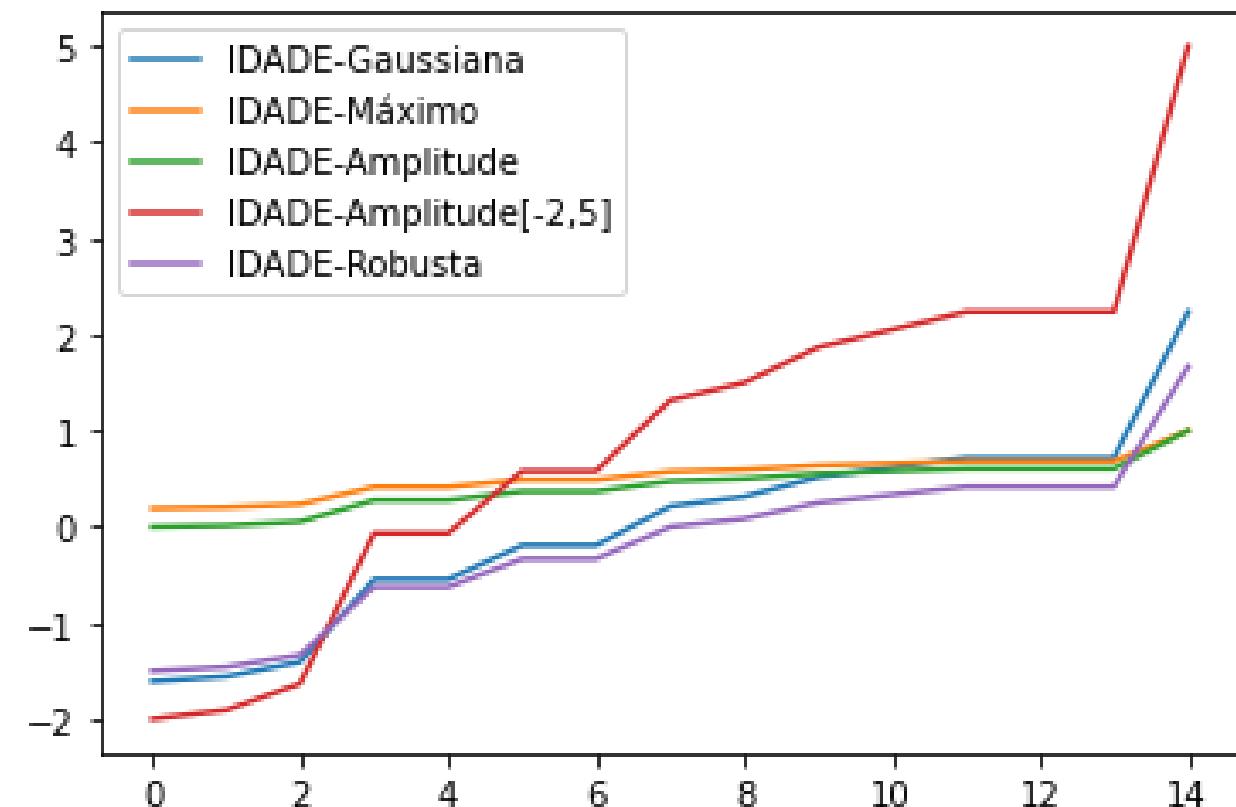
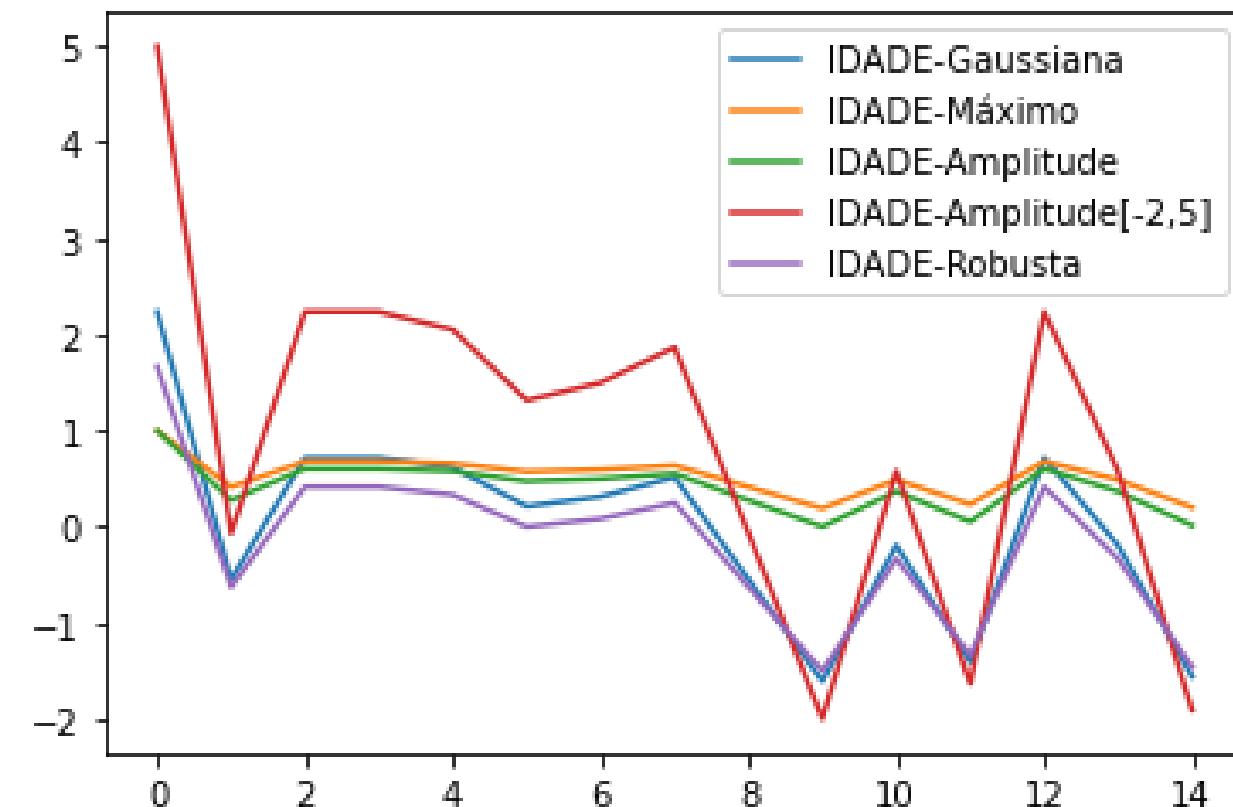
Preparação de dados

