ANÁLISE DE DADOS APLICADA À COMPUTAÇÃO

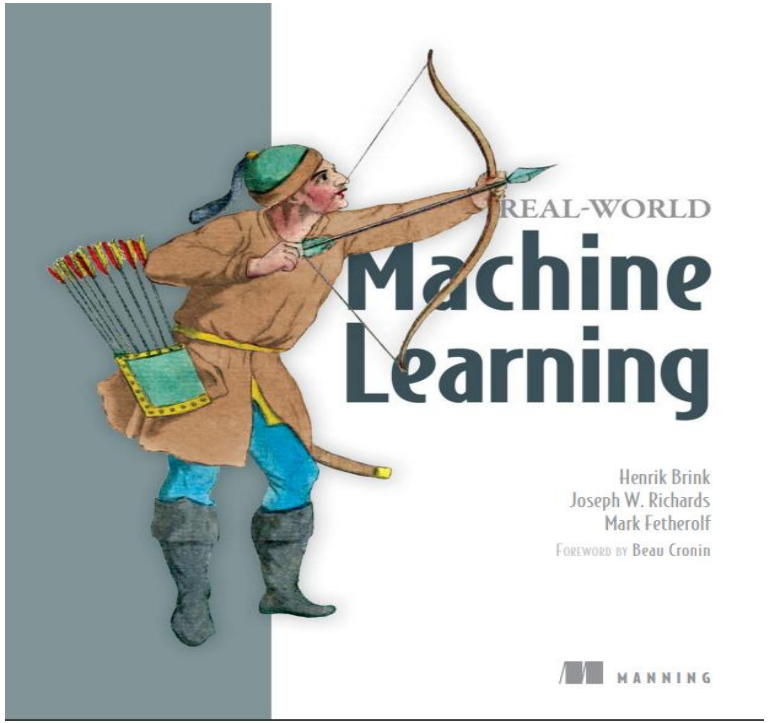
Feature Engineering

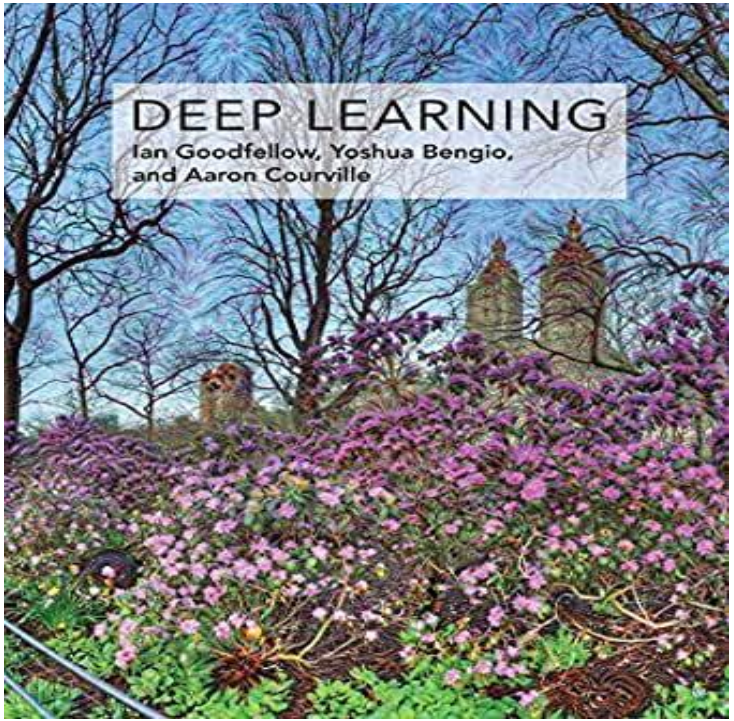
Prof. M.Sc. Howard Roatti

Sumário

1. **Introdução ao *Feature Engineering***
2. **Pré-Processamento**
3. **Pré-Processamento – Atributos Categóricos**
4. **Pré-Processamento – Atributos Numéricos**
5. **Pré-Processamento – Atributos Ausentes**

