data train

12 de dezembro de 2019

Descoberta de padrões úteis por meio da química analítica e do aprendizado de máquina

Nattane Luíza da Costa
nattane.luiza@ifgoiano.edu.br
@profa.nattaneluiza

data train

Quem sou eu

- Tecnóloga em Redes de Computadores
- Mestra em Ciência da Computação
- Doutoranda em Ciência da Computação
- Professora do Instituto Federal Goiano, Campus Urutaí

- Primeiro contato com aprendizado de máquina/mineração de dados/*data science*
  - Dados corporativos x dados gerados para estudos específicos

data train

Química Analítica e Análise de Dados

Análise de amostras físicas em laboratório $\rightarrow$
Análise de dados

data train

Química Analítica e Análise de Dados

Análise de amostras físicas em laboratório → Análise de dados

Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools [Brereton et al., 2017]

- **1949:** Uso de ANOVA para estudo sobre borrachas sintéticas
- **1960-1961:** uso de métodos multivariados do ponto de vista teórico
- **1968:** Lançamento do periódico *Pattern Recognition*

data train

Química Analítica e a análise de dados

- **1972:** Uso do termo "Quimiometria"
- **1980:** diferenciação entre quimiometria(específico) e reconhecimento de padrões (geral)
- *Século XXI:* avanço da computação e dos equipamentos analíticos permitiram análise de bases de dados grandes e complexos
  - 1970 $\rightarrow$ 30 amostras e 10 variáveis
  - Atual $\rightarrow$ número de amostras e variáveis pode variar entre 100 e 10 000 (a depender de amostras, reagentes, e tipo de análise química)

Exemplos de problemas

- Vinhos
  - *Terrior* $\rightarrow$ implica no preço (Syrah - Argentina x Chile [da Costa et al., 2018], Cabernet Sauvignon - Chile x Brasil [da Costa et al., 2016]
  - Tipo de uva utilizada na produção do vinho
  - Carménère x Merlot (Chile) [da Costa et al., 2019]

Mundovino

Sindicância busca indícios de fraude em rótulos de vinhos de Bordeaux

Marcas de vinho estariam colocando nomes de château que não condizem com o vinho da garrafa

Funcionários da Direccte, órgão de defesa do consumidor da França, investigam se algumas marcas de vinho de Bordeaux estão enganando os consumidores usando o "nome de marca" de um château quando o vinho na garrafa não foi produzido somente pela propriedade indicada ou as uvas sequer vieram do local apontado.

Da redação
Publicado em 3 de Setembro de 2019 às 16:00

Figure 1: https://revistaadega.uol.com.br/artigo/sindicancia-busca-indicios-de-fraude-em-rotulos-de-vinhos-de-bordeaux$_1$1919.*html*

Fraude no vinho: milhões de litros de rosé da Espanha eram vendidos como se fossem da França

Volume seria equivalente a dez milhões de garrafa da bebida

AFP
09/07/2018 - 11:12 / Atualizado em 09/07/2018 - 17:59

Figure 2:
https://oglobo.globo.com/economia/fraude-no-vinho-milhoes-de-litros-de-rose-da-espanha-eram-vendidos-como-se-fossem-da-franca-22867292

Exemplos de problemas

- Alimentos orgânicos x convencionais
  [de Lima and Barbosa, 2019]
  - Implica na saúde e no preço
  - Não possui selos de certificação como os vinhos
  - Conhecimentos sobre a concentração de elementos

data train

Outros exemplos

- Azeite de oliva, mel, óleos essenciais, arroz, suco, café, leite, ovos, carnes ...

