CODING IN BRAZIL

Ciência de dados

Wellington Silva

Brasil
2021

Sumário

Lista de ilustrações

1 Introdução

Origem

A mineração de dados surgiu como área de pesquisa e aplicação independente em meados da década de 1990. Entretanto, as suas origens na matemática, estatística e computação são muito anteriores a esse período.

Objetivo

Preparação e análise das grandes massas de dados, tendo a finalidade de encontrar o conhecimento. Portanto, para cumprir tal finalidade, reuni áreas distintas, como estatística; matemática; engenharia; inteligência artificial; banco de dados; sistemas de informação; visualização; antropologia; e o especialista do domínio dos dados, que se complementam e formam a área de ciência de dados.

KDD

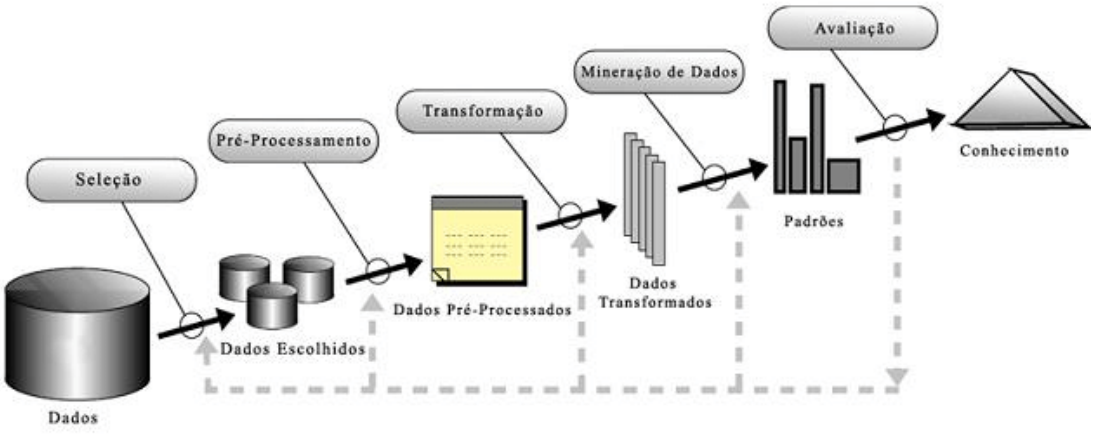

Figura 1 – Ilustração das etapas da extração de conhecimento

1. Dados: conjunto de dados organizados de forma *qualitativa* ou *quantitativa* sobre determinado tema, no qual possibilidade a extração de informação que pode resultar em conhecimento.

2. Pré-processamento dos dados: Selecionar os dados de acordo com a demanda do estudo, descartando assim dados irrelevantes, a fim de tornar a análise dos eficiente e eficaz. As etapas são distribuídas:

   - limpeza: remoção de ruídos de dados inconsistentes e ausentes;
   - integração:combinação dos dados de diferentes fontes;

- seleção: escolha de dados relevantes à análise; e
- transformação: consolidação dos dados em formato apropriado.

3. Mineração de dados: Utilização de métricas e medidas estatísticas, para representar o conjunto de dados e a sua distribuição. Tais medidas são análise descritiva, agrupamento, predição, associação e detecção de anomalias.

4. underlineAvaliação: Identificar os padrões obtidos pela representação do conhecimento são válidos, ou seja, representativo.

1.1 Ferramentas

- Weka (www.cs.waikato.ac.nz/ml/weka)
- Matlab
- R Studio (www.r-project.org)
  - Bioconductor (www.bioconductor.org)
- Wolfram Mathematica (www.wolfram.com/mathematica)
- RapidMiner (rapidminer.com)
- SAS (sas.com)
- SSPS by IBM (www-01.ibm.com/software/analytics/spss)
- Orange (orange.biolab.si)
- Mahout by Apache (mahout.apache.org)
- ELKI (elki.dbs.ifi.lmu.de): aprendizagem não supervisionado
- LIBSVM (www.csie.ntu.edu.tw/ cjlin/libsvm)

Visualização de dados

- Tableau: visualização dinâmica dos dados e dashboard personalizado
- PowerBI: integração com os serviços Microsoft
- Data Studio: integração com os serviços Google
- Qlik: licitação

SGBD

- PostgreSQL [1]
- ORACLE [2]
- SQL Server [3]
- DB2 [4]

Apache

Airflow [5]: plataforma de gerenciamento de fluxo de trabalho de código aberto

Integração dos dados

- Kafka[6]
- Spark[7]
- NiFi[8]

Microsserviços

- Docker
- Docker Compose
- Kubernet

1.2 Material complementar

Livros:

- Introdução a mineração de dados por Ferrari e Silva (2017)
- Data Science para Negócios por Fawcett e Provost (2018)
- Python para análise de dados por McKinney (2019)
- Introdução à Ciência de Dados Fundamentos e Aplicações [9]

[1] www.postgresql.org
[2] www.oracle.com
[3] www.microsoft.com/en-us/server-cloud/products/sql-server
[4] www-01.ibm.com/software/data/db2
[5] https://airflow.apache.org/
[6] https://kafka.apache.org/
[7] https://spark.apache.org/
[8] https://nifi.apache.org/
[9] https://www.ime.usp.br/ jmsinger/MAE5755/cdados2019ago06.pdf

Cursos:

- ML4all - UFPR [10]

Blog:

- DIKW by Towards Data Science [11]
- Curso R [12]
- Tests as linear by Lindeloev [13]
- JTemporal [14]

Base de dados:

- UCI Machine Learning Repository [15]
- KDnuggets [16]
- Governo Brasileiro [17]
  - Brasil IO [18]
  - Gasto de parlamentar [19]
- Governo Americano [20]
- Governo do Inglês [21]
- PyData Book [22]

[10] http://cursos.leg.ufpr.br/ML4all/1parte/
[11] https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb
[12] https://blog.curso-r.com/
[13] https://lindeloev.github.io/tests-as-linear/
[14] https://jtemporal.com/
[15] http://archive.ics.uci.edu/ml/index.php
[16] https://www.kdnuggets.com/datasets/index.html
[17] https://dados.gov.br/
[18] https://brasil.io/
[19] https://serenata.ai/
[20] https://www.data.gov/
[21] https://data.gov.uk/
[22] https://github.com/wesm/pydata-book

2 Python

Editores

- Vim
- Atom
- Sublime-text
- VSCode
- Spyder3
- PyDev
- PyCharm da JetBrains
- Komodi IDE
- Kite [23]

IPython é um interpretador interativo para várias linguagens de programação, mas especialmente focado em Python.

- JupyterLab
- Jupyter Notebook
- Colab Notebooks
- Kaggle [24]

Gerenciamento de pacote

Python Package Index - PIP [25] Sistema de gerenciamento de pacotes padrão de facto usado para instalar e gerenciar pacotes de software escritos em Python. Muitos pacotes podem ser encontrados na fonte padrão para pacotes e suas dependências.

Anaconda [26] Distribuição gratuita e de código aberto das linguagens de programação Python e R para computação científica, que visa simplificar o gerenciamento e a implantação de pacote.

Conda [27] Gerenciador de pacotes e sistema de gerenciamento de ambiente de código aberto, plataforma cruzada e independente de linguagem.

[23] https://www.kite.com/
[24] https://www.kaggle.com/
[25] https://pypi.org/project/pip
[26] https://www.anaconda.com
[27] https://docs.conda.io/en/latest

Essencial

NumPy é um pacote para a linguagem Python que suporta arrays e matrizes multidimensionais, possuindo uma larga coleção de funções matemáticas para trabalhar com estas estruturas.

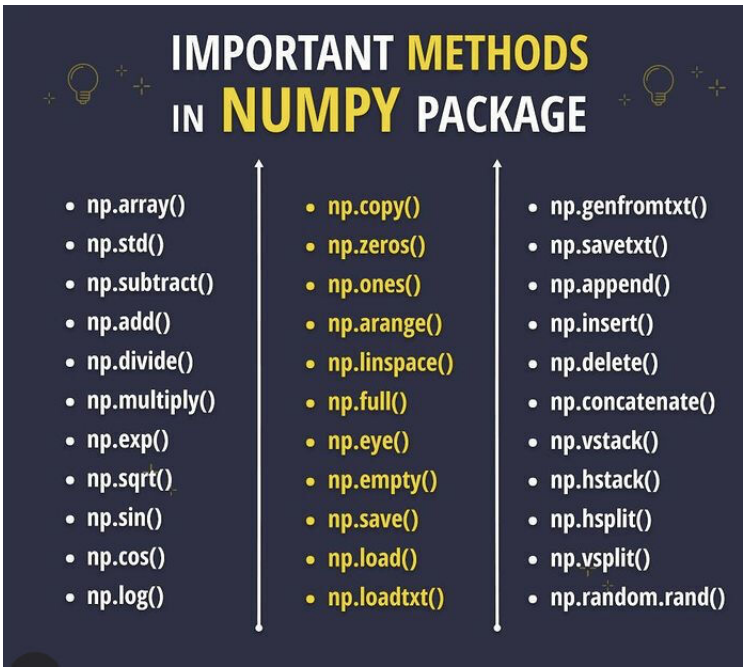

Figura 2 – NumPy

Pandas é uma biblioteca de software criada para a linguagem Python para manipulação e análise de dados. Em particular, oferece estruturas e operações para manipular tabelas numéricas e séries temporais. O nome é derivado de painel data.