Bibliografia



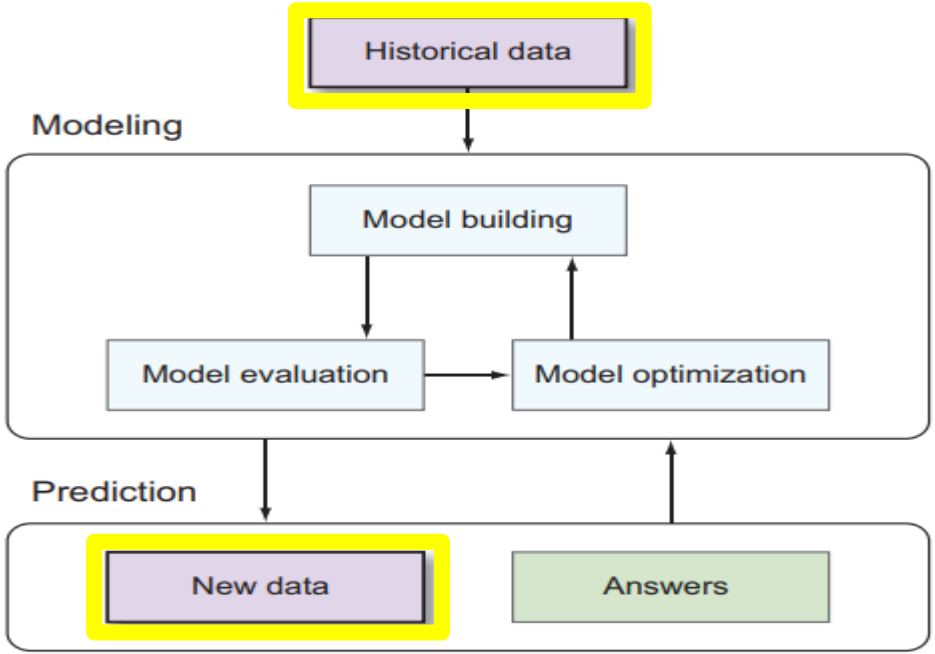


Feature Engineering

- No *Machine Learning* (ML), usamos os dados para ensinar os sistemas como realizar decisões melhores.
- Algoritmos de ML são projetados para descobrir padrões e associações em dados históricos gerando um modelo para prever um atributo de importância para novos dados.
- Com dados de alta qualidade, sistemas preditivos de alta fidelidade podem ser construídos.
- Porém, se os dados de treinamento forem de baixa qualidade, os esforços dos melhores algoritmos de ML podem ser inúteis.
- Boa parte da arte do ML está na exploração dos dados para avaliar sua qualidade orientando o processo de aprendizado.

Feature Engineering

Nesse momento, nossa preocupação está nos dados. Destacamos então onde eles estarão participando do Workflow do Machine Learning.



Feature Engineering

- **Pré-Processando para modelagem**
- De acordo com a composição do *dataset*, é necessário executar alguns passos de pré-processamento.
- Muitos algoritmos de ML funcionam apenas sobre dados numéricos: inteiros (***int***) e reais (***float***).
- Muitos *datasets* incluem outros tipos de atributos, como textos (representados por TF ou TF-IDF), variáveis categóricas e valores faltantes.
- Alguns atributos numéricos podem ser reescalados para torná-los comparáveis ou para alinhá-los com uma distribuição de frequência.