Análise Exploratória de Dados: padronização





Análise Exploratória de Dados: padronização



Em geral, as padronizações **gaussiana** e **robusta** resultam em uma variabilidade maior do que as demais (desconsidere o exemplo **amplitude[-2,5]**, que foi criado para fins didáticos).

Análise Exploratória de Dados: padronização

	IDADE	SALDO	vizinho	IDADE-Gaussian	SALDO-Gaussian	vizinho	IDADE-Amplitude	SALDO-Amplitude	vizinho
0	94	2000	5	2.235786	0.175995	3	1.000000	0.812500	3
1	39	-5000	6	-0.546301	-0.761001	13	0.276316	0.593750	13
2	64	8000	4	0.718284	0.979134	4	0.605263	1.000000	4
3	64	0	9	0.718284	-0.091719	6	0.605263	0.750000	6
4	62	7000	2	0.617117	0.845277	2	0.578947	0.968750	2
5	54	1000	9	0.212450	0.042138	6	0.473684	0.781250	6
6	56	-2000	13	0.313617	-0.359432	5	0.500000	0.687500	5
7	60	5000	12	0.515951	0.577564	12	0.552632	0.906250	12
8	39	3500	14	-0.546301	0.376779	10	0.276316	0.859375	10
9	18	500	5	-1.608552	-0.024790	14	0.000000	0.765625	14
10	46	6000	12	-0.192217	0.711421	8	0.368421	0.937500	8
11	22	-24000	1	-1.406218	-3.304276	1	0.052632	0.000000	1
12	64	5400	7	0.718284	0.631107	7	0.605263	0.918750	7
13	46	-1245	6	-0.192217	-0.258370	5	0.368421	0.711094	5
14	19	4123	8	-1.557968	0.460172	9	0.013158	0.878844	9

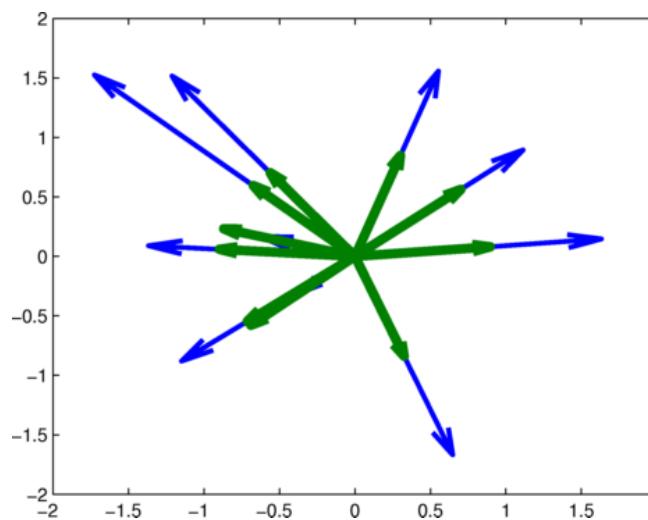
Observa-se que a identificação do vizinho mais próximo de cada registro muda antes e depois das padronizações devido ao peso exagerado da variável ‘RENDAS’ em relação à ‘IDADE’. Porém, pouco ou nada muda entre as diferentes padronizações. Ao se interpretar as vizinhanças após as padronizações, nota-se maior razoabilidade na associação.



Análise Exploratória de Dados: normalização

A normalização de dados transforma cada registro em um vetor de norma unitária.

	IDADE	SALDO	IDADE	SALDO
0	94	2000	0.046948	0.998897
1	39	-5000	0.007800	-0.999970
2	64	8000	0.008000	0.999968
3	64	0	1.000000	0.000000
4	62	7000	0.008857	0.999961
5	54	1000	0.053921	0.998545
6	56	-2000	0.027989	-0.999608
7	60	5000	0.011999	0.999928
8	39	3500	0.011142	0.999938
9	18	500	0.035977	0.999353
10	46	6000	0.007666	0.999971
11	22	-24000	0.000917	-1.000000
12	64	5400	0.011851	0.999930
13	46	-1245	0.036923	-0.999318
14	19	4123	0.004608	0.999989



