Introdução ao Aprendizado de Máquina para Físicos

Marcelo Vargas dos Santos
Aula 2

Entendendo um problema em ML

O ovo ou a galinha?

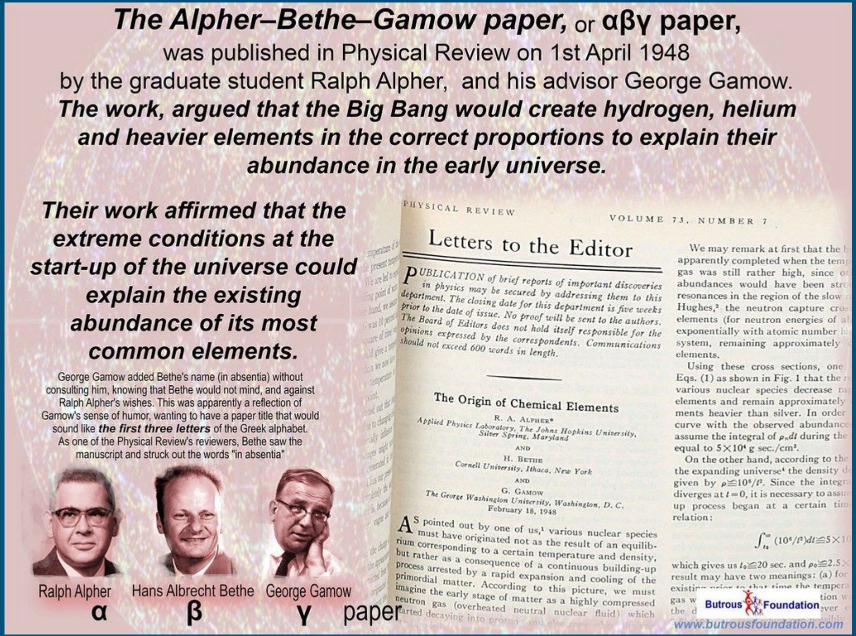
Qual caminho seguir?

1. Formular uma pergunta a ser respondida com os dados que temos?

2. Ou buscar dados para responder uma pergunta já formulada?

Tanto faz! Ambos os caminhos são válidos.

Radiação Cósmica de Fundo



The Alpher–Bethe–Gamow paper, or **αβγ paper,**
was published in Physical Review on 1st April 1948
by the graduate student Ralph Alpher, and his advisor George Gamow.
The work, argued that the Big Bang would create hydrogen, helium and heavier elements in the correct proportions to explain their abundance in the early universe.

Their work affirmed that the extreme conditions at the start-up of the universe could explain the existing abundance of its most common elements.

George Gamow added Bethe's name (in absentia) without consulting him, knowing that Bethe would not mind, and against Ralph Alpher's wishes. This was apparently a reflection of Gamow's sense of humor, wanting to have a paper title that would sound like ***the first three letters*** of the Greek alphabet. As one of the Physical Review's reviewers, Bethe saw the manuscript and struck out the words "in absentia"

PHYSICAL REVIEW VOLUME 73, NUMBER 7

Letters to the Editor

PUBLICATION of brief reports of important discoveries in physics may be secured by addressing them to this department. The closing date for this department is five weeks prior to the date of issue. No proof will be sent to the authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents. Communications should not exceed 600 words in length.

The Origin of Chemical Elements

R. A. ALPHER*
Applied Physics Laboratory, The Johns Hopkins University, Silver Spring, Maryland
AND
H. BETHE
Cornell University, Ithaca, New York
AND
G. GAMOW
The George Washington University, Washington, D. C.
February 18, 1948

AS pointed out by one of us,[1] various nuclear species must have originated not as the result of an equilibrium corresponding to a certain temperature and density, but rather as a consequence of a continuous building-up process arrested by a rapid expansion and cooling of the primordial matter. According to this picture, we must imagine the early stage of matter as a highly compressed neutron gas (overheated neutral nuclear fluid) which started decaying into proton

We may remark at first that the
apparently completed when the tem
gas was still rather high, since o
abundances would have been str
resonances in the region of the slow
Hughes,[2] the neutron capture cro
elements (for neutron energies of a
exponentially with atomic number
system, remaining approximately
elements.

Using these cross sections, one
Eqs. (1) as shown in Fig. 1 that the
various nuclear species decrease ra
elements and remain approximately
ments heavier than silver. In order
curve with the observed abundance
assume the integral of $\rho_n dt$ during the
equal to 5×10^4 g sec./cm^3.