Feature Engineering

- **Pré-Processando para modelagem - Atributos Categóricos**
- Atributos categóricos são aqueles que podem ser colocados em um conjunto finito e a ordem não é importante.
- **Podem ser óbvios:** solteiro ou casado, fraco ou forte, sim ou não.
- **Ou não tão óbvios assim:**
  - Classes: 0, 1, 2 ou 3

Person	Name	Age	Income	Marital status
1	Jane Doe	24	81,200	Single
2	John Smith	41	121,000	Married

Categorical features

PassengerId	Survived	Pclass	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Male	22	1	0	A/5 21171	7.25		S
2	1	1	Female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Female	26	0	0	STON/02. 3101282	7.925		S
4	1	1	Female	35	1	0	113803	53.1	C123	S
5	0	3	Male	35	0	0	373450	8.05		S
6	0	3	Male		0	0	330877	8.4583		Q

Feature Engineering

- **Pré-Processando para modelagem - Atributos Categóricos**
- Para codificar os atributos categóricos existem algumas possibilidades, porém três são mais utilizadas:
- (1) LabelEncoder:

Breakfast
Every day
Never
Rarely
Most days
Never

Breakfast
3
0
1
2
0

- (2) OneHotEncoder:

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

- (3) Remover as colunas

Feature Engineering

- **Pré-Processando para modelagem - Atributos Categóricos**
- **Praticando:** baixe o arquivo **datasets01082020.zip**, utilizando o dataset do Titanic, aplique os métodos a seguir para os atributos categóricos existentes. Valide os resultados com a aplicação de um modelo de Machine Learning.
- (1) LabelEncoder
- (2) OneHotEncoder
- (3) Remover as colunas

Feature Engineering

- **Pré-Processando para modelagem - Atributos Numéricos**
- Uma prática que ajuda a evitar o enviesamento (bias) dos algoritmos de ML para números de maiores grandezas é a utilização de transformação de dados.
- Há duas técnicas importantes que têm o mesmo objetivo: transformar os atributos na mesma ordem de grandeza:
- (1) Padronização: remove a média e dimensiona os dados para a variação da unidade. Gera números dentro de uma faixa de valores positivos e negativos. $$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$
- (2) Normalização: redimensiona o vetor de cada amostra para ter norma unitária, independentemente da distribuição das amostras. Gera números entre 0 e 1. $$z = \frac{x_i - \mu}{\sigma}$$

Feature Engineering

- **Pré-Processando para modelagem - Atributos Numéricos**
- **Praticando:** utilizando o dataset **house-prices**, aplique os métodos a seguir para os atributos numéricos existentes. Valide os resultados com a aplicação de um modelo de Machine Learning.
- (1) Padronização

