UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS

Introdução à Ciência de Dados

Prof. Dr. Francisco Carlos Souza
Prof. Dr. Anderson Carniel

Sumário

- Fases genéricas da Ciência de Dados
- Tecnologias
  - Captura de dados
  - Armazenamento
  - Processamento
  - Aprendizado de máquina
  - Visualização de dados
- Entendendo os dados
  - Exemplos de csv, json e xml
- Estatística
- Dica para ser um cientista de dados de sucesso

Fases Ciência de dados

- Como todos projetos de software, projetos de ciência de dados possuem um início, meio e fim.
- Um projeto de ciência de dados possui algumas fases genéricas que foram baseadas no processo de data mining chamado CRISP-DM (*Cross Industry Standard Process for Data Mining*).
- Esse processo é composto por *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation e Deployment.*

Fases Ciência de dados

- O processo de ciência de dados é composto basicamente por 6 fases

Fase 1 - Descoberta

- Consiste na aquisição de dados de todas as fontes internas e externas identificadas, como:
  - Logs de servidores da web
  - Dados coletados das mídias sociais
  - Conjuntos de dados do censo
  - Dados transmitidos de fontes usando APIs

Fase 2 – Preparação dos dados

- Os dados podem ter muitas inconsistências, como:
  - valor ausente,
  - colunas em branco
  - formato de dados incorreto que precisa ser limpo.
- Processar, explorar e condicionar dados antes da modelagem.
- Quanto mais limpos os dados, melhores são as previsões.

Fase 3 – Planejamento do modelo

- Necessário determinar o método e a técnica para desenhar a relação entre as variáveis de entrada.
- É realizado usando diferentes fórmulas estatísticas e ferramentas de visualização.
- Ferramentas: serviços de análise SQL, R e SAS/access.

Fase 4: Construção do Modelo

Construção do modelo utilizando Aprendizado de máquina

- Cientista de dados distribui conjuntos de dados para treinamento e teste.
- Técnicas como associação, classificação e clustering são aplicadas ao conjunto de dados de treinamento.
- O modelo construído é testado com o conjunto de dados "teste".

Fase 5 - Operação

- Ocorre a entrega, o modelo final de linha de base com relatórios, códigos e documentos técnicos.
- O modelo é implantado em um ambiente de produção em tempo real após testes completos.

Fase 6 – Comunicação dos Resultados

- As principais conclusões são comunicadas a todas as partes interessadas.
- Essa comunicação ajuda a decidir se os resultados do projeto são bem-sucedidos ou fracassados, com base nas entradas do modelo.

Tecnologias em Ciência de Dados

- O Cientista de dados deve ter a capacidade de construir soluções e otimizar modelos para responder questionamentos de organizações.
- Para isso, somado ao conhecimento teórico é necessário o desenvolvimento de algoritmos e artefatos para gerar Insight.
- Conhecer tecnologias para auxiliar neste processo, é uma tarefa crucial para um Cientista de dados

Tecnologias: Captura de dados

- São métodos para obtenção de dados externos e ocorre na fase 1, os mais comuns são como:
  - *Web crawler*
  - IOT (Internet das Coisas)
  - *Logs* de sistema
  - *Wearables*
  - Equipamentos de rede

Web crawlers

- Tratam-se de *scripts* e algoritmos para **coleta** de **dados** e **conteúdos** na internet.
- Deve-se tomar **cuidado** com **disponibilização** de dados obtidos por meio desses mecanismos.
- Apesar de está na internet, **nem** todos os **dados** são **públicos**.
- Diversas linguagens possuem **ferramentas para web crawler**, tais como:
  - *Scrapy* em Python
  - *Crawler4j* para Java
  - *Mechanize* em Ruby

Tecnologias: Armazenamento

- As tecnologias de armazenamento como **SQL** e **NoSQL** são essenciais para guardar grandes volumes de dados.
- Além disso **mecanismos de busca** e de **serialização** de dados, como *elasticsearch* e *json* facilitam o dia-a-dia do Cientista de dados.