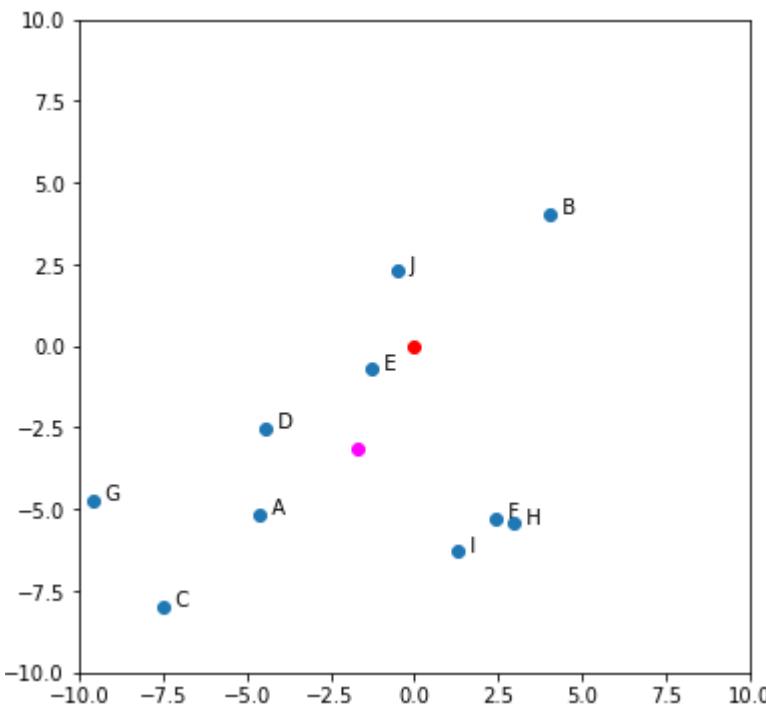
$$x'_i = \frac{x_i}{\|x\|}$$

No exemplo, observa-se que muitos registros se aproximaram após a normalização, um efeito indesejável e que ocorre devido aos domínios bastante diferentes (amplitudes e elementos).

Neste caso, uma padronização realizada antes da normalização evitaria tal comportamento.

Análise Exploratória de Dados : análise de componentes principais

A análise de componentes principais é uma metodologia de transformação do espaço original dos dados (possivelmente correlacionados) em um espaço de dimensões não correlacionadas chamadas de **componentes principais**. Espera-se gerar um espaço de dimensões reduzidas mantendo-se a maior informação estatística possível (variabilidade).



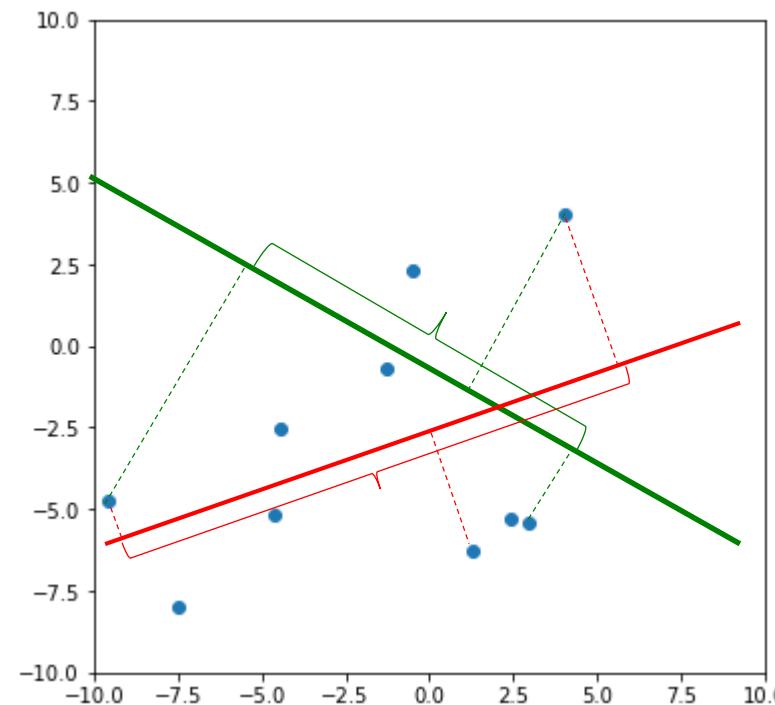
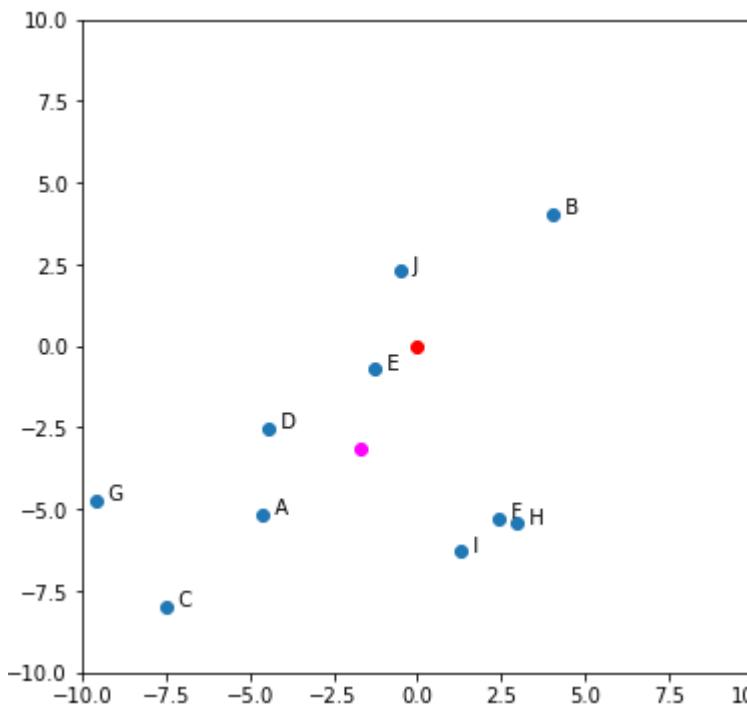
A “manutenção” da variabilidade original será alcançada com novas variáveis, as componentes principais, ortogonais entre si, que são combinações lineares das variáveis originais e que são obtidas a partir da maximização de suas variâncias.