On the other hand, according to the
the expanding universe[4] the density d
given by $\rho\cong10^6/t^2$. Since the integr
diverges at $t=0$, it is necessary to assu
up process began at a certain tim
relation:

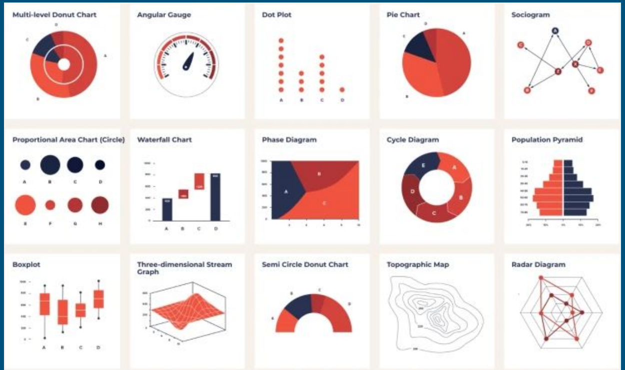
$$\int_{t_0}^{\infty}(10^6/t^2)dt\cong5\times10$$

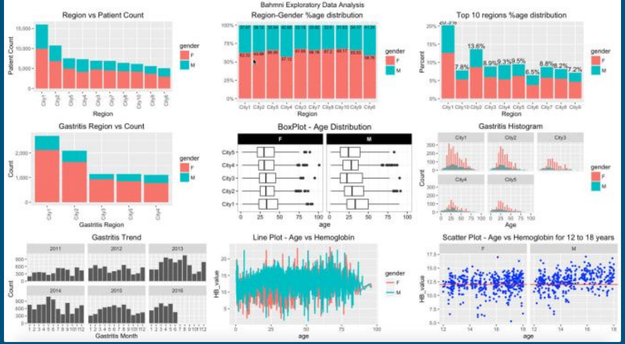
which gives us $t_0\cong20$ sec. and $\rho_0\cong2.5\times$
result may have two meanings: (a) for
existi
gas w tion w
the d ever

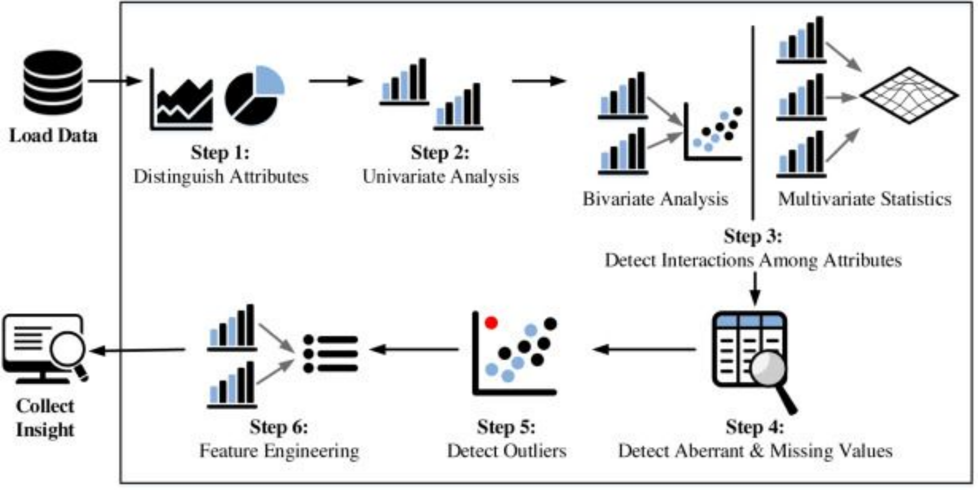
Onde encontrar dados

1. Google Dataset Dearch:datasetsearch.research.google.com
2. Kaggle: kaggle.com (Competições em ciência de dados)
3. Drivendata: drivendata.org (Competições)
4. Portal Brasileiro de Dados: dados.gov.br
5. 538: fivethirtyeight.com (Opinião Pública)
6. Quandl: quandl.com (Dados Financeiros)
7. Reddit: reddit.com/r/datasets

