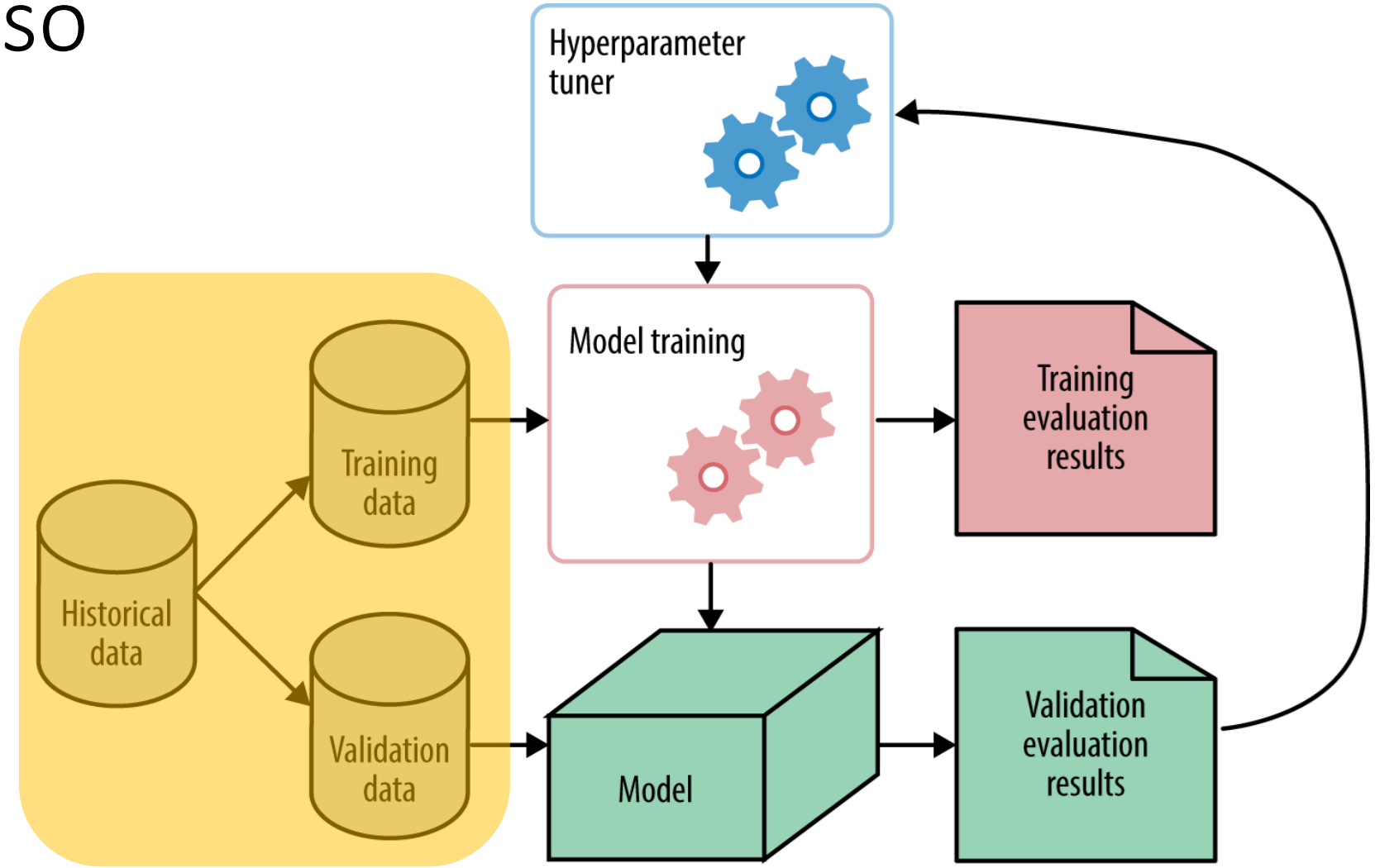
Aprendizado de Máquina
e Deep Learning

Preparação da base de dados

Prof. Dr. Thiago Meirelles Ventura

Processo



Deep Learning Bible - A. End-to-End Pipelines

Preparação dos dados

- Carregamento
- Seleção
- Limpeza
- Pré-processamento
- Separação

Carregamento

- Para o modelo ser construído ele deve ter acesso aos dados
- Essa etapa pode ser feita por lotes dependendo do tamanho dos dados
  - Operação por lotes é comum em dados maiores, como imagens
  - Caso contrário, todos os dados podem ser carregados em memória

Seleção

- Não são todos os atributos que devem ser utilizados na construção do modelo
- O tipo de dado pode ser um limitante para sua utilização
- Alguns atributos podem prejudicar o aprendizado do modelo
- Alguns atributos podem dar a falsa impressão que o modelo aprendeu
- Alguns atributos podem enviesar o modelo

Quais desses atributos podem enviesar um modelo de predição de crime?

slido.com

código: 3424889

Limpeza

- Remoção de colunas desnecessárias
- Detecção e tratamento de outliers
- Preenchimento de falhas
- Remoção de registros desnecessários

Pré-processamento

- Processamentos necessários dependendo do problema
- Transformação de dados
- Data augmentation
- Normalização

Pré-processamento - transformação

- Às vezes os dados de um atributo não está preparado para ser processado na construção de um modelo
- Dados textuais são um exemplo
- Dependendo do dado, é necessário uma transformação para que o mesmo possa ser útil para o aprendizado do modelo