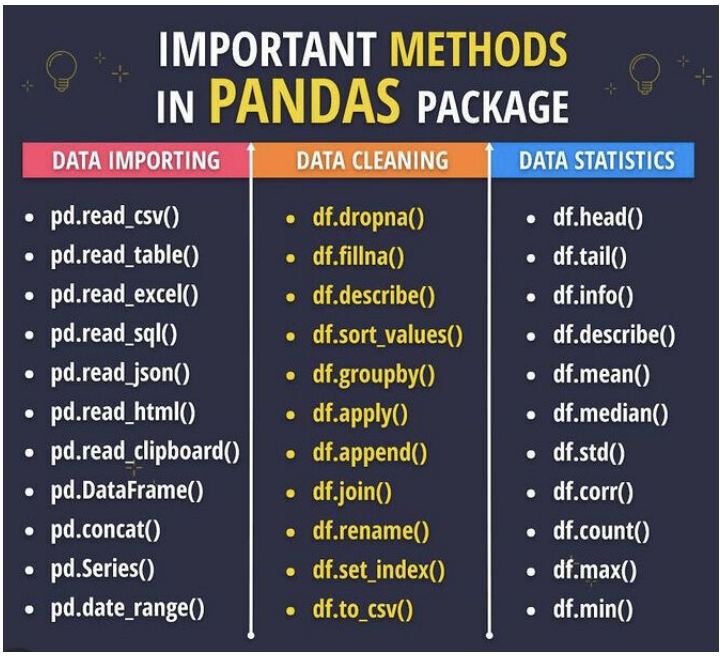

Figura 3 – Pandas

Matemática

SymPy é uma biblioteca Python para computação simbólica. Ela fornece ferramentas de álgebra computacional tanto como uma aplicação independente como, também, uma biblioteca para outras aplicações.

SciPy é uma biblioteca Open Source em linguagem Python que foi feita para matemáticos, cientistas e engenheiros. Também tem o nome de uma popular conferência de programação científica com Python.

StatsModels é um pacote Python que permite aos usuários explorar dados, estimar modelos estatísticos e executar testes estatísticos

Visualização de dados

Matplotlib [28] é uma biblioteca para geração de gráficos e visualizações de dados em geral, feita para e da linguagem de programação Python e sua extensão de matemática NumPy.

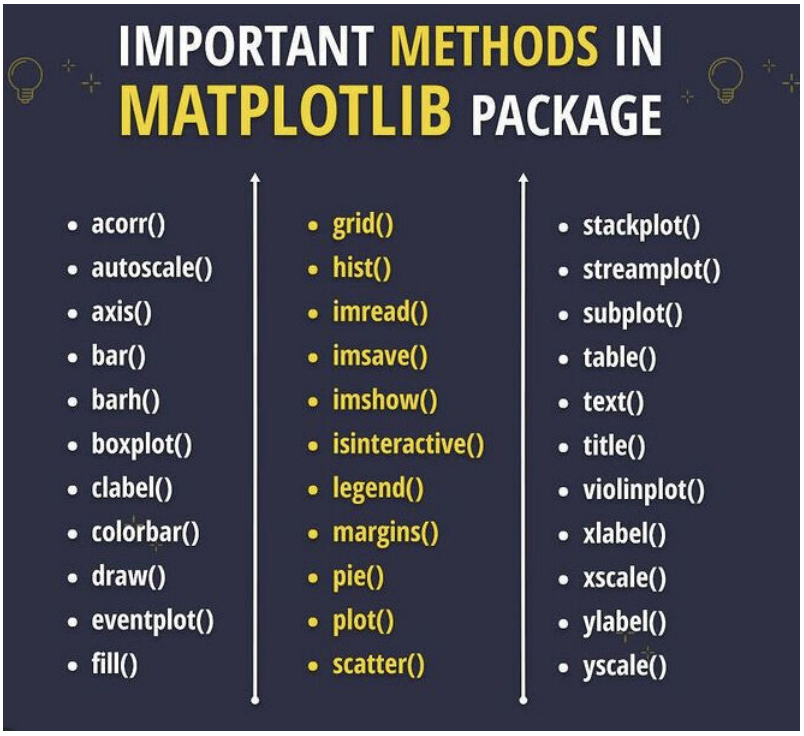

Figura 4 – Matplotlib

Seaborn [29]

Web Scraping

- LXML [30]

[28] https://matplotlib.org/
[29] https://seaborn.pydata.org/
[30] https://lxml.de

Figura 5 – Seaborn

- HTLM5LIB [31]
- Beautiful Soup [32]

Aprendizagem de máquina

Scikit-learn [33] Biblioteca de aprendizado de máquina de código aberto para modelagem preditiva pela linguagem de programação Python.