E depois? Análise Exploratória de dados







Load Data

Step 1: Distinguish Attributes

Step 2: Univariate Analysis

Bivariate Analysis

Multivariate Statistics

Step 3: Detect Interactions Among Attributes

Step 4: Detect Aberrant & Missing Values

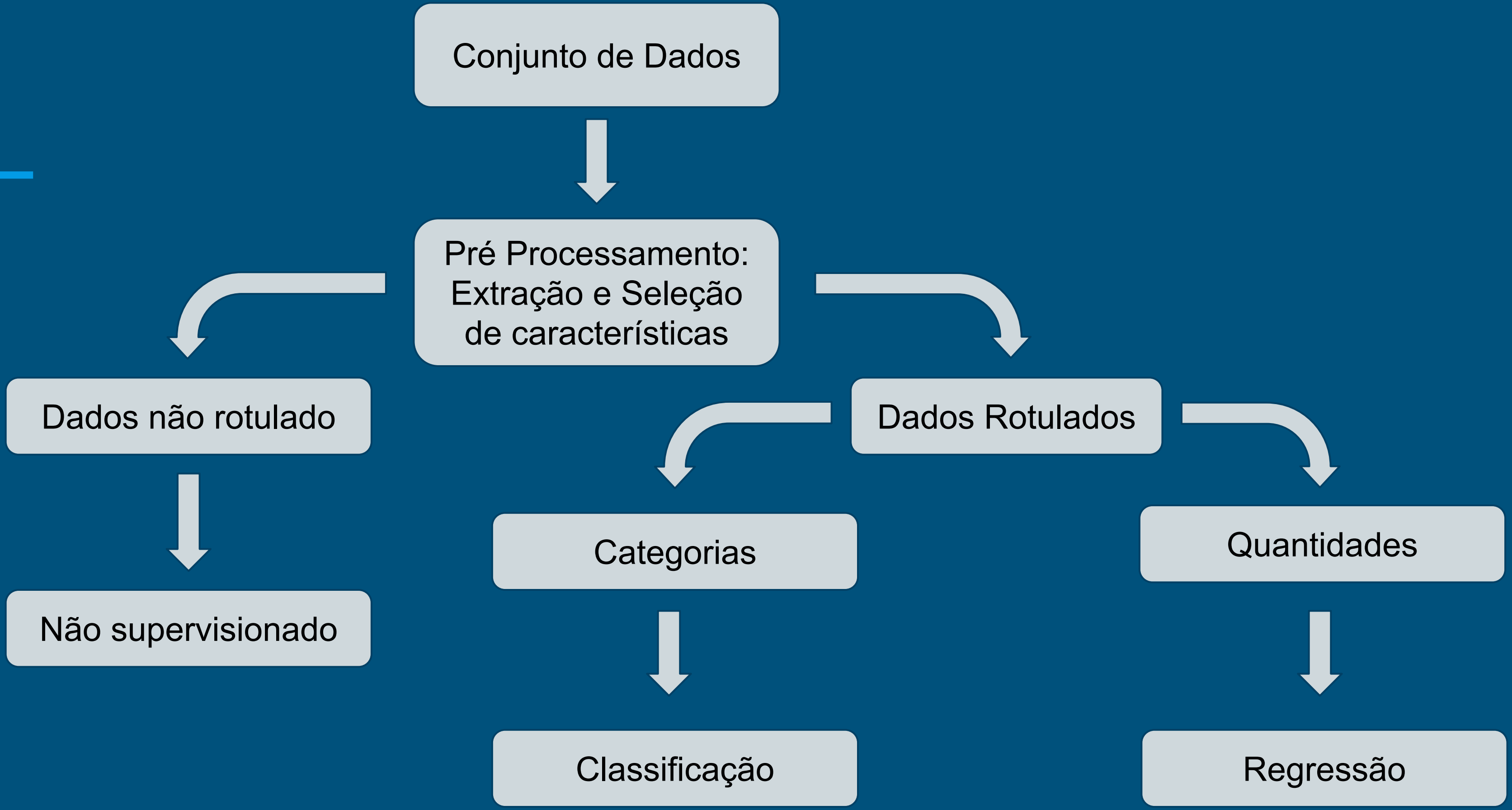
Step 5: Detect Outliers

Step 6: Feature Engineering

Collect Insight

A comprehensive review of tools for exploratory analysis of tabular industrial datasets
https://doi.org/10.1016/j.visinf.2018.12.004