PostgreSQL

- Atualmente o P*ostgre* é um dos SGBDs mais utilizados por se tratar de uma tecnologia **open-source** e **estável**.
- Suporte excelente a **Full-text search.**
- **Geração** nativa de **UUID** (*universally unique identifier*)
- Permite a **manipulação** de dados **JSON** e **JSONB** (versão binária)
- Versão 12.2

Tecnologias: Processamento

- Para **processar dados** a partir de tecnologias de armazenamento, é **possível realizar** com a grande **maioria** das **linguagens** de **programação**.
- Contudo, para garantir maior produtividade o ***Python*** e o **R** são aquelas **mais utilizadas**.
- Além disso para processamento de dados não-estruturados é necessária o uso de tecnologias específicas, como:
  - **OpenCV** para imagens
  - **Tesseract** para reconhecimento de caracteres
  - **Google Speech Recognition** para fala
  - **NLTK** em processamento de linguagem natural

Tecnologias: Processamento

R e Python

- Ambas linguagens são:
  - Gratuitas
  - Simples de instalar
  - Escrita mais próxima da linguagem natural
  - Possuem uma extensa quantidade de pacotes e bibliotecas para análise de dados
  - Cada linguagem possuí seus prós e contras em diferentes cenários

Linguagem R

- Criada em 1995
- Originada a partir da implementação da linguagem S da Bell Labs
- Versão atual 4.0 em 2020
- Modelos estatísticos podem ser escritos com poucas linhas de código
- A mesma funcionalidade pode ser escrita de diversas formas diferentes

- Fácil para escrever fórmulas complexas
- Grande números de pacotes para análise de dados e ML
- CRAN é o repositório do R
- R possui IDE bastante consolidada, chamada Rstudio

Linguagem R

- R tem credibilidade devido sua história ao longo dos anos e possui uma **comunidade confiável e forte no setor de dados**
- R é considerada também uma **ferramenta de visualização e gráficos.**
- Permite que os Cientistas de Dados criem **gráficos interativos** a partir dos **resultados** das análises de dados.
- O R **possui diversos pacotes** que facilitam o processo de manipulação dos dados

Linguagem Python

- Criada em 1991
- Inspirada na linguagem C, Modula-3 e ABC
- Versão atual 3.8.2 em 2020
- Codificação e *debugging* eleva nível de produtividade
- Possui um padrão definido, permitindo que diferentes tipos de funcionalidades sejam escritas da mesma forma

- É flexível e permite manipular dados de diferentes maneirais
- Não foi criada inicialmente para análise de dados e ML
- Pip é um repositório da linguagem com diversas bibliotecas e ferramentas
- O python possui diversas IDEs amplamente utilizadas, como Pycharm e Spyder

Linguagem Python

- Criada para produzir código limpo é fácil de manter de maneira rápida
- Linguagem open-source e multiplataforma
- Com a linguagem é possível
  - Construção de sistemas web
  - Construção de aplicativos para celular
  - Construção de sistemas desktop
  - Análise de dados e Inteligência Artificial

Tool-kit

- A programação faz parte do trabalho de um cientista de dados.
- Assim, para melhorar a qualidade dos resultados e a produtividade é necessário ferramentas
- Escolha seu ambiente de desenvolvimento favorito
  - Linux e Windows
    - Ambientes de dev. no windows costumam ser mais trabalhosos de configurar, então o prepare com antecedência
  - IDE e editores de texto
  - Instalação das linguagens e pacotes
  - Dockers e VirtualEnv

Bibliotecas

Linguagem R	
Biblioteca	**Função**
Database drivers	Conexão com BD
reshape2	Ajustar formato de datasets
dplyr	Tratamento de datasets
stringr	Manipulação de texto
ggplot2	Visualização de dados
caret	Modelagem Estatística

Bibliotecas

Linguagem Python	
Biblioteca	**Função**
Pandas	Tratamento de datasets
Numpy	Manipulação de arrays multidimensionais
Scikit-learn	Aprendizado de máquina
Seaborn	Visualização de gráficos estatísticos
Matplotlib	Criação de gráficos

Tecnologias: Aprendizado de máquina

- Aprendizado de máquina **não é um método de ciência de dados**, porém é um conceito que é **aplicado com frequência** em problemas de análise de dados.
- Aprendizado de máquina é uma subárea da inteligencia artificial que **possui capacidade de ensinar máquinas a partir de dados**.
- Com AM é possível **reconhecer padrões**, **realizar predições** e **tomar decisões**.

Tecnologias: Aprendizado de máquina

- Dependendo do **propósito**, **diferentes algoritmos** podem ser utilizados. Os mais comuns são:
  - Redes Neurais (*neural networks*)
  - Máquina de Vetores Suporte (*support vector machine*)
  - Regressão Logistica (*logistic regression*)
  - Clusterização K-means
  - Árvore de decisão (*decision trees*)

Tecnologias: Aprendizado de máquina

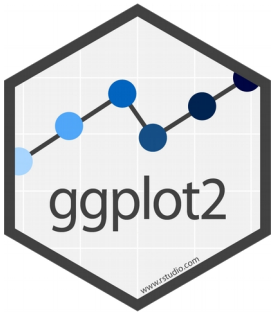
- Ferramentas e bibliotecas também são utilizadas para aumentar a **produtividade** e a **qualidade** dos resultados.

