

Machine Learning 101

Autor: Dirso

4 de abril de 2019

Sumário

- 1 Processo de Data Science
- 2 Processo Modelagem
- 3 Amostragem - principais erros
- 4 Engenharia de features
- 5 Overview de modelos
- 6 Treinamento de modelos
- 7 Ferramentas
- 8 Contato

Pessoal

- Adilson Khouri, jogador de Magic the Gathering, nerd, apaixonado por computação, machine learning, brasileiro mas não sei jogar futebol nem sambar!



Figura: Eu no Peru palestrando e na Argentina trabalhando

Formação Acadêmica

- Bacharel em Sistemas de Informação (2011 - USP)
- Mestre em Sistemas de Informação (2016 - USP)
- Doutorando em Sistemas de Informação (cursando - USP)

Experiência de Mercado

- Programador na consultoria Arbit (2010-2011)
- Programador Itaú-Unibanco (2011-2013)
- Cientista de dados Sr. PagSeguro (2016 - 2018)
- Cientista de dados Sr. NuvemShop (Atual)
- Professor de Programação - SENAC (Atual)

The diagram illustrates the Machine Learning lifecycle, a continuous process from Start to End, passing through four main stages: Business Understanding, Data Acquisition & Understanding, Modeling, and Deployment.

Start leads to **Business Understanding**.

Business Understanding leads to **Data Acquisition & Understanding**.

Data Acquisition & Understanding leads to **Modeling**.

Modeling leads to **Deployment**.

Deployment leads to **Customer Acceptance**, which leads to **End**.

Modeling is further detailed with the following components:

- Feature Engineering:** Transform, Binning, Temporal, Text, Image, Feature Selection.
- Model Training:** Algorithms, Ensemble, Parameter Tuning, Retraining, Model management.
- Model Evaluation:** Cross Validation, Model Reporting, A/B Testing.

Deployment includes **Scoring, Performance monitoring, etc.**

Data Acquisition & Understanding includes the following considerations:

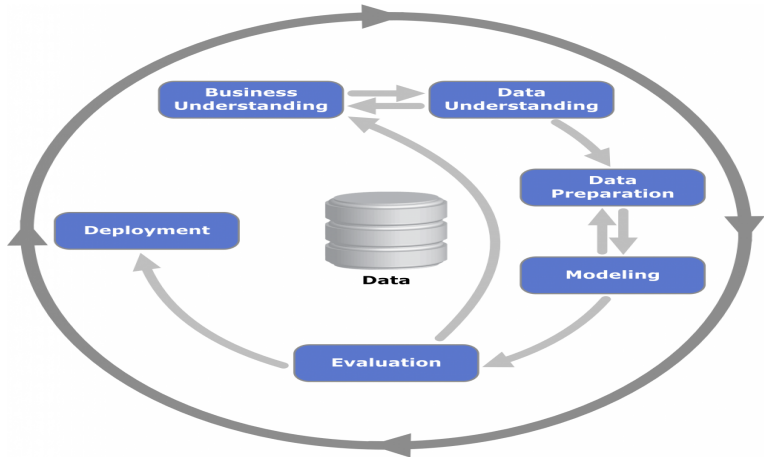
- Data Source:** On-Premises vs Cloud, Database vs Files.
- Pipeline:** Streaming vs Batch, Low vs High Frequency.
- Environment:** On-premises vs Cloud, Database vs Data Lake vs ... Small vs Medium vs Big Data.
- Wrangling, Exploration & Cleaning:** Structured vs Unstructured, Data Validation and Cleanup, Visualization.

The **Deployment** stage is supported by **Model Store**, **Web Services**, and **Intelligent Applications**.