Escolhendo o método



Conjunto de Dados

Pré Processamento: Extração e Seleção de características

Dados não rotulado

Não supervisionado

Dados Rotulados

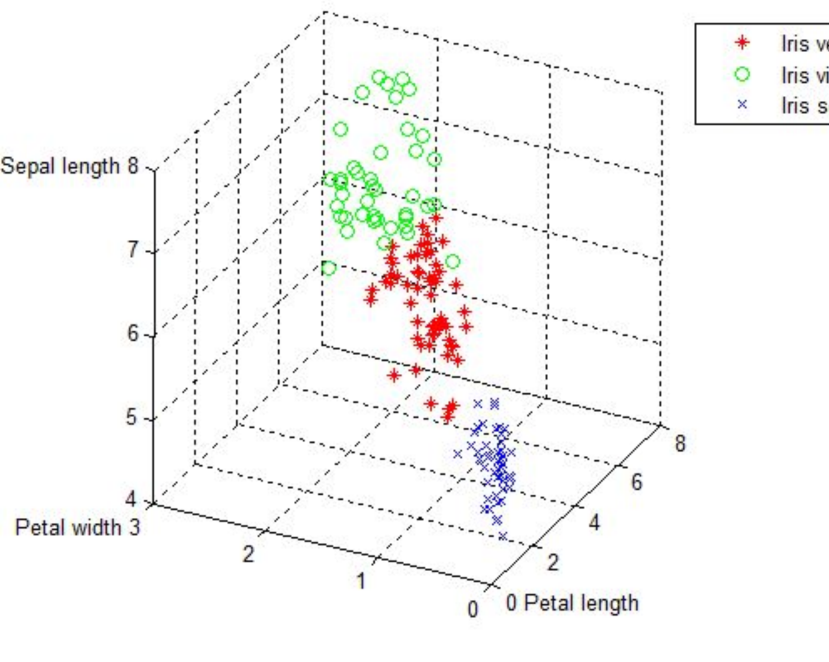
Categorias

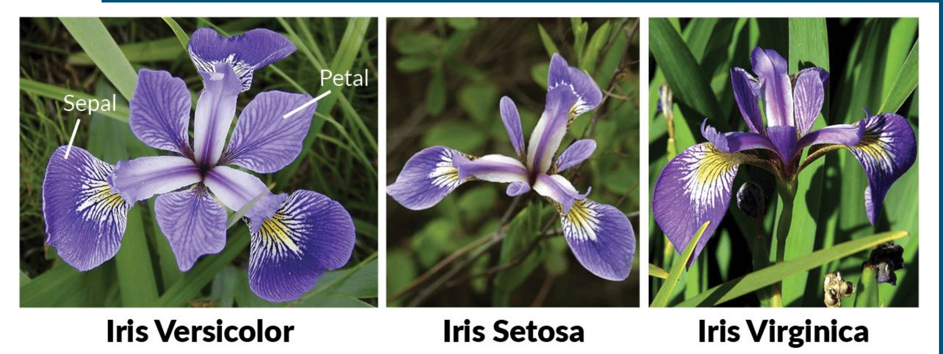
Classificação

Quantidades

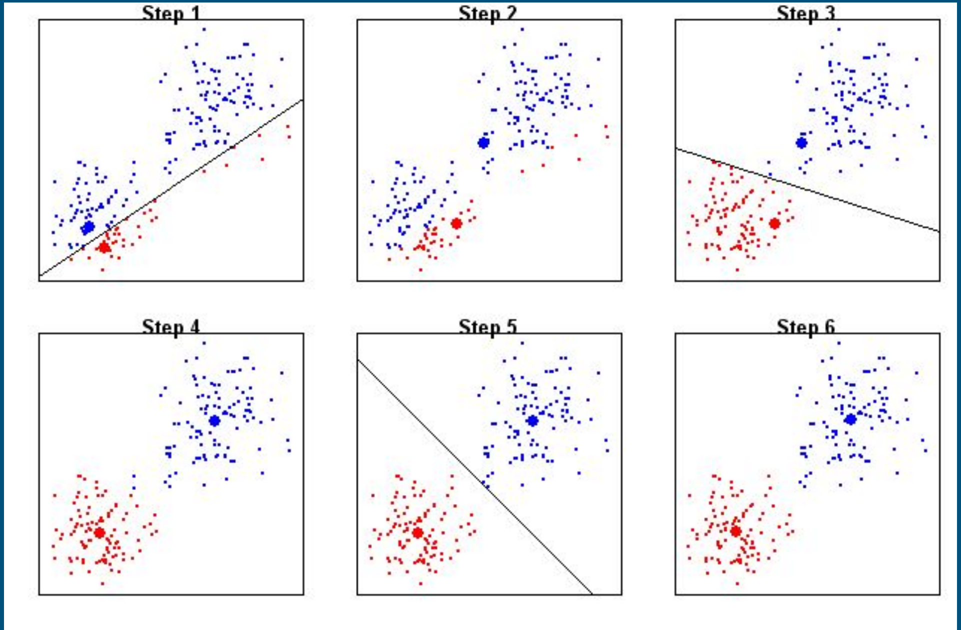
Regressão

Não supervisonado (Agrupamento)