data train

Outros exemplos

- Azeite de oliva, mel, óleos essenciais, arroz, suco, café, leite, ovos, carnes ...
- Biomarcadores para o diagnóstico de doenças (câncer colorretal [Nakajima et al., 2018], câncer de bexiga [Kouznetsova et al., 2019], entre outros)

data train

Outros exemplos

- Azeite de oliva, mel, óleos essenciais, arroz, suco, café, leite, ovos, carnes ...
- Biomarcadores para o diagnóstico de doenças (câncer colorretal [Nakajima et al., 2018], câncer de bexiga [Kouznetsova et al., 2019], entre outros)
- Ecstasy [Maione et al., 2018]

data train

Outros exemplos

- Azeite de oliva, mel, óleos essenciais, arroz, suco, café, leite, ovos, carnes ...
- Biomarcadores para o diagnóstico de doenças (câncer colorretal [Nakajima et al., 2018], câncer de bexiga [Kouznetsova et al., 2019], entre outros)
- Ecstasy [Maione et al., 2018]
- Dano celular por bisfenóis, clorofenóis, parabenos e benzofenonas [Rocha et al., 2018]

Os desreguladores hormonais presentes em plásticos e cosméticos e que foram encontrados em crianças brasileiras

Alessandra Goes Alves
De São Paulo para a BBC News Brasil

23 setembro 2018

Conhecidas como "desreguladores endócrinos", algumas destas substâncias podem interferir na síntese e ação de hormônios, responsáveis por funções como metabolismo, crescimento, desenvolvimento, reprodução, sono e estado de ânimo.

A fim de verificar o nível de exposição de crianças brasileiras a essas substâncias, um grupo de pesquisadores analisou a concentração de 65 desreguladores endócrinos em urinas de 300 crianças das cinco regiões do país, com idades entre 6 e 14 anos.

Figure 4: https://www.bbc.com/portuguese/geral-45555524

No estudo do pesquisador da USP, a mineração dos dados – metodologia que verifica padrões em uma grande quantidade de informações – apontou que urinas com altas concentrações de diferentes desreguladores também apresentaram maiores concentrações de 8OHDG, molécula que, em grande quantidades, indica a ocorrência de lesão nas células.

"Este processo danifica o DNA e favorece a ocorrência de obesidade, infertilidade, doenças cardiovasculares e alguns tipos de câncer, principalmente aqueles relacionados ao sistema reprodutor", explica Rocha.

Figure 5: https://www.bbc.com/portuguese/geral-45555524

data train

Como fazer a análise?

