Prof. Carlos Andrés Ferrero

Introdução a Aprendizado de Máquina e Ciência de Dados

Conteúdo

- Processo de Descoberta de Conhecimento
- Aprendizado Supervisionado e Não-supervisionado
- Paradigmas de Aprendizado

Introdução a Aprendizado de Máquina

- Em inglês *Machine Learning*
- Aprendizado de Máquina é uma subárea de Inteligência Artificial

- Definições

> "*The field of machine learning is concerned with the question of how to| construct computer programs that automatically improve with experience*", Tom Mitchel in 1997.
>
> O campo da aprendizado de máquina está preocupado com a questão de como construir programas de computador que melhoram automaticamente com a experiência.

> "*Machine learning research is part of research on artificial intelligence, seeking to provide knowledge to computers through data, observations and interacting with the world. That acquired knowledge allows computers to correctly generalize to new settings*", Yoshua Bengio recentemente.
>
> A pesquisa em aprendizado de máquina é parte da pesquisa em inteligência artificial, buscando fornecer conhecimento aos computadores através de dados, observações e interagindo com o mundo. Esse conhecimento adquirido permite que os computadores generalizem corretamente para novas configurações/situações.

- Outras definições:
  - What is Machine Learning? by Machine Learning Mistery
  - What is Machine Learning? by EMERJ

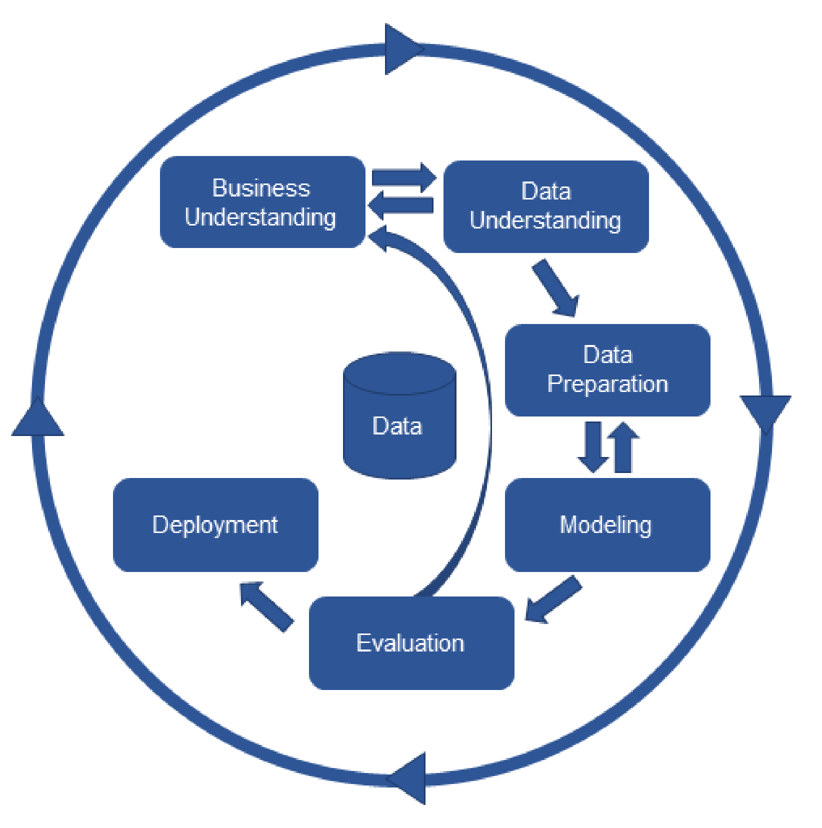
Processos

- KDD (*Knowledge Discovery in Databases*), proposto em 1989
- CRISP-DM (*Cross-industry standard process for data mining*), proposto em 1996
- Data Science Process, mais recente