Inicialmente, será buscada a combinação linear das variáveis originais que possua a máxima variabilidade. Essa será a 1^a componente e as demais são obtidos ortogonalmente aos anteriores, sucessivamente capturando a máxima variabilidade.

Análise Exploratória de Dados : análise de componentes principais

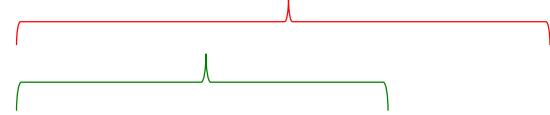
Assim, procura-se uma transformação linear que promova a redução de dimensionalidade do problema sem perder informações sobre a variabilidade dos dados.

Seja A esta transformação dos dados originais X , resultando nos dados projetados Y : $Y = AX$.
 As covariâncias de Y podem ser escritas em função das covariâncias de X : $Cov_Y = A Cov_X A^T$



$$Cov(Y) = E[(Y - E[Y])(Y - E[Y])^T]$$

Qual reta é a que melhor captura a variabilidade dos dados?



Análise Exploratória de Dados : análise de componentes principais

Como uma covariância estima o grau de relacionamento linear entre duas variáveis, quanto maior o seu valor, maior a redundância entre elas.

Observando a matriz de covariâncias Cov_Y , deseja-se que ela tenha:

- valores da diagonal principal (variâncias) não nulos, porque indicam variáveis com dispersão;
- valores fora da diagonal principal (covariâncias) nulos ou bem próximos de zero, o que indica baixa redundância entre as variáveis.

Em uma transformação ideal, em que as variáveis resultantes são perfeitamente não correlacionadas, Cov_Y seria uma matriz diagonal e é isso o que se deseja. Como Y não está disponível, o objetivo será alcançado a partir da diagonalização de Cov_X .

Como Cov_X é simétrica e real, o teorema espectral diz que ela pode ser diagonalizada com uma matriz V de autovetores, neste caso **ortonormais**, e por uma matriz diagonal Λ de autovalores:

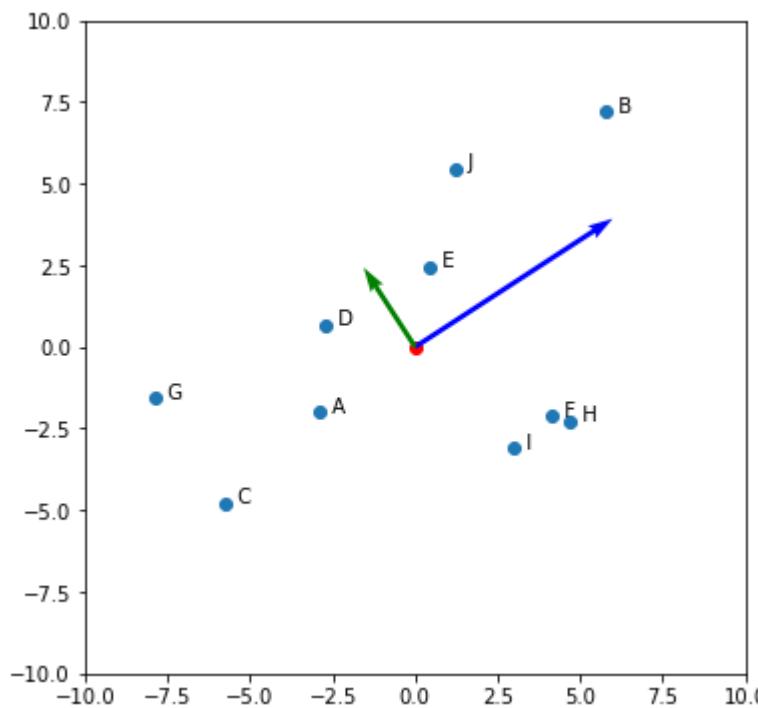
$$\text{Cov}_X = V \cdot \Lambda \cdot V^{-1} = V \cdot \Lambda \cdot V^T$$

$$\text{Cov}_Y = V^T \text{Cov}_X V = V^T V \cdot \Lambda \cdot V^T V = \Lambda$$

Análise Exploratória de Dados : análise de componentes principais

Dessa forma, o problema se torna encontrar os autovetores e autovalores de Cov_X . $\text{Cov}_X \mathbf{V} = \Lambda \cdot \mathbf{V}$

Feita a diagonalização de Cov_X , os autovetores ortonormais encontrados são os componentes principais que formam a base vetorial do espaço transformado. Por sua vez, os autovalores informam as quantidades de variabilidade capturada por cada autovetor.



Assim, com a normalidade dos autovetores, evita-se problemas de mudança de escala após a transformação, além de promover a interpretação relativa dos componentes principais.