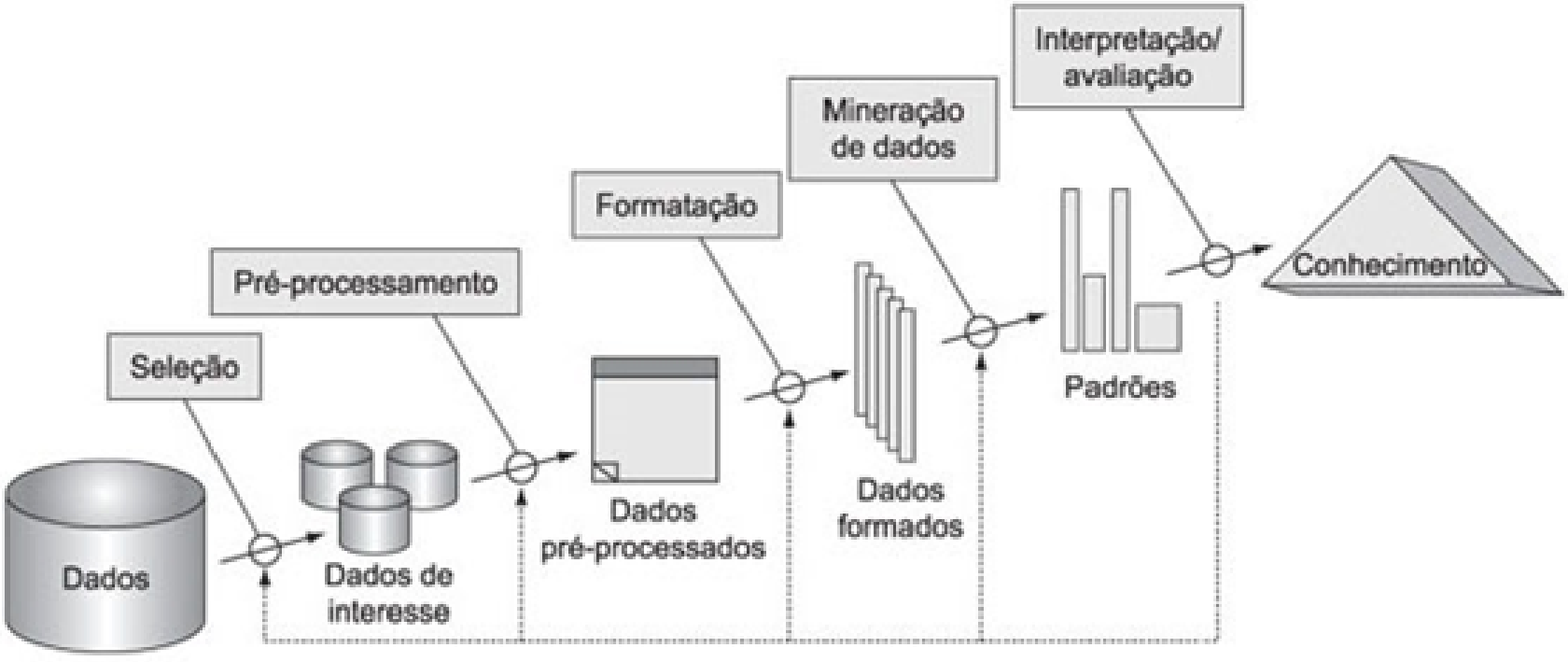

Figure 6: Processo de KDD

data train

Como fazer a análise?

