Inteligência Artificial

Preparação e Pré-processamento dos dados
Parte III

Prof. Dr. Ivan Carlos Alcântara de Oliveira

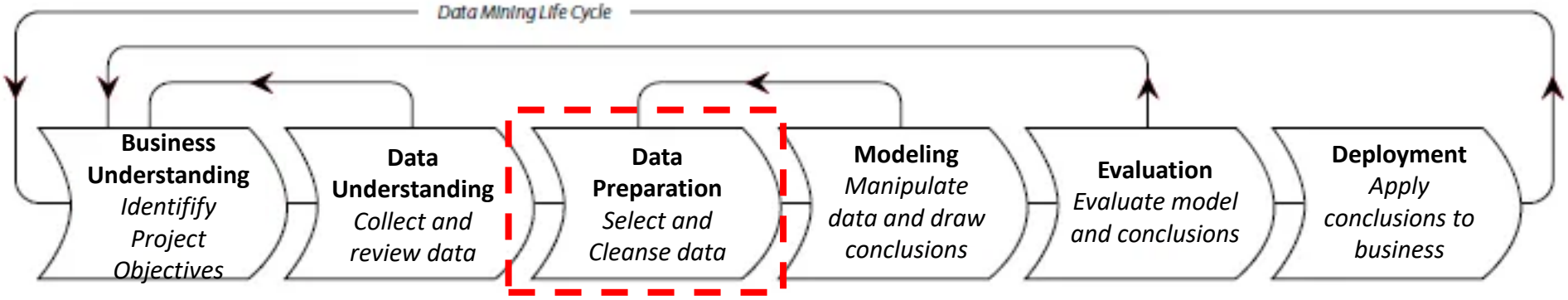
https://orcid.org/0000-0002-6020-7535

Preparação dos dados e Pré-Processamento

Ciclo de Vida de Projeto de Data Science

CRISP-DM

- **CRISP-DM (*Cross Industry Standard Process for Data Mining* -** Processo Padrão Interindústrias para Mineração de Dados) é um processo de fases bastante aceito na indústria para representar um ciclo completo de Ciência e Análise de Dados, incluído a aplicação de modelos de Aprendizado de Máquina.



Auxiliar na Identificação	Obtenção dos dados	Pré-processamento	Seleção do modelo	Validação do modelo	Aplica solução obtida ao negócio
Compreensão	Integração	Transformação de Variáveis	Cross-Validation	Otimização	Desenvolvimento de Aplicação Analítica
Pesquisas e Estudos	Análise Exploratória	Seleção de Atributos	Métricas de performance		Geração de relatório
Reunião com Especialistas		Redução de Dimensionalidade	Otimização		Construção de Dashborads etc.
Definição dos Atributos		Amostragem			
Definição medida de erro					

Ciência de Dados

Ciclo de Vida de Projeto de *Data Science*

3. **Preparação dos Dados**

Data Wrangling: pré-processamento, transformação, limpeza dos dados, etc.

Até aqui, aproximadamente 70% do tempo de um Cientista de Dados.

Preparação dos Dados

- O **desempenho** dos algoritmos de AM geralmente são **afetado**s pelo **estado dos dados**.
  - Conjunto de dados **podem apresentar diferentes características**, **dimensões** ou **formatos**.
  - Por exemplo, podem ser numéricos ou categóricos/simbólicos.

Preparação dos Dados

- **Conjunto de dados podem também**:
  - Estar limpos ou **conter ruídos** e imperfeições
  - Com valores incorretos
  - **Inconsistentes**
  - **Duplicados ou ausentes**
  - Ter atributos independentes ou relacionados
  - **Apresentar poucos ou muitos objetos**
  - Ter um **número pequeno ou elevado de atributos**.

Preparação dos Dados

- **Técnicas de Preparação dos** dados são frequentemente **utilizadas** para **melhorar** a **qualidade dos dados** por meio da eliminação ou minimização desses problemas.

Preparação dos Dados

- Essa melhora pode:
  - **facilitar o uso de técnicas de AM**
  - **permitir a construção de modelos mais fiéis à distribuição real dos dados**
  - **Reduzir a complexidade computacional**
  - **Tornar mais fáceis e rápidos o ajuste de parâmetros** do modelo e seu posterior uso.
  - **Facilitar a interpretação dos padrões extraídos pelo modelo**

Preparação dos Dados

- Técnicas de Preparação dos dados também **podem tornar os dados mais adequados para** sua utilização por um determinado **algoritmo de AM**.
- **Por exemplo**, **alguns** algoritmos de AM **trabalham apenas com valores numéricos**.
- **Não existe** uma **ordem** fixa **para aplicação** das diferentes **técnicas** de **pré-processamento**.

Preparação dos Dados

Amostragem de Dados

- **Algoritmos de AM** podem ter **dificuldades em lidar com grande quantidade de objetos**.

- **Por exemplo,** Algoritmos baseados em instâncias, como o k-vizinhos mais próximos (**K-NN**, do inglês *k-nearest neighbours*), que podem apresentar problemas de **saturação de memória**.