Visualização de Dados

- A **conversão** dos dados em **artefatos visuais** é uma atividade crucial na ciência de dados. Essa atividade pode ocorrer em todo processo para otimizar análise.
- Por meio de artefatos visuais é possível também identificar **padrões**, **tendências** e **conclusões** que **não foi** identificado pela máquina.
- A visualização de dados pode ser por meio de **gráficos**, **tabelas**, **infográficos**, **mapas**, etc.

Visualização de Dados

- **Ferramentas e bibliotecas**



matplotlib

Seaborn

Data-Driven Documents

pentaho™
open source business intelligence™

Visualização de Dados

- A **conversão** dos dados em **artefatos visuais** é uma fase crucial na ciência de dados.
- Por meio de artefatos visuais é possível também identificar **padrões**, **tendências** e **conclusões** que **não foi** identificado pela máquina.
- A visualização de dados pode ser por meio de **gráficos**, **tabelas**, **infográficos**, **mapas**, etc.

Entendendo os dados

- Dados estruturados
  - Lista e Matrizes
  - Tabelas
  - Redes (rotas de viagem)
- Não estruturados
  - Textos
  - Imagens
  - Vídeos
  - Sons

Outras extensões comuns

{JSON}

JavaScript Object Notation

CSV - Comma-separated values

- Valores que devem ser separados por vírgulas
- Exemplo:

```
comprimento_sepala, largura_sepala, comprimento_petala, largura_petala, especie
5.1 , 3.5 , 1.4 , 0.2 , sedosa
4.9 , 3 , 1.4 , 0.2 , sedosa
4.7 , 3.2 , 1.3 , 0.2 , sedosa
4.6 , 3.1 , 1.5 , 0.2 , sedosa
```

JavaScript Object Notation

- JSON é um formato baseado em texto estruturado para representação de dados.

```
{
    "primeiroNome": "Alan",
    "ultimoNome": "Smith",
    "estaVivo": verdadeiro,
    "idade": 25,
    "endereco": {
        "rua": "Presidente de moraes 222",
        "cidade": "Dois Vizinhos",
        "estado": "PR",
        "cep": "85660-000"
    },
    "numerosTelefone": [
        {
            "tipo": "casa",
            "numero": "3436-1212"
        },
        {
            "tipo": "trabalho",
            "numero": "3436-2522"
        },
        {
          "tipo": "celular",
          "numero": "99999-9997"
        }
    ],
    "filhos": [],
    "conjuge": null
}
```

Extensible Markup Language

- XML é uma linguagem de marcação que define um conjunto de regras para codificação e estruturação de documentos

```
<pessoa>
    <primeiroNome> Alan </primeiroNome>
    <ultimoNome> Smith </ultimoNome>
    <idade> 25 </idade>
    <endereco>
        <rua> Presidente de moraes 22 </rua>
        <cidade> Dois Vizinhos </cidade>
        <estado> PR </estado>
        <cep> 85660-000 </cep>
    </endereco>
    <numeroTelefone>
        <tipo> casa </tipo>
        <numero> 3436-1212 </numero>
    </numeroTelefone>

    <numeroTelefone>
        <tipo> trabalho </tipo>
        <numero> 3436-2522 </numero>
    </numeroTelefone>
    <numeroTelefone>
        <tipo> celular </tipo>
        <numero> 99999-9997 </numero>
    </numerosTelefone>
    <genero>
        <tipo> masculino </tipo>
    </genero>
</pessoa>
```

Estatística

- A estatística na ciência de dados tem um papel de dar confiabilidade nos resultados e gerar *insights* de mais valor.
- Técnicas e algoritmos de aprendizado de máquina utilizam conceitos de estatística, portanto é também utilizado na construção de um modelo de AM
- Para a ciência de dados, se precisa conhecer conceitos básicos de:
  - coleta de dados em estatística
  - representatividade da população
  - distribuição
  - normalidade dos dados, e
  - hipóteses para confirmação dos resultados

Estatística

- Conceitos que devemos conhecer

Média, mediana e moda
Distribuição de dados (Normal, exponencial, binominal)
Desvio Padrão e Variância
Teste de Hipóteses
Teste de significância
Análise de Variância
Níveis de confiança

Como ser um cientista de dados de sucesso?

- Unir teoria e prática
  - Conhecer os aspectos teóricos e aplicá-los por meio das tecnologias dezenas de vezes
  - Com isso se adquire, experiência para decidir e como usar uma técnica no momento certo
  - Saber questionar se um resultado bom está correto ou foi um erro de modelagem
  - Tudo depende do nosso esforço para buscar conhecimentos extras