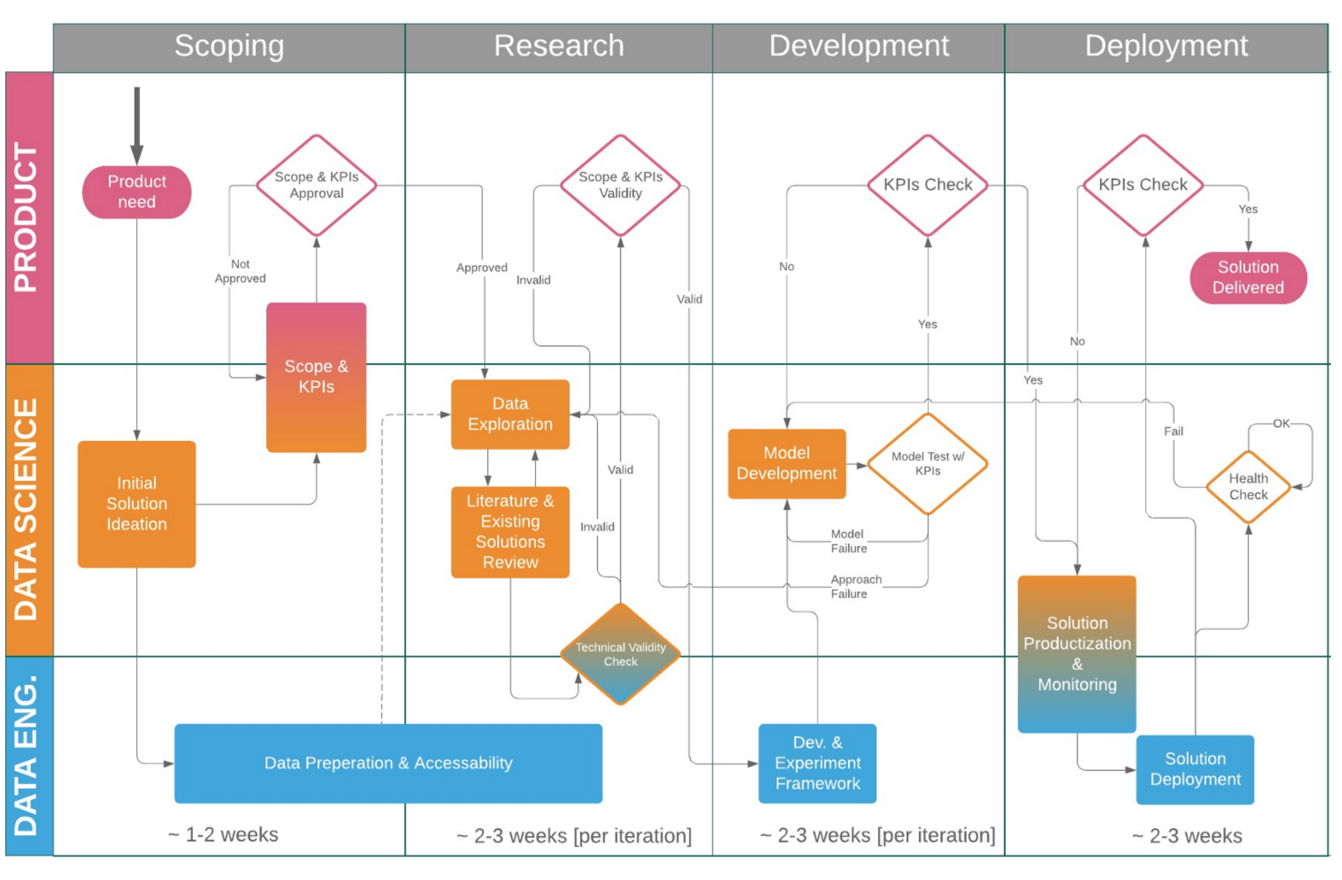
Processo de Descoberta de Conhecimento (KDD)

- Discussão da relação entre KDD e CRISP-DM

Processo CRISP-DM



Proceso de Ciência de Dados (for Startups)



Fonte: Palachy, Shya (2019). Data Science Project Flow for Startups. Towards in Data Science. Disponível em: https://towardsdatascience.com/data-science-project-flow-for-startups-282a93d4508d

- Principais Atores
  - Data Engineer (Engenheiro de Dados)
  - Data Scientist (Cientista de Dados)
  - Machine Learning Engenieer (Engenheiro de Machine Learning)

Preparação dos Dados

- Visa preparar os dados para uma análise eficiente e eficaz

- Algumas tarefas de pré-processamento:

  - Limpeza (remoção de ruídos e dados inconsistentes)

  - Integração (combinação de dados obtidos a partir de múltiplas fontes)

  - Seleção de Dados ou Redução (escolha dos dados relevantes para análise)

  - Transformação (transformação dos dados para um formato adequado para análise)

Modelagem

- Aplicação de algoritmos capazes de extrair padrões e/ou informações e conhecimento relevantes a partir dos dados pré-processados

- Algumas tarefas de mineração:

  - Análise descritiva estatística

  - Agrupamento

  - Modelos de Predição (classificação e regressão)

  - Associação (extração de regras de associação)

  - Detecção de Anomalias

Avaliação de modelos

- Avaliação da qualidade dos modelos ou conhecimento extraído na etapa de Mineração de Dados

- Algumas tarefas de pós-processamento:

  - Verificar a precisão dos modelos construídos em novos dados

  - Identificar conhecimentos verdadeiros e conhecimentos triviais

Aprendizado Supervisionado e Não-Supervisionado

As técnicas para aprender modelos a partir de dados são classificadas de acordo com o tipo de aprendizado em:

- **Aprendizado Supervisionado**

- **Aprendizado Não-supervisinado**

- **Aprendizado Semi-Supervisionado**

- Aprendizado por Reforço (*reinforcement learning*)

Aprendizado Supervisionado

No aprendizado supervisionado os exemplos estão constituídos por atributos, onde um desses atributos é *classe*

O atributo *classe* permite orientar o processo de aprendizando para ajustar um modelo aos dados

Exemplo	Atributo 1	Atributo 2	...	Atributo m	Atributo classe
e_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$	y_1
e_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,m}$	y_2
...	...	...	...	...	...
e_n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$	y_n

Neste aprendizado o objetivo consiste em encontrar uma função $f(X) = \hat{y}$, em que, dado os valores dos m atributos, estime y.

As tarefas de interesse do aprendizado supervisionado são preditivas e definidas por:

- **Classificação**: quando o valor da *classe* é um valor categórico (ou discreto)
- **Regressão**: quando o valor da *classe* é um valor numérico.

Aprendizado Não-superivsionado

No aprendizado não-supervisionado os exemplos estão constituídos por atributos, mas não há um atrbiuto *classe*

Isso indica que não tem como orientar o processo de aprendizado tentando acertar algum valor de atributo específico

Exemplo	Atributo 1	Atributo 2	...	Atributo m
e_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,m}$
e_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,m}$
...	...	...	...	...
e_n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,m}$

Neste tipo de aprendizado tentamos encontrar grupos de dados que apresentem comportamento similar

As tarefas de interesse dentro do aprendizado não-supervisionado são:

- **Agrupamentos**: agrupar exemplos de acordo com algum critério de distância ou similaridade
- **Extração de regras de associação**: encontrar relaçoes entre atributos e os valores dos atributos

Tarefas de Mineração de Dados

- As principais tarefas de mineração são:
  - Predição (seja classificação ou regressão)
  - Agrupamento
  - Associação
- A seguinte figura temos a relação dessas tarefas em um conjunto de dados convencional:

Fonte: Silva, L. A., Peres, S. M., & Boscarioli, C. (2017). **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil.

Aprendizado Semi-Supervisionado

- No aprendizado semi-supervisionado apenas alguns exemplos do conjunto de dados possuem o valor do atributo *classe*.
- Muitas vezes é muito caro classificar elementos. Alguns exemplos:
  - Transcrever segmentos de áudio
  - Identificar observações de experimentos físicos
  - Identificar os rostros de fotos do Facebook manualmente
- Nestas situações, com algumas instâncias rotuladas (exemplos classificados) é possível inferir os rótulos e classes de outras instâncias.

Fonte: Wang, F., & Zhang, C. (2007). **Robust self-tuning semi-supervised learning**. Neurocomputing, 70(16-18), 2931-2939.

Aprendizado por Reforço

- No aprendizado por reforço o sistema inteligente deve realizar uma sequência de ações permitidas (*policies*) em um ambiente para atingir um objetivo. Uma ação única não costuma ser tão importante, mas a sequência de ações pode levar a alcançar um objetivo com sucesso.
- O aprendiz (ou agente) interage com o ambiente e recebe recompensas ou penalidades de acordo com suas ações. Por tentativa e erro o agente aprende quais sequências de ações maximizam as recompoensas, i.e., quais tem mais chance de atingir o objetivo com sucesso.

Fonte: Alpaydin, E. (2007). **Introduction to Machine Learning**. MIT Press, p. 448 (579), 2010.

- Exemplos: aprender a jogar Games, veículos autônomos, mercado de ações, cuidados de saúde.

Paradigmas de Aprendizado

Alguns autores chamam de paradigmas os aprendizados supervisionado e não-supervisionado (CASTRO e FERRARI, 2016).

No entanto a literatura clássica divide os paradimgas de aprendizado da seguinte forma:

- Simbólico
- Baseado em Instância
- Estatístico
- Conexionista
- Genético

Leitura deste capítulo de livro:

- MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Conceitos sobre aprendizado de máquina**. Sistemas Inteligentes-Fundamentos e Aplicações, v. 1, n. 1, p. 32, 2003. Disponível em: http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. (até o conceito sobre completude).

Leituras Recomendadas (online)

- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. **From data mining to knowledge discovery in databases**. AI magazine, 17(3), pp.37-37. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

- Palachy, Shya (2019). **Data Science Project Flow for Startups**. Towards in Data Science. Disponível em: https://towardsdatascience.com/data-science-project-flow-for-startups-282a93d4508d