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- (2) Normalização

$$z = \frac{x_i - \mu}{\sigma}$$

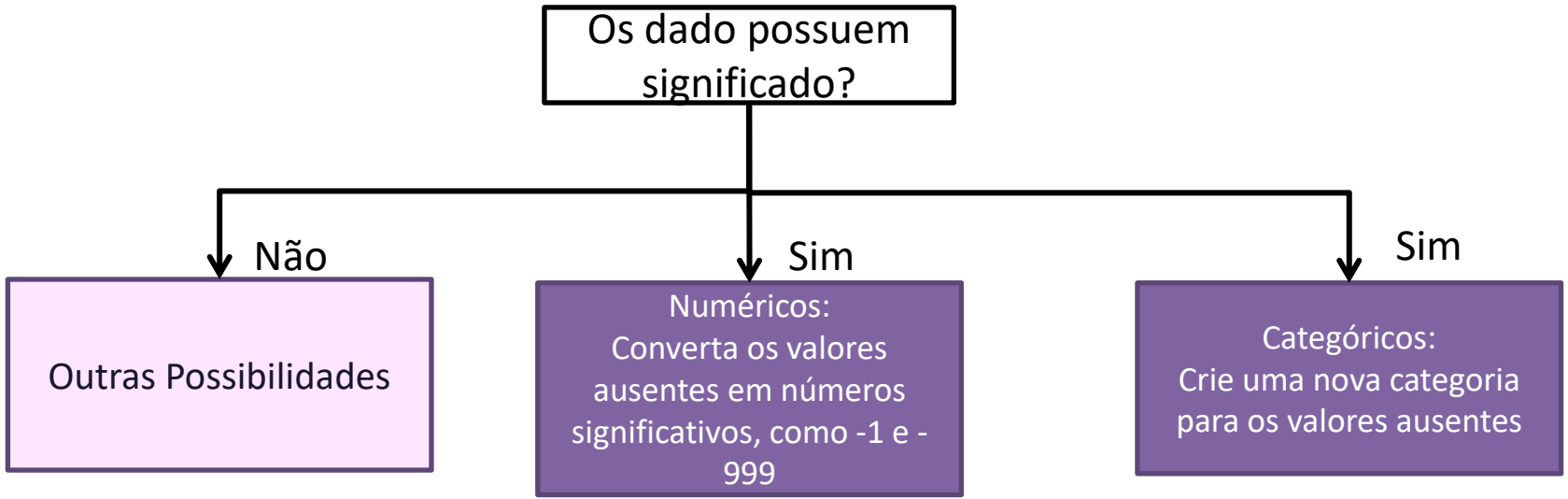
Feature Engineering

- **Pré-Processando para modelagem - Atributos Ausentes**
- Os dados ausentes geralmente aparecem como células vazias ou células com **NaN** (Não é um número), **N/A**, **NaT** ou **None**.
- Para atributos que faltam dentro de uma coleção, algumas abordagens podem ser aplicadas para tornar o desempenho do algoritmo melhor.

PassengerId	Survived	Pclass	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Male	22	1	0	A/5 21171	7.25		S
2	1	1	Female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Female	26	0	0	STON/02. 3101282	7.925		S
4	1	1	Female	35	1	0	113803	53.1	C123	S
5	0	3	Male	35	0	0	373450	8.05		S
6	0	3	Male		0	0	330877	8.4583		Q