Pré-processamento - transformação

- Exemplo com dado de dia da semana

Dia
Sábado
Terça
Segunda
Sexta

Dia
6
2
1
5

Domingo	Segunda	Terça	Quarta	Quinta	Sexta	Sábado
0	0	0	0	0	0	1
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	1	0

Pré-processamento - data augmentation

- Técnica para aumentar artificialmente os dados de uma base
- Adiciona variabilidade
- Aumenta o aprendizado do modelo
- Pode auxiliar para balancear classes

Pré-processamento - data augmentation

- Novos dados são criados a partir de dados originais
- Cada tipo de dados pode ter técnicas diferentes para realizar a criação
  - Imagens: rotações, zoom, coloração
  - Sons: volume, adição de ruídos, mudança de velocidade
  - Texto: troca de palavras, remoção de palavras
- Inclusive outros métodos de IA podem servir para gerar novos dados

Exercício 1

- Você foi contratado por uma pizzaria para criar um modelo que consiga estimar quantas pizzas serão vendidas no dia seguinte
- A base de dados da pizzaria possui apenas o atributo de data (dd/mm/aaaa) e quantas pizzas foram vendidas no respectivo dia (inteiro)
- Aplique data augmentation para criar novos atributos e, assim, aumentar a precisão do modelo que será criado

Pré-processamento - normalização

- Tem o objetivo de deixar todos os dados em um mesmo intervalo
- Atributos com escalas diferentes prejudicam o aprendizado do modelo
  - Um atributo pode parecer ter mais importância que outro
  - A calibração dos pesos pode ser mais demorada

Pré-processamento - normalização

- Analise:

$$y = x_1 * w_1 + x_2 * w_2$$

Pré-processamento - normalização

- Uma forma de normalizar:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Exercício 2

- Normalize os dados abaixo

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Salário
10.000,00
5.000,00
2.500,00
14.000,00
8.250,00

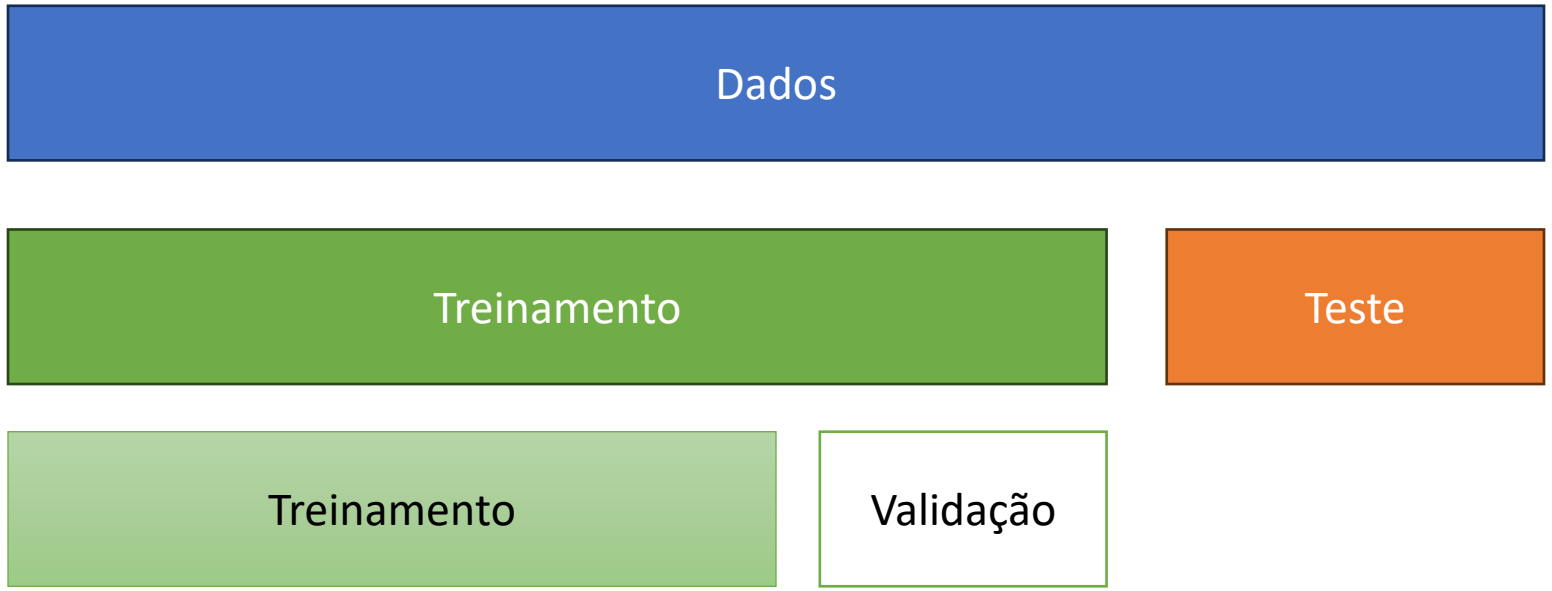
Separação

Como saber o quanto o modelo aprendeu?

Separação

- É necessário avaliar o modelo durante e após o seu treinamento
- Os dados não podem ser os mesmos
- Deve ser uma avaliação justa

Separação



Separação

- Divisão comuns:
  - 70% / 30%
  - 80% / 20%
  - Ano de previsão
- Pode depender do domínio dos dados

A era da fé cega no Big data tem de acabar

TED 2017, Cathy O'Neil

https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end