Figura: Obtido em: <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview>

Processo Modelagem

- Existem muitos processos para usar na área de big data, um dos mais simples e práticos é o **CRISP-DM**



Processo Modelagem

- Entendimento de negócio
- Entendimento dos dados
- Preparação dos dados
- Modelagem
- Avaliação do modelo
- Deploy

Método científico

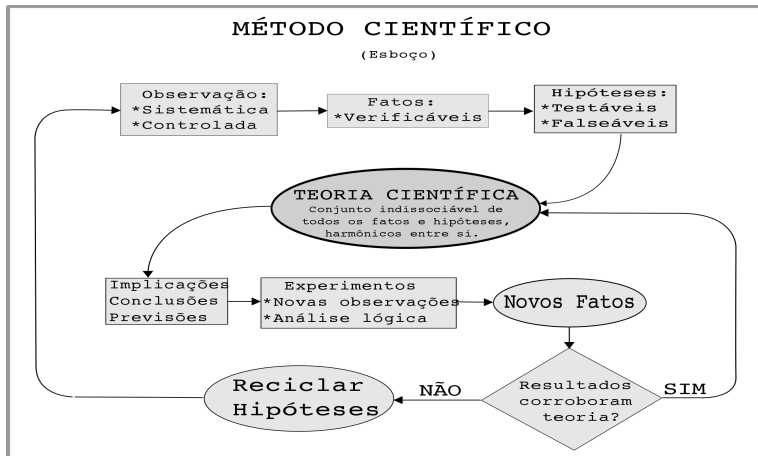


Figura: Método científico

Método científico

- Então método científico nos impede de cometer erros?

Método científico

- Por que Dirso ficou doente na Argentina em 2018?



Figura: Na minha cabeça a Argentina é tipo wall de game of thrones!!!!

Explicar o problema de monty hall

- Alguém conhece o problema de monty hall?

Método científico

- Como essas três perguntas se relacionam com essa apresentação?

Amostras viesadas

- Precisamos de informação precisa e sem viés para tomarmos boas decisões.
- Se você “cria conhecimento” ou “toma decisões” usando informação viesada você não está sendo # datadriven
- A probabilidade de tomar decisões ruins aumenta... e decisões ruins costumam ser caras...

processo de KDD

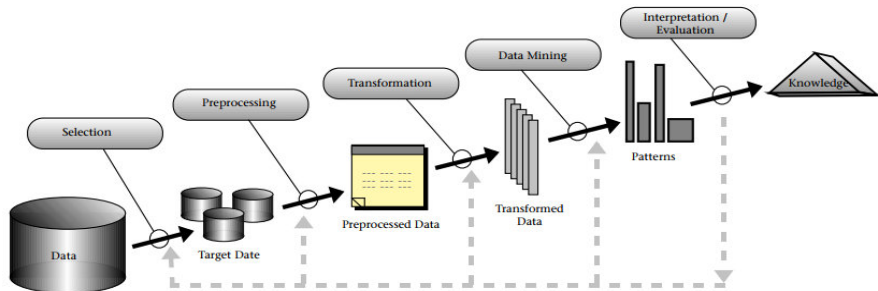


Figura: Processo de KDD

- Se você cometer um erro durante a etapa de: “seleção” os passos seguintes e suas conclusões estarão erradas.

Amostragem 101

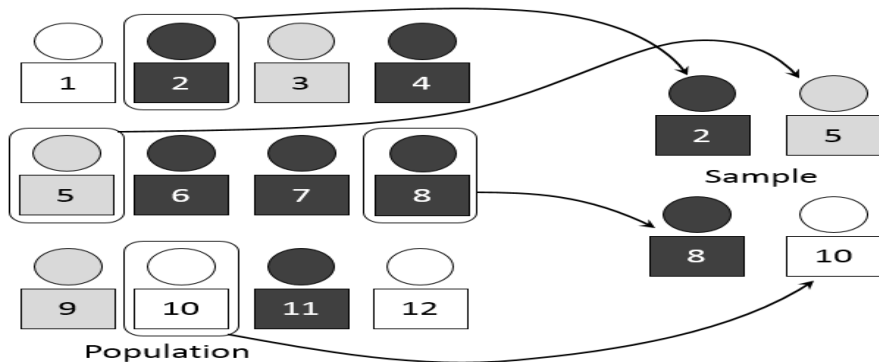


Figura: Overview de amostragem

- O subconjunto (amostra) de elementos deve ser representativa da população.

