Introdução

Mineração de Dados

Universidade Federal do ABC

Introdução

Mineração de Dados

Dados, dados e mais dados

"*Estamos nos afogando em informação, mas famintos por conhecimento*" - John Naisbitt

- Progresso na coleta e armazenamento de dados tornaram comuns bases de dados enormes
  - Desde o início dos anos 90
  - Não há indícios de que isso irá parar tão cedo
  - Como tirar o melhor proveito possível dos dados

Dados, dados e mais dados

Figura 1: Comércio Eletrônico

- O que você procura
- O que você compra
- O que você critica

Dados, dados e mais dados

Figura 2: Twitter

- O que você escreve
- De onde você escreve
- Com quem você conversa
- Quem você lê

Dados, dados e mais dados

Figura 3: Facebook

- Quem é você
- O que você faz
- Com quem e quando você faz
- O que você acha sobre o que os outros fazem

Dados, dados e mais dados

Figura 4: *Smartphones*

- Com quem você fala
- Por onde você anda
- Como você se locomove
- Sem entrar em detalhes sobre *apps*

Dados, dados e mais dados

Figura 5: Internet das coisas

- Acompanhamento 24/7
- Diversas questões preocupam nessa área
  - Não iremos abordar esses assuntos *nesta* disciplina

Dados, dados e mais dados

- Esses são casos extremos, mas a situação é comum:
  - Concessionárias
  - Hospitais
  - Escolas
  - Bancos
  - ...
- Empresas ainda tomam decisões importantes considerando apenas em intuição
  - "HiPPO" *the highest-paid person's opinion*
- Pessoas se apoiam muito em experiência e intuição mas pouco em dados

Introdução

Mineração de Dados

Mineração de Dados

- Amadureceu conforme as bases de dados cresceram em tamanho e complexidade.
- Disciplina interdisciplinar relacionada a:
  - Estatística
  - Bancos de dados
  - Aprendizado de Máquina
  - Reconhecimento de Padrões
- Fronteiras entre as áreas **não** são rígidas
- Cervejas e fraldas (*fun fact*): https://www.kdnuggets.com/news/2000/n14/8i.html

Mineração de Dados

- Inicialmente era conhecida como a etapa de extração de padrões dentro do Processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*)
  - Com o tempo alguns pesquisadores começaram a utilizar como sinônimos

Mineração de Dados

Definição (Hand, Manilla & Smith) *Mineração de dados é a análise de bases de dados observacionais (frequentemente grandes) para encontrar relações desconhecidas e para sumarizar os dados em formas que sejam compreensíveis e úteis para o dono dos dados.*

Mineração de Dados

- Alguns termos na definição chamam atenção:
  - dados observacionais
  - relações desconhecidas
  - formas compreensivas
  - formas úteis

Etapas

- Identificação e formalização do problema
  - possivelmente a mais difícil
- Pré-processamento
  - Costuma ser a que leva mais tempo
- Extração de Padrões
- Pós-processamento