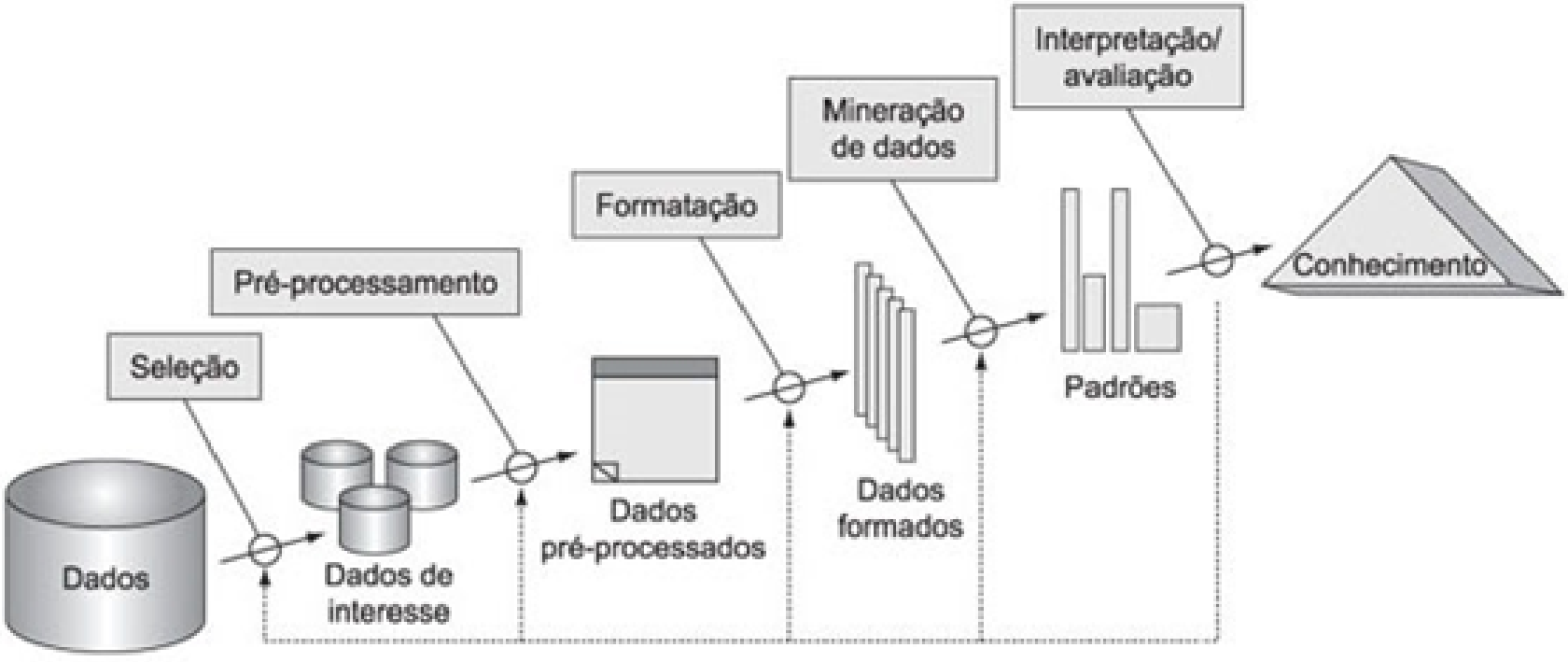

Figure 6: Processo de KDD

FOCO NA SELEÇÃO DE VARIÁVEIS

data train

Complicações

- Dados desbalanceados
- Número de amostras pequenos
- Valores ausentes
- Diferentes pessoas preencherem a planilha de dados
- Comunicação entre o especialista no assunto (dono dos dados) e quem faz a análise de dados

data train

Pacotes R

- caret

data train

Pacotes R

- caret
- FSelector

Pacotes R

- caret
- FSelector
- unbalanced, DMwR

data train

Pacotes R

- caret
- FSelector
- unbalanced, DMwR
- ggplot2

Pacotes R

- caret
- FSelector
- unbalanced, DMwR
- ggplot2
- NbClust

data train

Pacotes R

- caret
- FSelector
- unbalanced, DMwR
- ggplot2
- NbClust

https://rstudio.com/wp-content/uploads/2019/01/Cheatsheets$_2$019.*pdf*

Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., Roger, J. M., Walczak, B., and Tauler, R. (2017).
Chemometrics in analytical chemistry—part i: history, experimental design and data analysis tools.
Analytical and bioanalytical chemistry, 409(25):5891–5899.

da Costa, N. L., Castro, I. A., and Barbosa, R. (2016).
Classification of cabernet sauvignon from two different countries in south america by chemical compounds and support vector machines.
Applied Artificial Intelligence, 30(7):679–689.

da Costa, N. L., Llobodanin, L. A. G., Castro, I. A., and Barbosa, R. (2019).
The use of data mining to classify carménère and merlot wines from chile.
Expert Systems, 36(2):e12361.

da Costa, N. L., Llobodanin, L. A. G., de Lima, M. D., Castro, I. A., and Barbosa, R. (2018).
Geographical recognition of syrah wines by combining feature selection with extreme learning machine.
Measurement, 120:92–99.

de Lima, M. D. and Barbosa, R. (2019).
Methods of authentication of food grown in organic and conventional systems using chemometrics and data mining algorithms: a review.
Food Analytical Methods, 12(4):887–901.

Kouznetsova, V. L., Kim, E., Romm, E. L., Zhu, A., and Tsigelny, I. F. (2019).
Recognition of early and late stages of bladder cancer using metabolites and machine learning.
Metabolomics, 15(7):94.

Maione, C., de Oliveira Souza, V. C., Togni, L. R., da Costa, J. L., Campiglia, A. D., Barbosa, F., and Barbosa, R. M. (2018).
Establishing chemical profiling for ecstasy tablets based on trace element levels and support vector machine.
Neural Computing and Applications, 30(3):947–955.

Nakajima, T., Katsumata, K., Kuwabara, H., Soya, R., Enomoto, M., Ishizaki, T., Tsuchida, A., Mori, M., Hiwatari, K., Soga, T., et al. (2018).
Urinary polyamine biomarker panels with machine-learning differentiated colorectal cancers, benign disease, and healthy controls.
International journal of molecular sciences, 19(3):756.

Rocha, B. A., Asimakopoulos, A. G., Honda, M., da Costa, N. L., Barbosa, R. M., Barbosa Jr, F., and Kannan, K. (2018).

Advanced data mining approaches in the assessment of urinary concentrations of bisphenols, chlorophenols, parabens and benzophenones in brazilian children and their association to dna damage.
Environment international, 116:269–277.

data train

12 de dezembro de 2019

Descoberta de padrões úteis por meio da química analítica e do aprendizado de máquina

Nattane Luíza da Costa
nattane.luiza@ifgoiano.edu.br
@profa.nattaneluiza