Preparação dos Dados

Amostragem de Dados

- **Associado ao número de objetos em um conjunto de dados**, existe um **balanço entre eficiência computacional e acurácia** (taxa de predições corretas).

- Quanto **mais dados** são utilizados, **maior** tende a ser a **acurácia do modelo** e **menor a eficiência computacional** do **processo indutivo**.

- **Para** se **obter** um **balanceamento entre eficiência e acurácia**, **trabalha-se com uma amostra ou subconjunto de dados**.

Preparação dos Dados

Amostragem de Dados

- Uma pequena amostra pode não representar bem o problema que se deseja modelar.

- **A amostra deve ser representativa do conjunto de dados original**.

- O **ideal** é que a **amostra** não seja grande, mas que seus dados **obedeça**m **à mesma distribuição estatística que gerou o conjunto de dados original**.

Preparação dos Dados

Amostragem de Dados

- **Com isso**, seria capaz de fornecer uma estimativa da informação contida na população original, **extrai**ndo **conclusões** de um todo **a partir de uma parte**.
- No entanto, **não** é **possível garantir** que **isso** aconteça.
- O **especialista no domínio** pode **decidir** também que **um subconjunto dos objetos deve ser utilizado** para suas análises.
- Por **exemplo**, em uma análise de **pacientes de um hospital** podem ser utilizados apenas os **objetos** do **sexo feminino**.

Preparação dos Dados

Amostragem de Dados

- Existem **técnicas de amostragem estatística** que **aumentam a chance** de **obter uma boa estimativa**, são elas:
    - **Amostragem aleatória simples**
    - **Amostragem estratificada**
    - **Amostragem progressiva**

Preparação dos Dados

Amostragem de Dados **– Aleatória simples**

- Amostragem aleatória simples **sem reposição**: **exemplos são extraídos do conjunto original** para a amostra a ser utilizada e **cada exemplo pode ser selecionado apenas uma vez**.

- Amostragem aleatória simples **com reposição**: quando **uma cópia dos exemplos selecionados é mantida** no conjunto original.

- É **mais fácil de analisar**, pois **a probabilidade de escolher qualquer objeto se mantém constante**.

Preparação dos Dados

Amostragem de Dados – **Estratificada**

- Utilizada **quando as classes apresentam propriedades diferentes**, por **exemplo**, **números de objetos bastante diferente**s.

- Em problemas de classificação, um cuidado que deve ser tomado na amostragem diz respeito à distribuição dos dados nas diferentes classes.

Preparação dos Dados

Amostragem de Dados – **Estratificada**

- A existência de classes com uma quantidade significativamente maior de exemplos que as demais pode levar à indução de classificadores tendenciosos para as classes majoritárias.

- A opção mais simples é manter o mesmo número de objetos para cada classe.

- Uma opção é **manter o número de objetos em cada classe proporcional ao número de objetos da classe no conjunto original**.

Preparação dos Dados

Amostragem de Dados – **progressiva**

- Nesse caso, **começa com uma amostra pequena** e **aumenta progressivamente o tamanho da amostra extraída**, **enquanto a acurácia preditiva** continuar a **melhorar**.

- Como **resultado**, é possível definir a **menor quantidade de dados necessária**, **reduzindo ou eliminando a perda de acurácia.**

- Essa abordagem geralmente **fornece uma boa estimativa para o tamanho da amostra**.

Preparação dos Dados

Redução da Dimensionalidade

- **Muitos problemas** que podem ser tratados por técnicas de AM **apresentam um número elevado de atributos**.
- Porém, **poucas técnicas podem lidar** com essa **situação**.
- Esse problema é descrito como **maldição da dimensionalidade**.
- Uma forma de **minimizar o impacto** do problema da dimensionalidade **é combinar ou eliminar parte dos atributos irrelevantes**.

Preparação dos Dados

Redução da Dimensionalidade

- A redução da dimensionalidade **pode** ainda **melhorar o desempenho do modelo induzido**, **reduzir seu custo computacional** e **tornar os resultados** obtidos **mais compreensíveis**.

- Algumas **técnicas para a redução** da dimensionalidade podem ser divididas em:
  - **Agregação (Ex.:** *Principal Component Analysis* **(PCA))**
  - **Seleção de atributos (Ex.:** *Random Forest***)**

Preparação dos Dados

Redução da Dimensionalidade - **Agregação**

- **Combinam os atributos originais por meio de funções lineares ou não lineares**.

- Porém, **levam à perda dos valores originais**.

- **Algumas áreas** como: biologia, finanças, medicina e monitoramento ambiental, geralmente **é importante preservar os valore**s dos atributos **para que os resultados** obtidos **possam ser interpretados**.