O autovetor de Cov_X associado ao maior autovalor sempre aponta para a direção onde os dados originais mais variam. A magnitude desse vetor é o autovalor associado.

O autovetor relacionado ao segundo maior autovalor indica a direção da segunda maior variação dos dados e será ortogonal ao primeiro (devido à simetria de Cov_X). E assim sucessivamente.

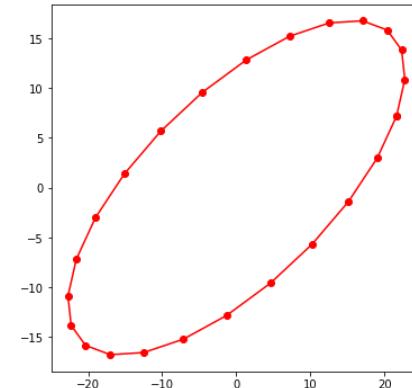
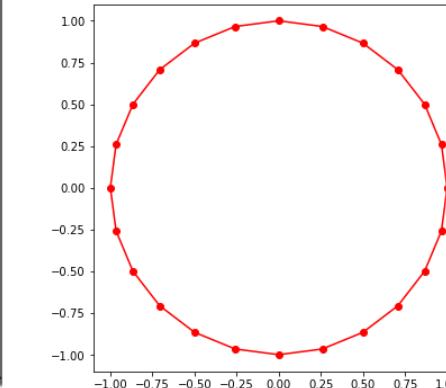
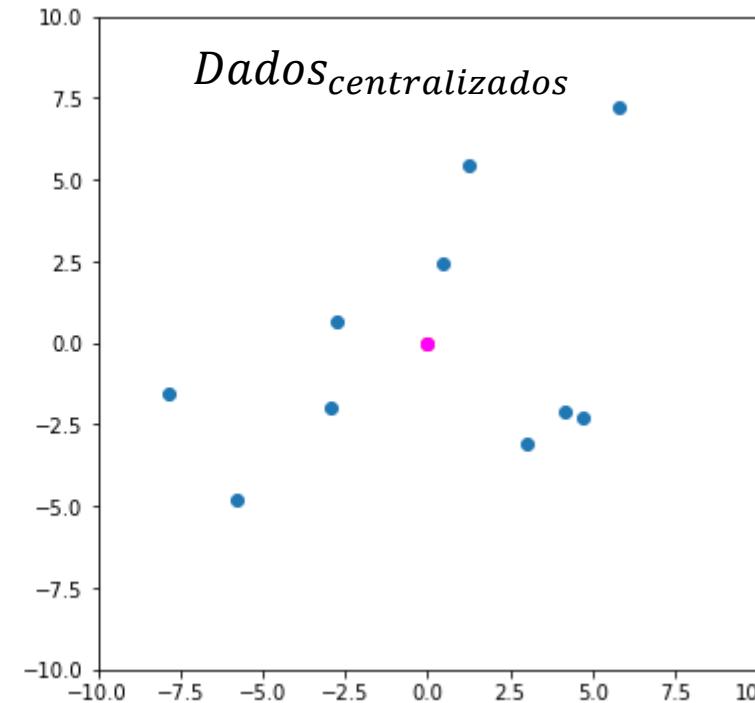
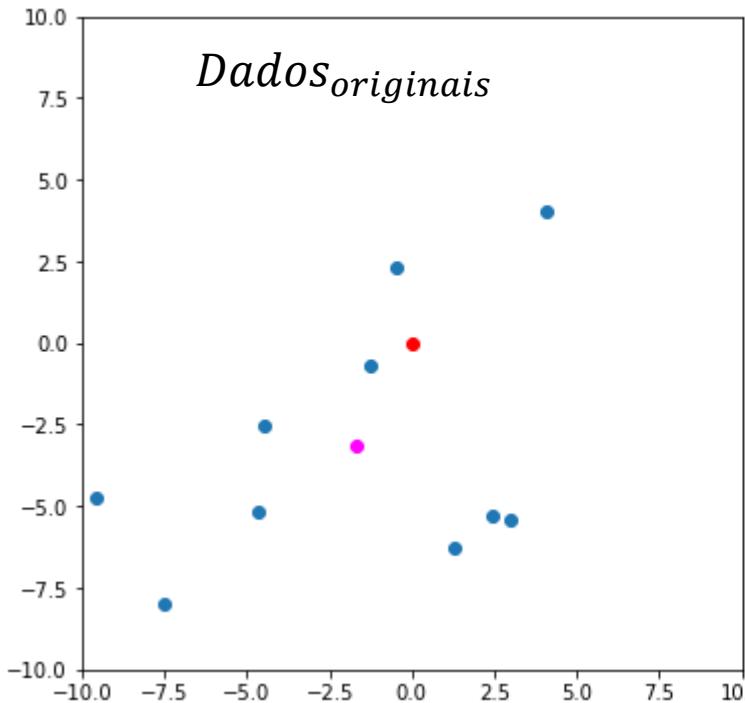
RPE
2025

Preparação de dados

Análise Exploratória de Dados : análise de componentes principais

Passos:

1. centralizar os dados e calcular a matriz de covariâncias (Σ)



$$\mathbf{Cov}_X = \Sigma = \begin{bmatrix} Var(x) & Cov(x, y) \\ Cov(x, y) & Var(y) \end{bmatrix} = \begin{bmatrix} 21.5920 & 7.1726 \\ 7.1726 & 15.2018 \end{bmatrix}$$