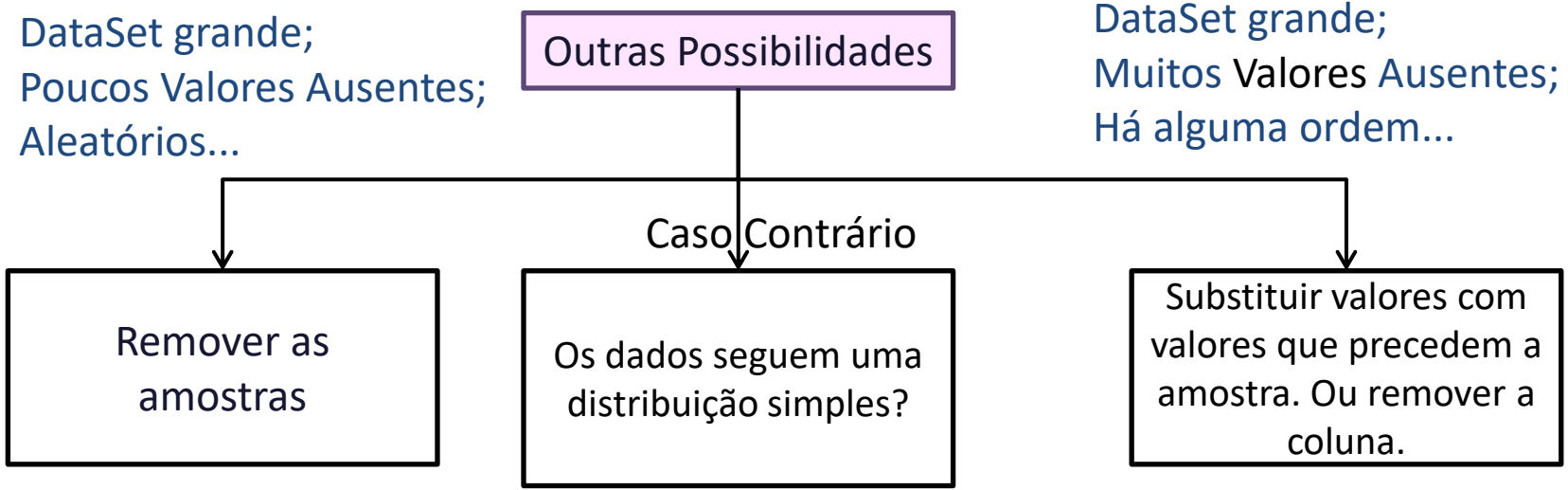
Feature Engineering

- **Pré-Processando para modelagem - Atributos Ausentes**



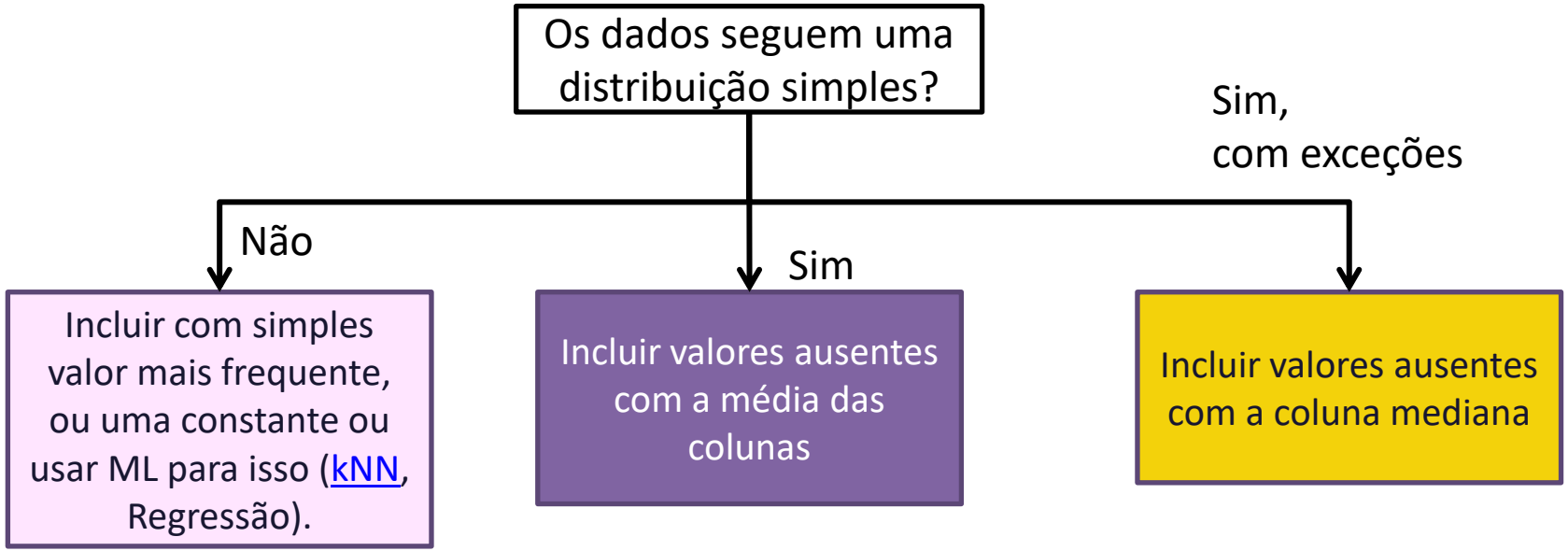
Feature Engineering

- **Pré-Processando para modelagem - Atributos Ausentes**



Feature Engineering

- **Pré-Processando para modelagem - Atributos Ausentes**



Feature Engineering

- **Pré-Processando para modelagem - Atributos Ausentes**
- **Praticando:** utilizando o dataset Titanic, aplique os métodos para tratamento de atributos ausentes nos dados existentes. Valide os resultados com a aplicação de um modelo de Machine Learning.
- **Dica:** a biblioteca sklearn possui um biblioteca para esses tipos de tratamento, chamada sklearn.impute

Feature Engineering

REFERÊNCIA

- brink et al, Real-World Machine Learning

- Kaggle Intermediate Machine Learning Course

- VAZ, A. L. NORMALIZAR OU PADRONIZAR VARIÁVES. DISP. LINK ACESSADO EM: 28/07/2020