Preparação dos Dados

Redução da Dimensionalidade – **Seleção de Atributos**

- **Permite**:
  - **Identificar atributos importantes**
  - Melhorar o desempenho de várias técnicas de AM
  - Reduzir a necessidade de memória e tempo de processamento
  - **Eliminar atributos irrelevantes e reduzir ruído**
  - **Lidar com a maldição da dimensionalidade**
  - **Simplificar o modelo gerado e tornar mais fácil sua compreensão.**
  - **Facilitar a visualização do dados.**

Preparação dos Dados

Redução da Dimensionalidade – **Seleção de Atributos**

- Na prática, **vários atributos passíveis de eliminação não são facilmente identificados**, o que torna pouco eficiente o uso de técnicas visuais.
- **Para avaliar a qualidade ou desempenho de um subconjunto de atributos**, **três abordagens são** utilizadas:
  - **Embutida**
  - **Baseada em filtro**
  - **Baseada em *wrapper***

Preparação dos Dados

Redução da Dimensionalidade – **Seleção de Atributos – Abordagem Embutida**

- A seleção do subconjunto **é embutida ou integrada no próprio algoritmo de aprendizado**.

- As **árvores de decisão realizam esse tipo de seleção** interna de atributos.

Preparação dos Dados

Redução da Dimensionalidade – **Seleção de Atributos – Abordagem Baseada em Filtro**

- Em uma etapa de pré-processamento, é **utiliza**do **um filtro sobre o conjunto** de atributos **original que filtra um subconjunto de atributos** do conjunto original, sem levar em consideração o algoritmo de aprendizado que utilizará esse subconjunto.

- As técnicas utilizam, por **exemplo**, **correlação entre atributos**.

Preparação dos Dados

Redução da Dimensionalidade **– Seleção de Atributos – Abordagem Baseada em *wrapper***

- **Utiliza o próprio algoritmo de aprendizado como uma caixa preta para a seleção**.

- Geralmente utilizada como uma técnica de amostragem.

- **Para cada possível subconjunto**, **o algoritmo é consultado** e **o subconjunto que apresentar a melhor combinação entre redução da taxa de erro e redução** do **número de atributos é** em geral **selecionado**.

Preparação dos Dados

Redução da Dimensionalidade – **Técnicas de Seleção de Subconjunto**

- **Pode ser vista como um problema de busca**.
- **Técnicas de partida para direção na busca** podem ser:
  - **Geração para trás** (*backward generation*), que **começa com todos os atributos e remove um por vez**.
  - **Geração para frente** (*forward generation*), que **começa sem nenhum atributo e inclui um atributo por vez**.
  - **Geração bidirecional** (*bidirectional generation*), em **que a busca pode começar de qualquer ponto e atributos podem ser adicionados e removidos**.
  - **Geração estocástica*** (*random generation*), quando **o ponto de partida da busca e atributos a serem removidos ou adicionados são decididos de forma estocástica**.

*Padrões estocásticos são aqueles que têm origem em processos não determinísticos, com origem em eventos aleatórios. Por exemplo, o lançar de dados resulta num processo estocástico, pois qualquer uma das 6 faces do dado tem iguais probabilidades de ficar para cima quando de seu arremesso. Fonte: https://educalingo.com/pt/dic-pt/estocastico. Data da consulta: 24/04/2018.

Preparação dos Dados

Redução da Dimensionalidade – **Técnicas de Seleção de Subconjunto**

- **Estratégias de busca** a serem adotadas são:
  - **Busca completa** (exponencial ou exaustiva), **que avalia todos os possíveis subconjuntos e é encerrada quando todos os subconjuntos forem testados**, **ou por um critério de parada**, que define terminar a busca **pelo melhor subconjunto de atributos**.
  - **Busca heurística** (sequencial), que **utiliza regras e métodos para conduzir a busca e que não garante que uma solução ótima seja encontrada**.
  - **Busca não determinística**, que está relacionada a geração estocástica.

Preparação dos Dados

Redução da Dimensionalidade – **Técnicas de Ordenação**

- Os **atributos são ordenados de acordo com sua relevância para um dado critério**, por exemplo, classificação dos objetos nas diferentes classes.

- Em problemas de classificação, **os atributos no topo da ordenação são selecionados** para utilização pelo classificador.

Bibliografia

BÁSICA:

- AGGARWAL, Charu C. **Artificial Intelligence: A Textbook**. New York: Springer: 2021.
- CHOLLET, François. **Deep Learning with Python, 2ed.** Shelter Island: Manning, 2021.
- GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**, 2 ed**.** Sebastopol: O'Reilly, 2019.

COMPLEMENTAR:

- GOODFELLOW, Ian; BENGIO, Yoshua, COURVILLE, Aaron. **Deep Learning**. Cambridge: MIT Press, 2016.
- RASCHKA, Sebastian; MIRJALILI, Vahid. **Python Machine Learning**. 3 ed. Birmingham: Packt, 2017.
- RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. 3 ed. Upper Saddle River: Pearson, 2010.
- TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. 2 ed. Upper Saddle River: Pearson, 2018.
- VANDERPLAS, Jake. **Python Data Science Handbook**. Sebastopol: O'Reilly, 2017.

ADICIONAIS:

- FACELI, Katti et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. 2ª Ed. Rio de Janeiro: LTC- Livros Técnicos e Científicos, 2021.
- LUGER, George F. **Inteligência Artificial** - 6ª ed. São Paulo: Pearson Education do Brasil, 2015.
- RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial: Uma Abordagem Moderna** - 4ª. Ed. GEN LTC, 2022.