Análise Exploratória de Dados : análise de componentes principais

Passos:

2. Encontrar os autovalores e autovetores da matriz de covariâncias

$$\Sigma = \text{Cov}_X$$

$$\Sigma \cdot V = \Lambda \cdot V$$

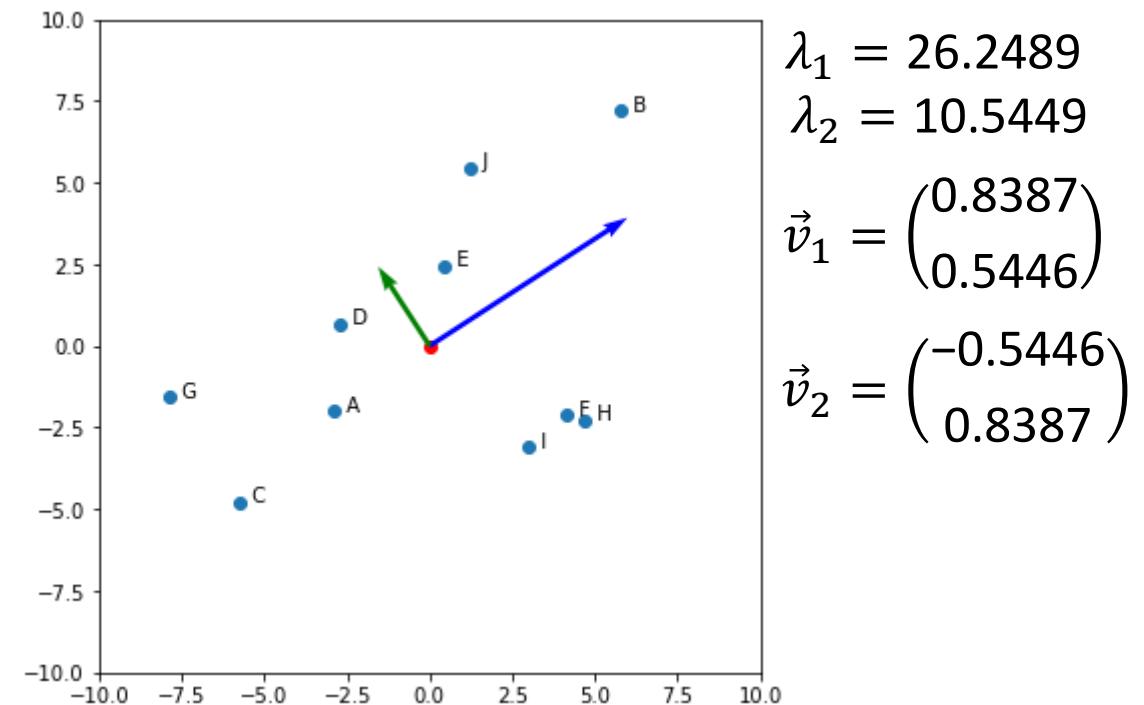
$$\Sigma \cdot V = \Lambda \cdot I \cdot V$$

$$\Sigma \cdot V - \Lambda \cdot I \cdot V = \mathbf{0}$$

$$(\Sigma - \Lambda \cdot I) \cdot V = \mathbf{0}$$

Como um autovetor v de V não pode ser um vetor nulo, a redução dimensional desejada somente é obtida com $\det(\Sigma - \Lambda \cdot I) = 0$, que é um polinômio cujas raízes são os autovalores de Σ .