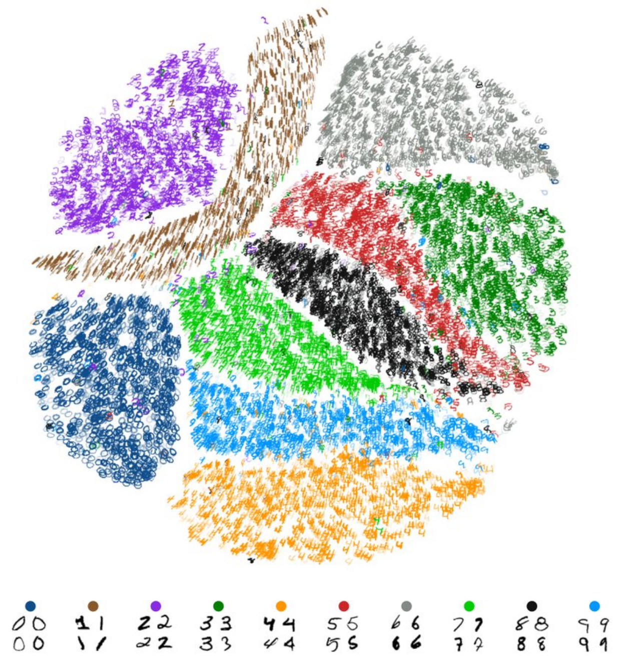




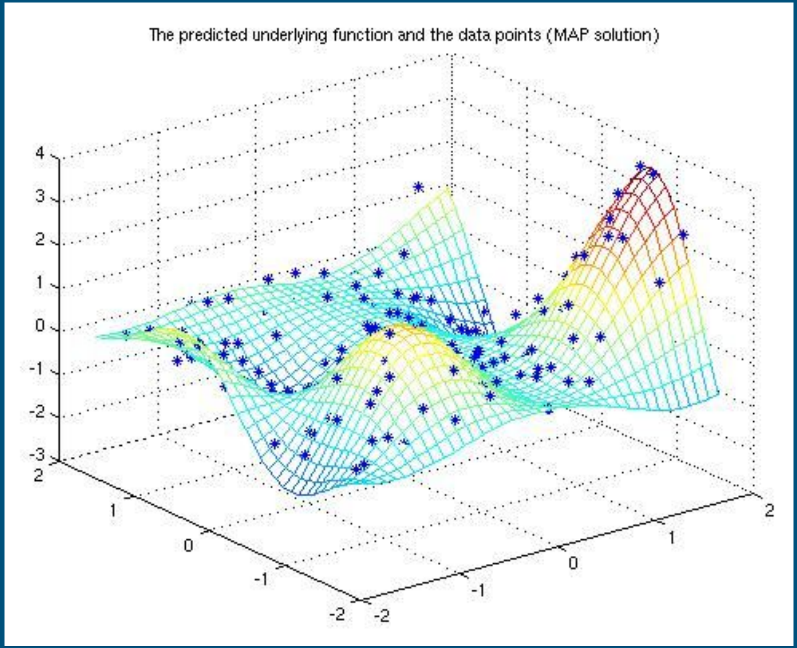
Modelo K-Means

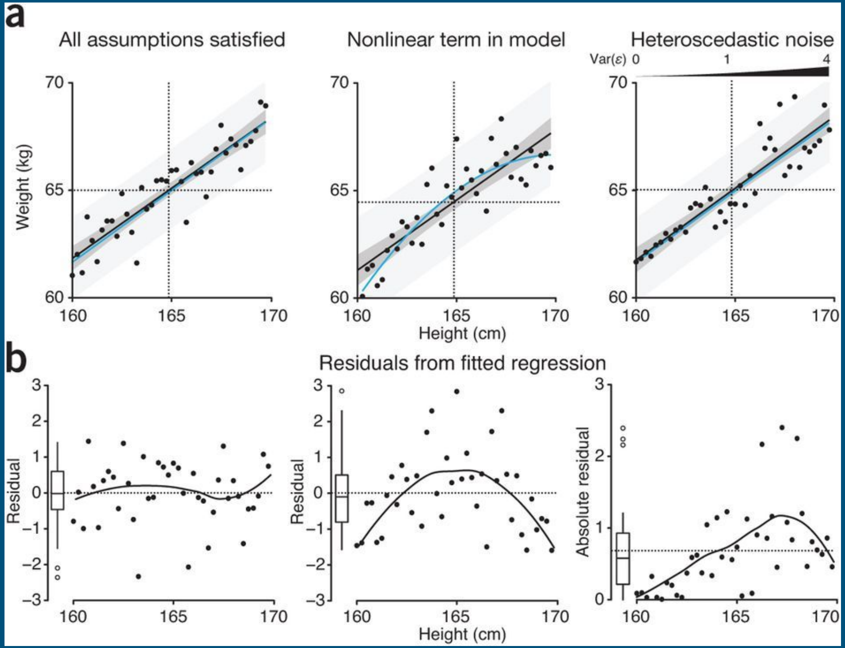


Classificação

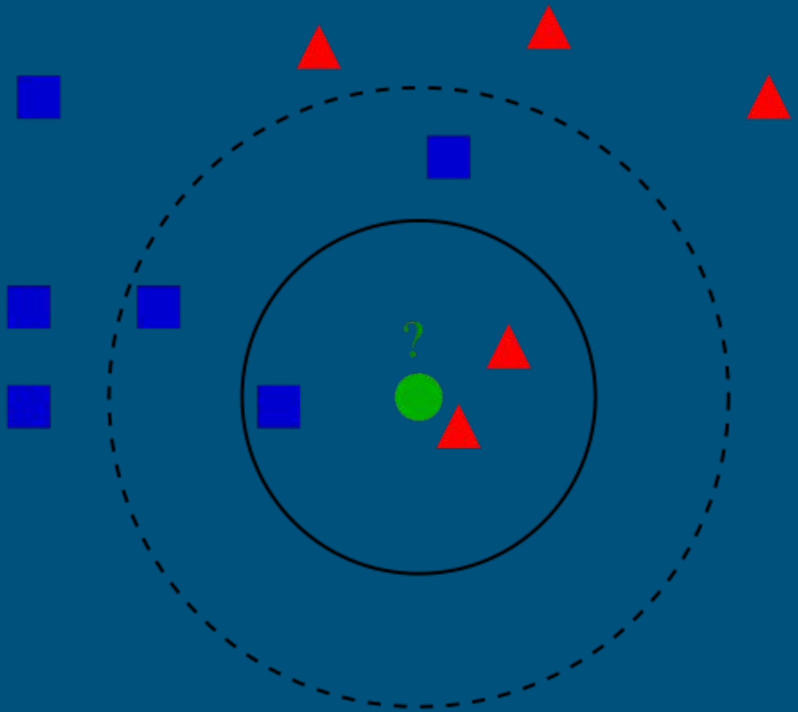


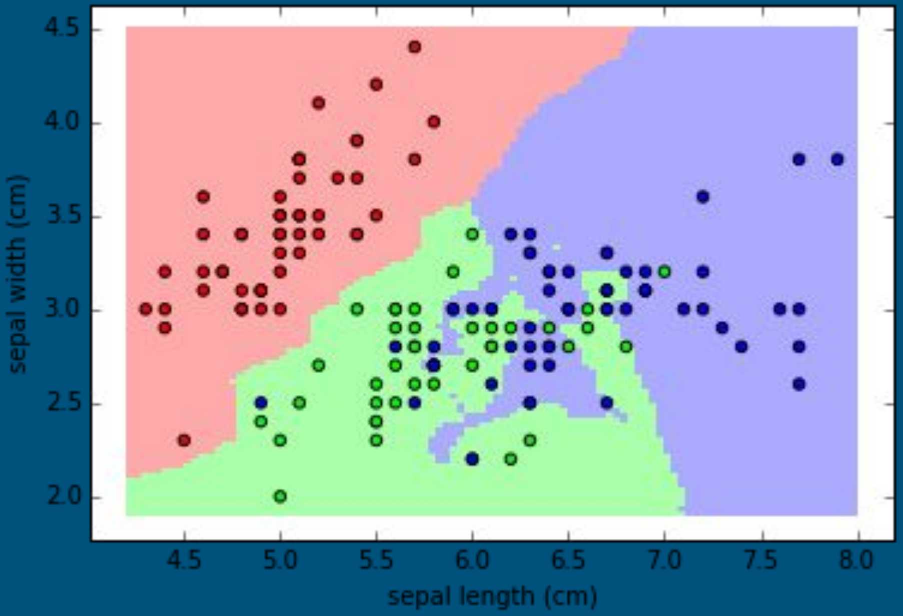
Regressão


The predicted underlying function and the data points (MAP solution)
