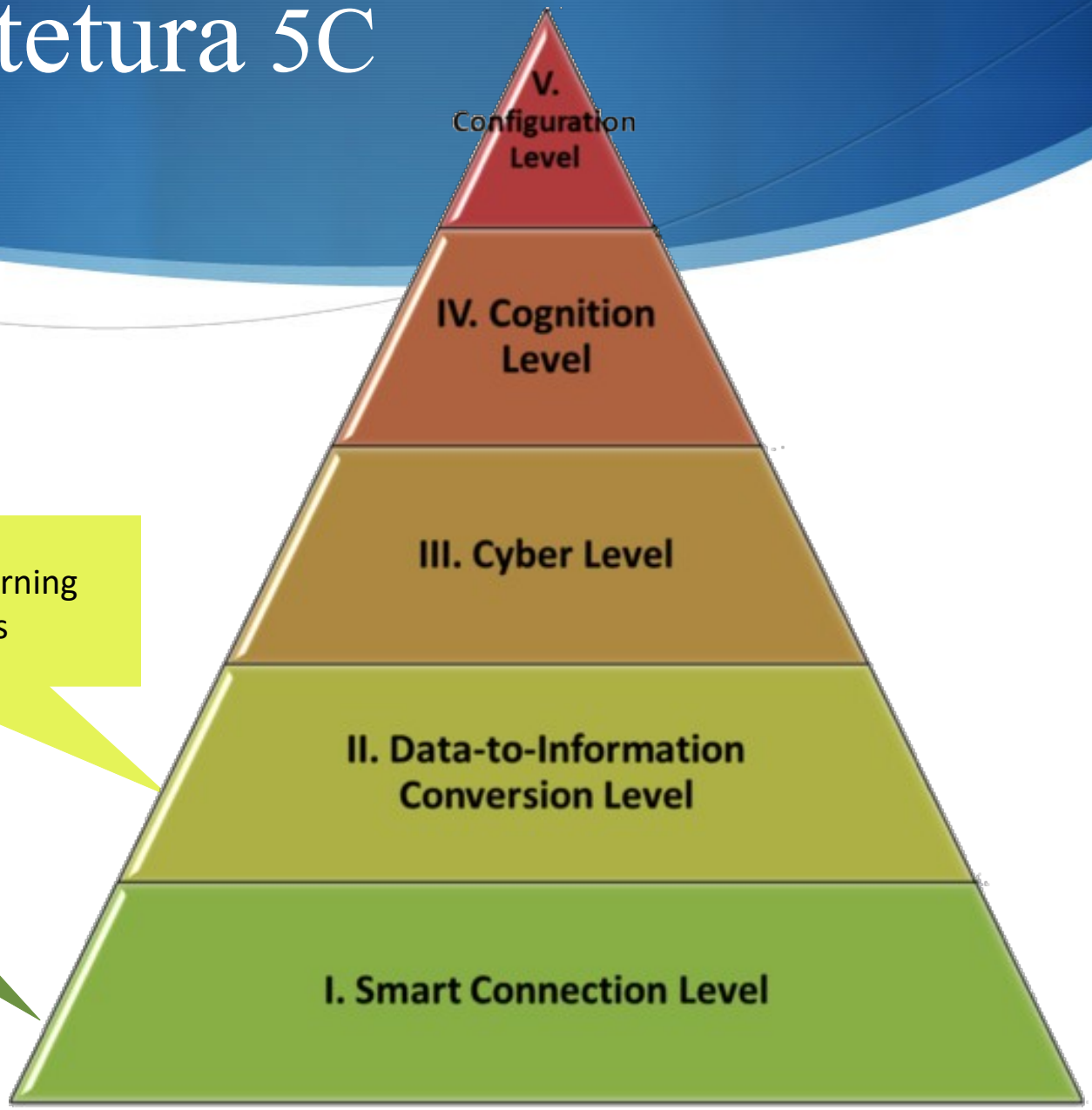
Big Data e Machine Learning

Daniela Maria Uez
dani.uez@gmail.com

1 AGOSTO 2019

UCS
UNIVERSIDADE
DE CAXIAS DO SUL

Arquitetura 5C



Big Data e Machine Learning
Manipular os dados

IIoT –
Adquirir dados

Por que analisar os dados?

Os dados armazenados só são úteis quando pode-se gerar informações a partir deles

Essas informações são utilizadas na tomada de decisões

Os dados não são completamente randômicos – possuem padrões

Encontrar padrões nos dados é o foco do ML

A aplicação de métodos de aprendizado de máquina em grandes bases de dados é chamada de Mineração de Dados(data mining)

Aprendizagem de máquina

- ML usa a teoria estatística para modelos matemáticos que permitam inferir um conhecimento - preditivo ou descritivo – a partir de uma amostra dos dados

- Existem muitos algoritmos - Como escolher o correto?
  - Supervisionados
  - Não supervisionados
  - Por reforço
  - ...

Qual algoritmo usar?