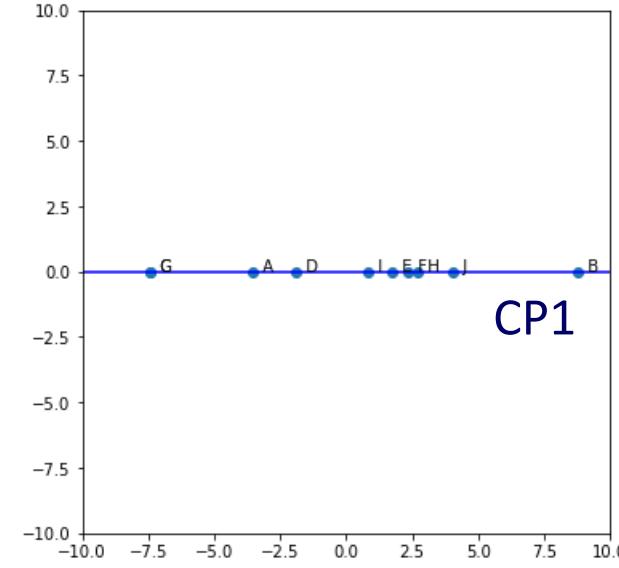
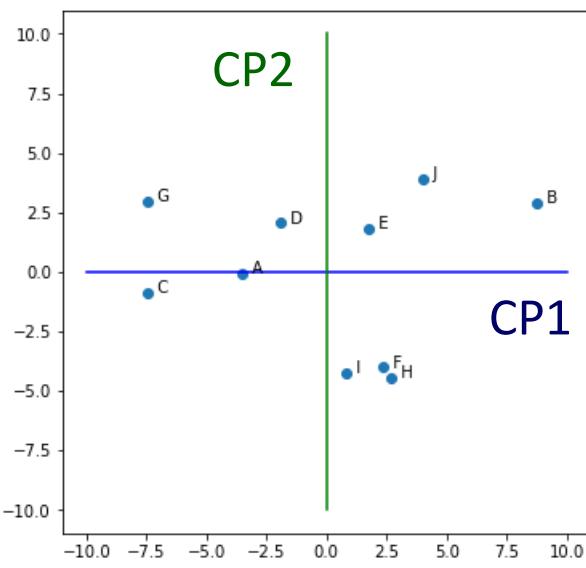
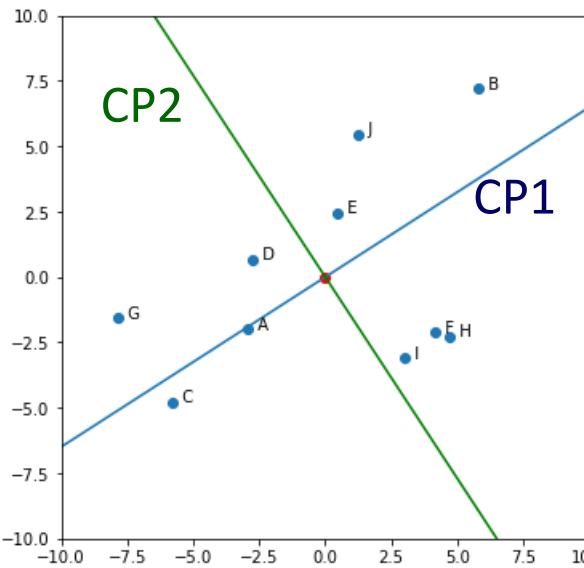
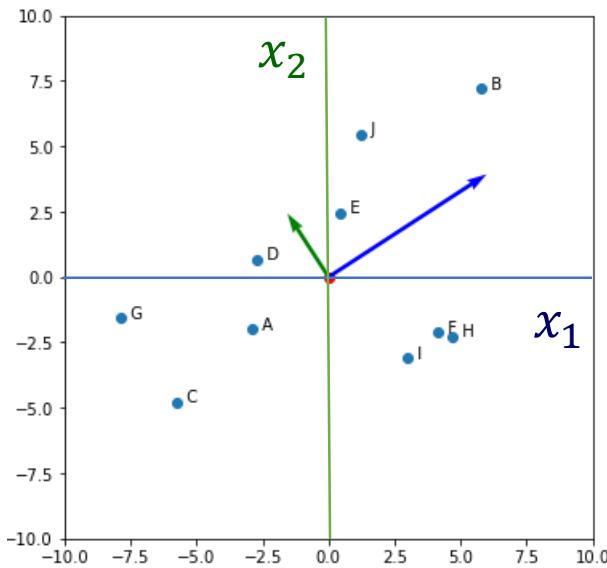
Com os autovalores calculados, descobrem-se os seus autovetores levando-os à equação $(\Sigma - \Lambda \cdot I) \cdot V = \mathbf{0}$.



Análise Exploratória de Dados : análise de componentes principais

Passos:

3. A partir das coordenadas dos autovetores, determinar as retas suporte que passam por eles (elas serão os novos eixos)



4. Para obter os valores dos registros no novo espaço (talvez reduzido), projetá-los nos componentes por meio dos autovetores.