4
3
2
1
0
-1
-2
-3
2
1
0
-1
-2
-2
-1
0
1
2

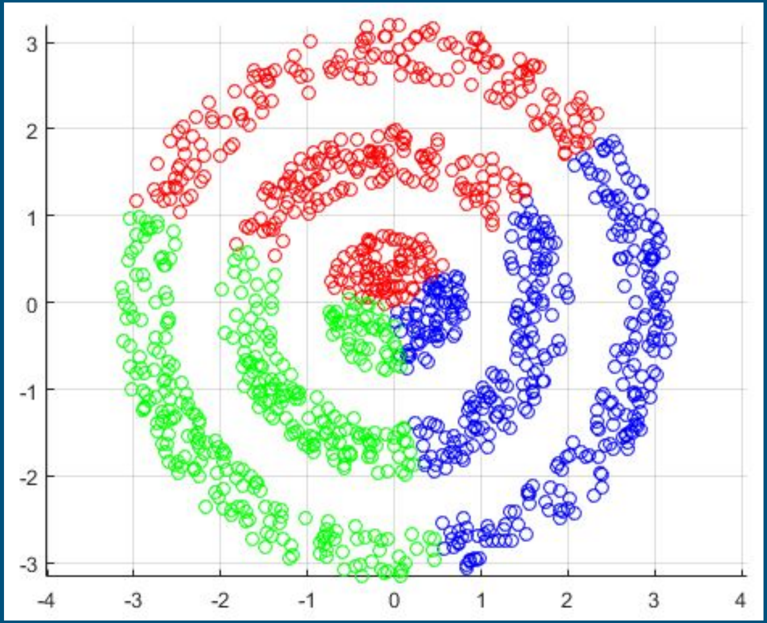

a
All assumptions satisfied
Nonlinear term in model
Heteroscedastic noise
Var(ε) 0 1 4
Weight (kg)
70
65
60
160
165
170
Height (cm)
b
Residuals from fitted regression
Residual
Absolute residual
3
2
1
0
−1
−2
−3
Height (cm)