Bias de auto seleção

- Suponha um estudo estatístico sobre detalhes íntimos da sexualidade de estudantes em universidades. Algumas pessoas provavelmente vão mentir.
- Uma pesquisa online sobre quem gosta de usar computador.
- Em ambos as pessoas selecionadas vão ter seus comportamentos diferentes da população geral.

Undercoverage Bias

- Digest em 1936 fez uma pesquisa eleitoral que previa vitória larga do candidato Landon em relação ao candidato Roosevelt. Roosevelt ganhou com uma margem larga, a pesquisa era feita por telefone, na época pessoas pobres (maioria da população que era a favor de Roosevelt) não tinha telefone. Essa foi uma das causas do erro estatístico.

Survivorship Bias

- Ocorre quando as observações estudadas no fim da investigação são não aleatórias em comparação as presentes no começo da observação.

Survivorship Bias

- Exemplo da segunda guerra mundial (tiros em avião)

Engenharia de features

- Modelos usam muitas variáveis para tomar decisões
- Encontrar boas variáveis é parte fundamental para um modelo
- Citar exemplo de variáveis de transações financeiras
- Citar exemplo de variáveis de pagamento de assinaturas
- Citar exemplo de um classificador de brasileiros e peruanos

Modelos



Figura: Brincadeira, cada modelo trabalha internamente de uma forma distinta!

Modelos

- Modelos tomam decisões baseados em diversas variáveis para, entre outras coisas, classificar dados
- Quem são peruanos e quem são brasileiros nessa sala?
- Há modelos para classificar em duas classes ou mais.

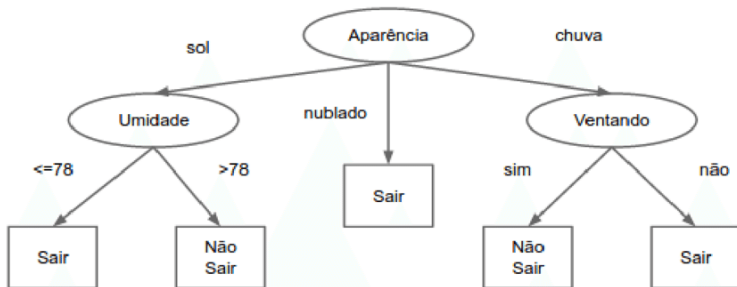


Figura: Exemplo de árvore de decisão para sair de casa

Treinamento

- O processo de treinamento é único para cada modelo mas a forma como se treina um modelo é parecida
- Os dados são divididos em treino (70%) e teste (30%)
- O conjunto de treino é apresentado ao modelo com os rótulos de cada observação
- Tipicamente usa-se uma validação cruzada para treinar o modelo

Validação

- O modelo é validado com o conjunto de teste, o qual não deve exibir os rótulos para o modelo

		PREDITO	
		Classe A	Classe B
VERDADEIRO	Classe A	VP	FN
	Classe B	FP	VN

Figura: Obtido no link: Scielo

Validação - outras métricas

- Se usarmos a matriz de confusão acima podemos obter outras métricas
- Citar o problema das classes de seller (relacionar com $F1$)

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

Ferramentas

- Na teoria pode-se usar qualquer linguagem de programação para trabalhar com Data Science
- Na prática usa-se, majoritariamente, a plataforma R e a linguagem Python (com alguns pacotes científicos)
- [Sci-kit learn](#)
- [blog sobre R](#)

Hands on

- Treinar modelo em Python com o time.

Fim!

Agradeço a Laura por me dar a possibilidade de fazer a apresentação e agradeço a vocês por assistirem :)

Contato

- E-mail: *adilson.khouri.usp@gmail.com*
- Phone: +55119444 – 26191
- [Linkedin](#)
- [Curriculum Lattes](#)
- [Código fonte GitHub](#)