$$v_j \cdot \text{Dados}_{\text{centralizados}} = \text{Dados}_{CP_j}$$

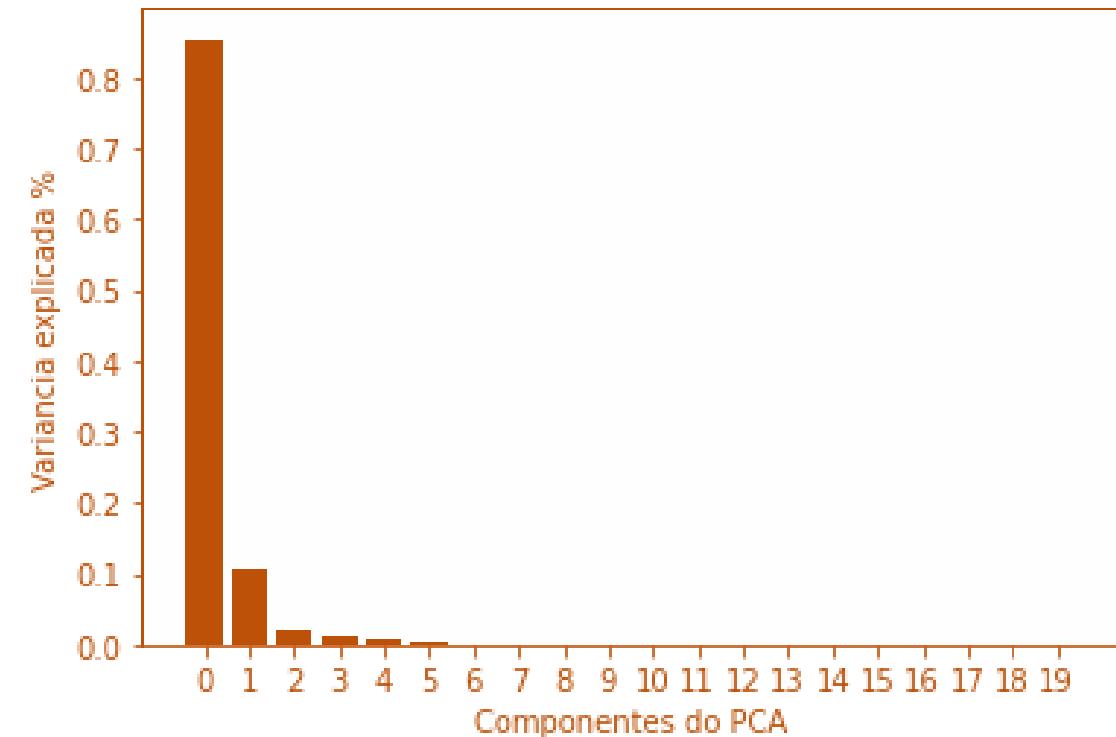
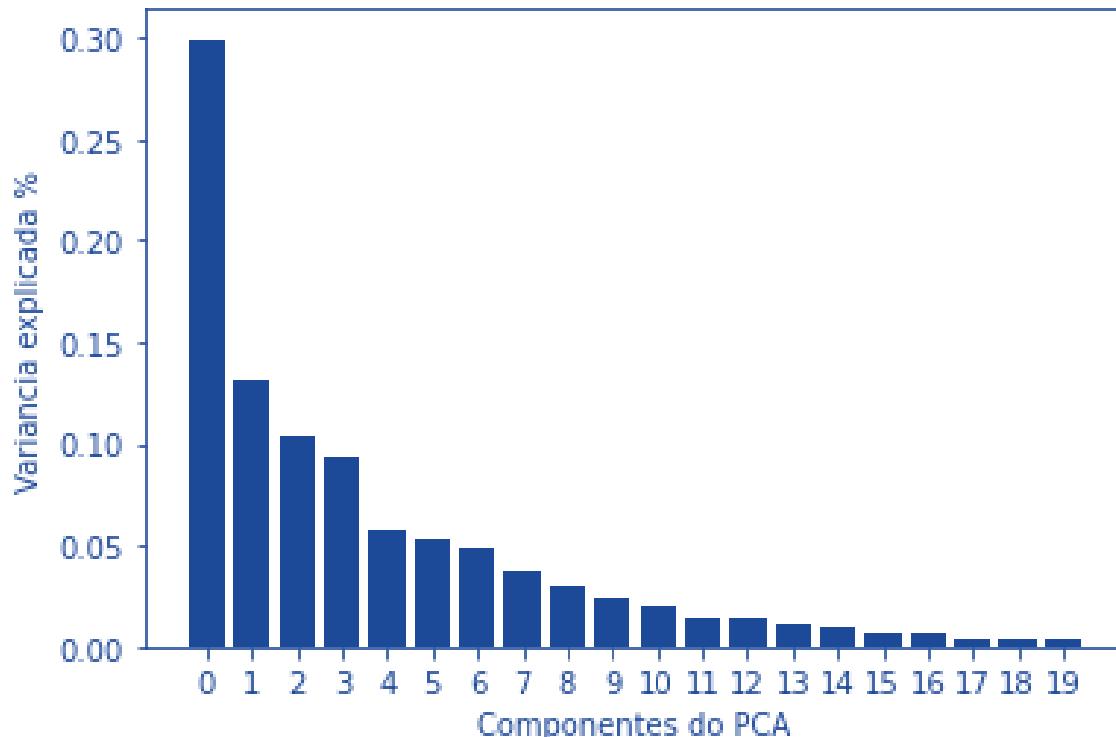
RPE
2025

Preparação de dados

Análise Exploratória de Dados : análise de componentes principais

Os autovalores são usados para estimar o quanto cada componente explica a variação dos dados.

$$\%_j = \frac{\lambda_j}{\sum \lambda_i}$$

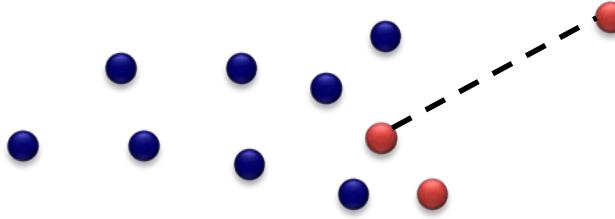


Desbalanceamento de classes: amostragem sintética

Uma técnica interessante para se reduzir o desbalanceamento entre classes é realizar a **subamostragem**, reduzindo a oferta dos registros das classes mais frequentes, ou a **superamostragem**, criando-se novas amostras das classes menos frequentes a partir dos dados existentes. Uma ação não inviabiliza a outra e é comum obter bons resultados adotando-se ambas.

Possivelmente a 1^a proposta que aliou ambas as estratégias foi a SMOTE e a implementação da superamostragem, sem reposição, é bastante simples:

Considerando os registros da classe que se deseja criar amostras sintéticas, escolher um elemento aleatoriamente e detectar alguns de seus vizinhos próximos. Desses, selecionar um aleatoriamente para a geração aleatória de um elemento sintético no caminho que une os elementos considerados.



[Journal of Artificial Intelligence Research 16 \(2002\) 321–357](#)

SMOTE: Synthetic Minority Over-sampling Technique

Nitesh V. Chawla

*Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA*

CHAWLA@CSEE.USF.EDU

Kevin W. Bowyer

*Department of Computer Science and Engineering
384 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA*

KWB@CSE.ND.EDU

Lawrence O. Hall

*Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA*

HALL@CSEE.USF.EDU

W. Philip Kegelmeyer

*Sandia National Laboratories
Biosystems Research Department, P.O. Box 969, MS 9951
Livermore, CA, 94551-0969, USA*

WPK@CALIFORNIA.SANDIA.GOV

imbalanced-learn