CRISP-DM

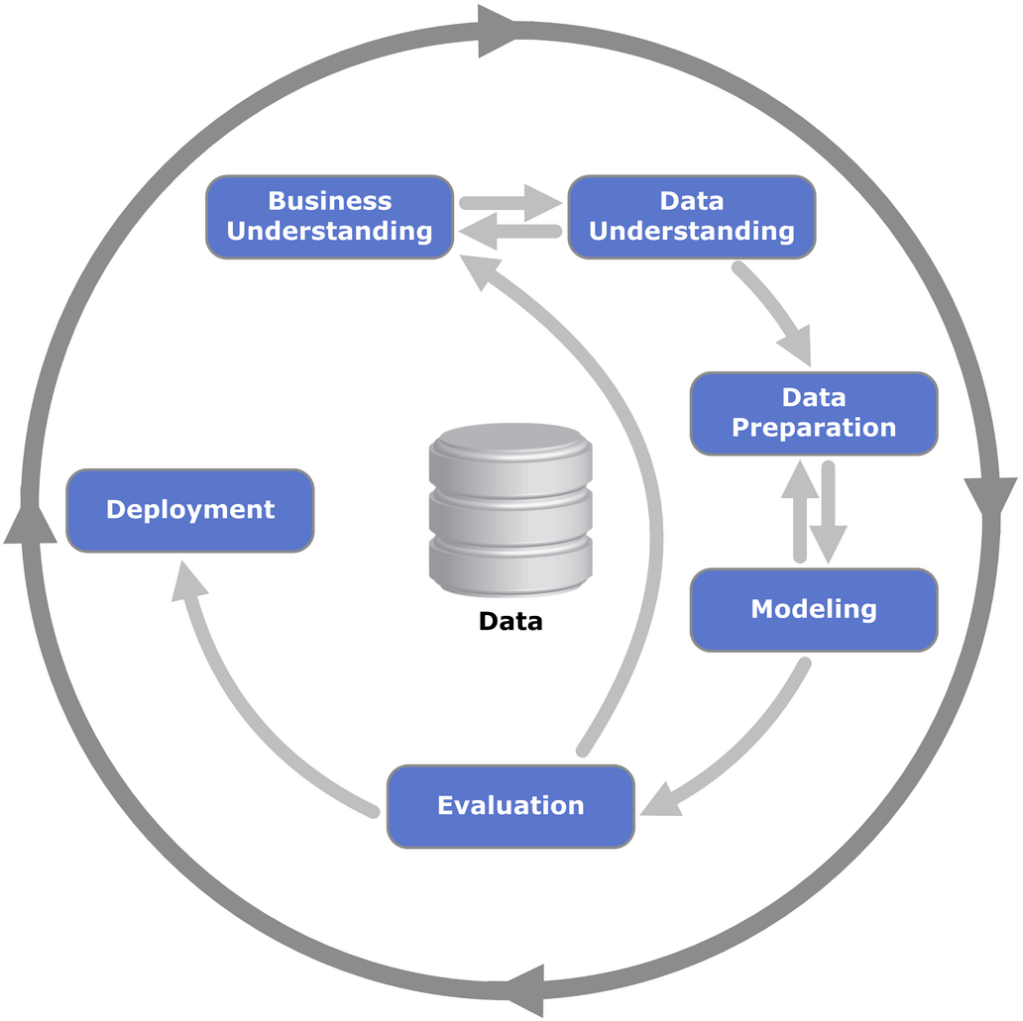

Figura 6: Cross-industry standard process for data mining

Identificação e formalização do problema

- Quais são as principais metas do processo?
- Como o desempenho será avaliado?
- A compreensibilidade do modelo deve ter peso maior do que seu desempenho?
- Qual o custo de um erro?

Pré-processamento

- Integração de dados: múltiplas fontes e formatos
- Transformação: normalização/adequação
- Limpeza: dados inválidos são frequentes
  - Por exemplo, pressão sanguínea = 0 [*Pima Indians Diabetes*]
- Seleção e Redução: formas de "simplificar" os dados, discutiremos em outra aula

Extração de Padrões

- Onde a mágica acontece
- Aplicação de algoritmos de AM, estatística etc.
  - Tarefas preditivas
  - Tarefas descritivas
- Não existe algoritmo que seja melhor para todos os problemas
  - *No free-lunch theorem* (Wolpert)

Pós-processamento

- O que o modelo achou que é interessante?
- Apresentação do modelo de forma compreensível

Áreas relacionadas

Business Intelligence

Data Science

Big Data

Predictive Analytics

...

O que não iremos cobrir

- Extração de dados
  - *Crawlers*
- Armazenamento e organização de dados
  - *Data Warehouse*
  - *Online Analytic Processing* (OLAP)
- Pós-processamento

Referências

- D. Hand, H. Manilla e P. Smith. Principles of Data Mining. **Capítulo 1**
- S. Rezende. Sistemas Inteligentes: Fundamentos e Aplicações. **Capítulo 12**
- P. Tan, M. Steinbach e V. Kumar, Introduction to Data Mining. **Capítulo 1**
- Breiman, Leo. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (2001), no. 3, 199–231. doi:10.1214/ss/1009213726. https://projecteuclid.org/euclid.ss/1009213726