- Depende do objetivo:
  - Se quer saber a previsão do tempo para o mês de agosto 2019
  - Se quer saber qual filme recomendar com base nos filmes que já foram assistidos

- Depende dos dados disponíveis
  - Dados podem ter atributos qualitativos e quantitativos

Usando ML

1) Coletar os dados
2) Preparar os dados de entrada
3) Analisar os dados de entrada
4) Treinar o algoritmo
5) Testar o algoritmo

Pré-processamento

- Normalmente os dados estão longe da perfeição para usar um algoritmo de ML
- Pré-processamento é composto de duas fases
  - Limpeza dos dados
  - Transformação dos dados

Limpeza dos dados

- Preencher dados ausentes
- Ajustar dados com ruídos
- Identificar e/ou remover valores aberrantes
- Resolver inconsistências
- Formatação de dados de forma a adequá-los à ferramenta de mineração

Características do conjunto de dados

- **Dimensão**: é o número de atributos que os objetos desse conjunto de dados possuem
  - Conjuntos com muitas dimensões não são bem classificados pelos algoritmos

- **Dispersão**: variabilidade da distribuição dos dados com relação à média - alguns algoritmos funcionam melhor com dados dispersos

Transformação dos dados

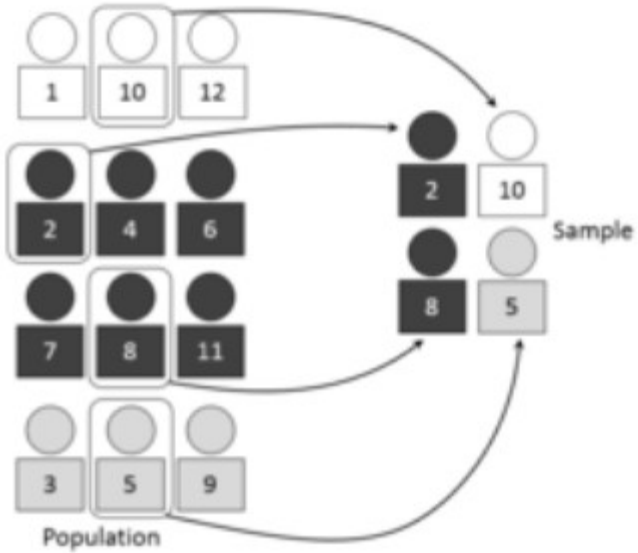
- É necessário para obter os dados numa forma mais apropriada para a mineração de dados.
- Em geral, transformação de dados envolve:
  - Agregação
  - Amostragem
  - Redução de dimensionalidade
  - Discretização e binarização
  - Transformação de variáveis

Agregação e Amostragem

- **Agregação**: combinação de dois ou mais objetos em um único
- **Amostragem:** seleciona um subconjunto dos objetos de dados a serem analisado
  - Usar uma amostra funcionará tão bem quanto usar o conjunto inteiro de dados se a amostra for representativa
  - Uma amostra é representativa se tiver aproximadamente as mesmas propriedade do conjunto original de dados.

Técnicas de Amostragem

- Amostragem estratificada: números proporcionais de objetos são selecionados de cada grupo



Binarização

- Binarização: Alguns algoritmos requerem que os dados estejam na forma de atributos binários (0 ou 1)
  - Tanto atributo contínuos quanto discretos podem precisar ser transformados em atributos binários
  - Pode ser necessário mais de um atributo binário

Discretização

- Discretização: transformação de um atributo contínuo em um categorizado
  - Por ex: atributo contínuo comprimento pode precisar ser transformado em um com categorias discretas: curto, médio ou longo

Transformação de variáveis

- Uma transformação que seja aplicada a todos os valores de uma variável
- Tipos:
  - Transformações funcionais simples: aplica uma função matemática a cada valor individualmente
    - Ex: x^k, log **x**
  - Normalização: faz o conjunto inteiro de valores ter uma determinada propriedade – usada em estatística

Referências
- Exemplos uso de ML empresas -

- Uso de Redes Neurais Artificiais para a Detecção das Doenças Olho de Boi e Manchas de Sarna em Maçãs - https://repositorio.ucs.br/xmlui/bitstream/handle/11338/3724/TCC%20Iago%20dos%20Passos.pdf?sequence=1&isAllowed=y

- Aplicação de Processo de Classificação e Técnica de Bayes na Base de Dados de Acidentes Ocupacionais de uma Empresa Metalúrgica - https://repositorio.ucs.br/xmlui/bitstream/handle/11338/3913/TCC%20Charles%20da%20Luz%20Pola.pdf?sequence=1&isAllowed=y

Redução de Dimensionalidade

- Reduzir a dimensionalidade elimina características irrelevantes
- Pode reduzir o ruído
- Existem técnicas que reduzem a dimensionalidade de um conjunto de dados criando novos atributos que sejam uma combinação dos atributos antigos

Referências

Orange - https://orange.biolab.si/

Tutorial sobre Orange: https://orange3.readthedocs.io/

Mais documentação sobre o Orange: https://docs.biolab.si/3/visual-programming/

Dados da aula: www.uez.com.br/ucs

Repositório GitHub: https://github.com/daniuez/courses