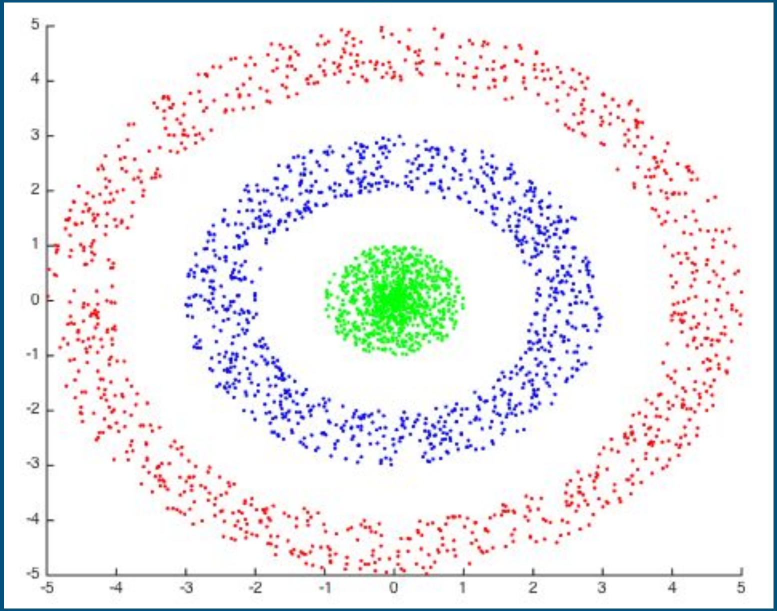
K-ésimo Vizinho mais Próximo
k-nearest neighbors





Pré Processamento



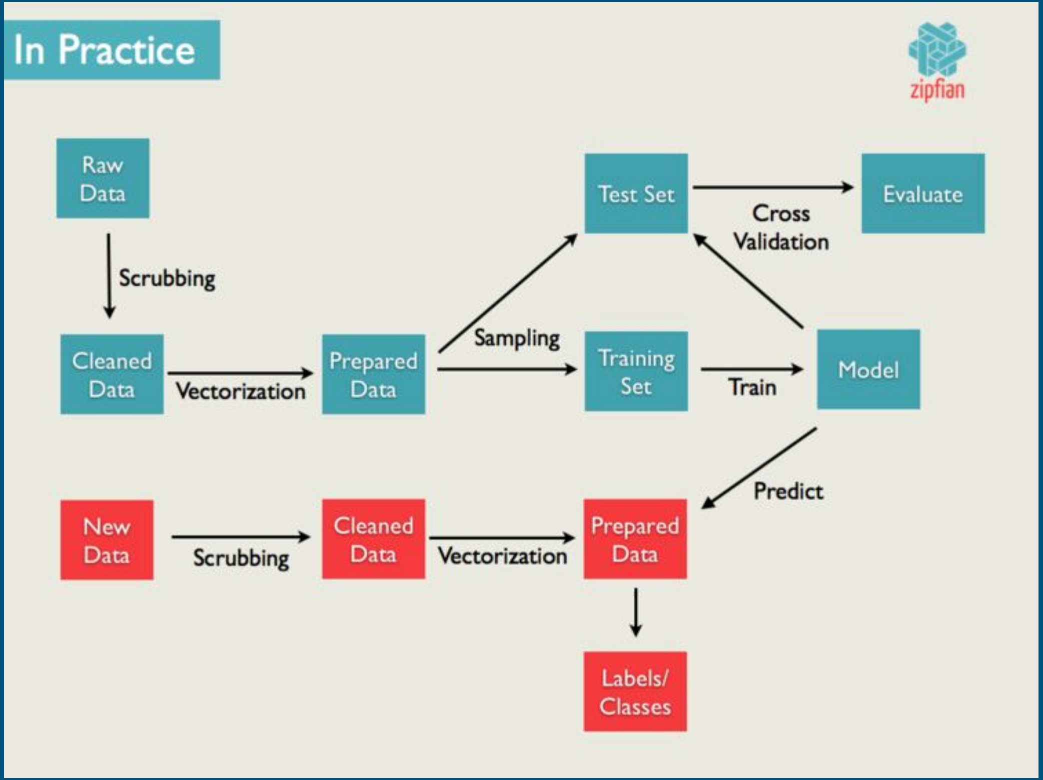


Processo de Aprendizado

Ingredientes

$\mathbf{X}$	:	Amostra
$\mathcal{G}(\cdot,\cdot)$	:	Algoritmo
$g_{\mathbf{w},\mathbf{X}}(\cdot) = \mathcal{G}(\mathbf{w},\mathbf{X})$	:	Modelo
$\mathbf{w}$	:	Hiperparâmetros
$\mathbf{Y} = g_{\mathbf{w},\mathbf{X}}(\bar{\mathbf{X}})$	:	Previsão
$\mathcal{C}(\bar{\mathbf{X}},\mathbf{Y})$	:	Custo

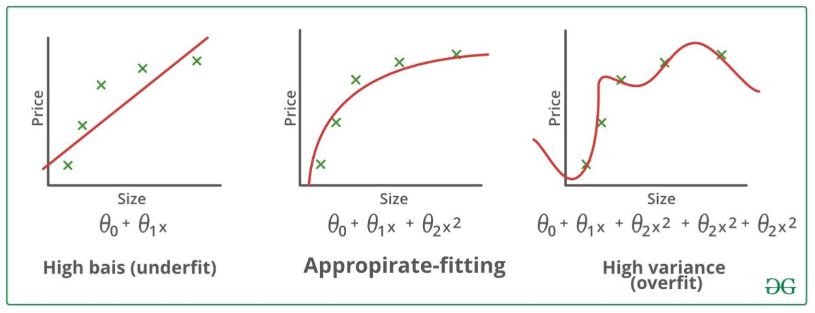
Processo iterativo

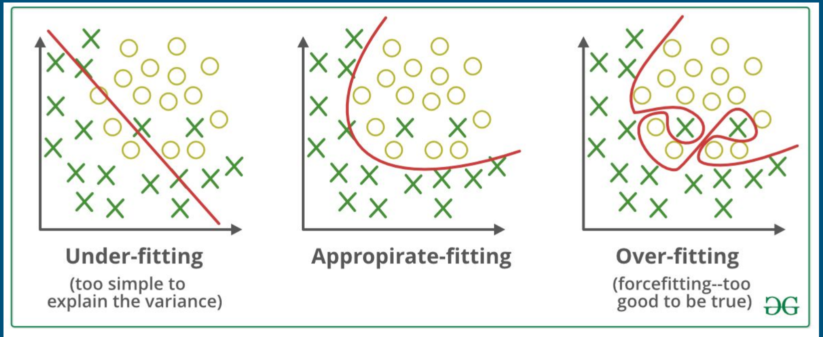


Pseudo-Código

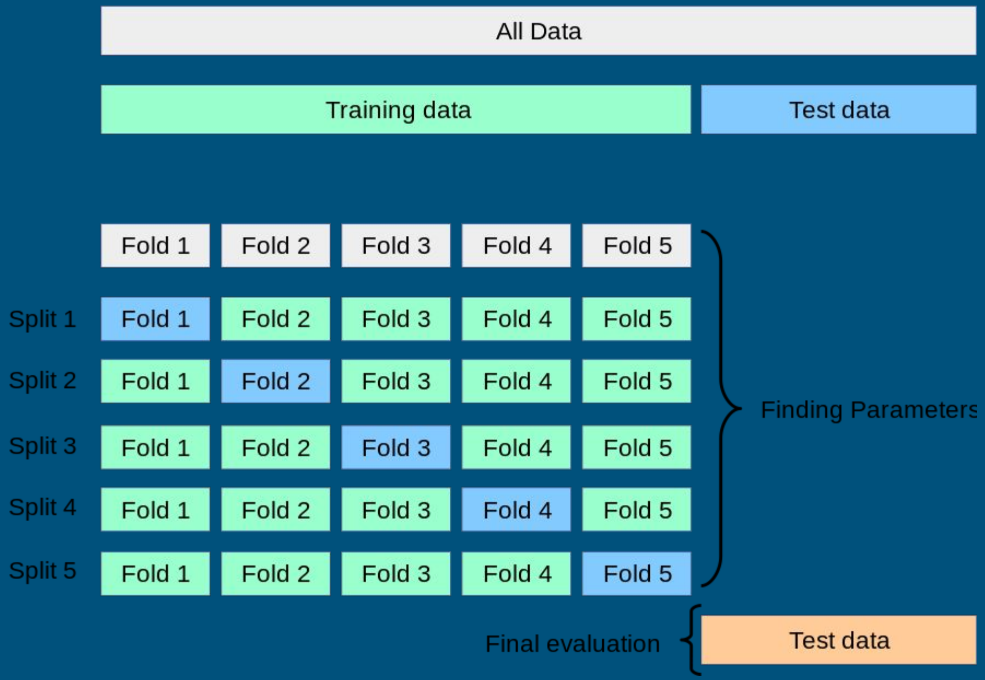
```
$\mathbf{X} = \mathbf{X}_{Test} + \mathbf{X}_{Train}$

Para $\mathbf{w}$ em $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ :
    $g = \mathcal{G}(\mathbf{w}, \mathbf{X}_{Train})$
    $c_i = \mathcal{C}(\mathbf{X}_{Test}, g(\mathbf{X}_{Test})$
```

Super-ajuste vs Sub-ajuste
Overfitting vs Underfitting




Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Validação cruzada



All Data

Training data | Test data

Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 1 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 2 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 3 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 4 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 5 | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Finding Parameters

Final evaluation | Test data

Pseudo-Código

```
$\mathbf{X} = \mathbf{X}_{Test} + \mathbf{X}_{Train}$

Para $\mathbf{w}$ em $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ :
    $g = \mathrm{CV}(\mathcal{G}(\mathbf{w}, \cdot), \mathbf{X}_{Train})$
    $c_i = \mathcal{C}(\mathbf{X}_{Test}, g(\mathbf{X}_{Test})$
```

Análise Exploratória de Dados

Próxima Aula