TensorFlow [34] Biblioteca de código aberto para aprendizado de máquina aplicável a uma ampla variedade de tarefas. É um sistema para criação e treinamento de redes neurais para detectar e decifrar padrões e correlações.

Keras [35] Biblioteca de rede neural de código aberto escrita em Python. Ele é capaz de rodar em cima de TensorFlow, Microsoft Cognitive Toolkit, R, Theano, ou PlaidML. Projetado para permitir experimentação rápida com redes neurais profundas, ele se concentra em ser fácil de usar, modular e extensível Izbicki e Santos (2020)

[31] https://html5lib.readthedocs.io/en/latest
[32] https://www.crummy.com/software/BeautifulSoup/bs4/doc
[33] https://scikit-learn.org
[34] https://www.tensorflow.org
[35] https://keras.io

Yellowbrick: Machine Learning Visualization [36] biblioteca de visualização para avaliação de modelos

Processamento de imagens

Python Imaging Library é uma biblioteca da linguagem de programação Python que adiciona suporte à abertura e gravação de muitos formatos de imagem diferentes.

OpenCV é uma biblioteca multiplataforma, totalmente livre ao uso acadêmico e comercial, para o desenvolvimento de aplicativos na área de Visão computacional

Scikit-image é uma biblioteca de processamento de imagens de código aberto para a linguagem de programação Python. Inclui algoritmos para segmentação, transformações geométricas, manipulação do espaço de cores, análise, filtragem, morfologia, detecção de recursos e muito mais.

PyTorch é uma biblioteca de aprendizado de máquina de código aberto baseada na biblioteca Torch, usada para aplicativos como visão computacional e processamento de linguagem natural

[36] https://www.scikit-yb.org/

3 Resenha do livro Introdução a mineração de dados

3.1 Principais tarefas da mineração de dados

Objetivo em especificar os tipos de informação a serem obtidas por intermédio das tarefas de mineração, sendo classificada em *descritivas* e *preditivas*, respectivamente, caracterizem as propriedades gerais dos dados; e fazem inferência a partir dos dados analisados.

Análise descritiva dos dados As análises descritivas permitem uma sumarização e compreensão dos objetos da base e seus atributos.

Predição: classificação e estimação terminologia usada para se referir à construção e ao uso de um modelo para avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de dado objeto. No primeiro caso, denominamos a tarefa de classificação e, no segundo, denominamos de regressão (em estatística) ou simplesmente estimação. Sob essa perspectiva, classificação e estimação constituem os dois principais tipos de problemas de predição, sendo que a classificação é usado para predizer *valores discretos*, ao passo que a estimação é usado para predizer *valores contínuos*.

Agrupamento

Análise de Associação Existem dois aspectos centrais na mineração de regras de associação: a proposição ou *construção* eficiente das regras de associação e a quantificação da *significância* das regras propostas. Ou seja, um bom algoritmo de mineração de regras de associação precisa ser capaz de propor associações entre itens que sejam estatisticamente relevantes para o universo representado pela base de dados.

3.2 Pré-processamento de dados

O processo de preparação da base de dados:

- Limpeza de dados: Imputação de valores ausentes, remoção de ruídos e correção de inconsistências;
- Integração dos dados: Unir dados de múltiplas fontes em um único local, como um armazém de dados (data warehouse);
- Redução dos dados: Reduzir a dimensão da base de dados, por exemplo, agrupando ou eliminando atributos redundantes, ou para reduzir a quantidade de objetos da base, sumarizando os dados;
- Transformação dos dados: Padronizar e deixar os dados em um formato passível de aplicação das diferentes técnicas de mineração;
- Discretização dos dados: Permitir que métodos que trabalham apenas com atributos nominais possam ser empregados a um conjunto maior de problemas. Também faz com que a quantidade de valores para um dado atributo (contínuo) seja reduzida.

Limpeza de dados: A baixa qualidade dos dados é um problema que afeta a maior parte das bases de dados reais. Assim, as ferramentas para a limpeza de dados atuam no sentido de imputar valores ausentes, suavizar ruídos, identificar valores discrepantes (outliers) e corrigir inconsistências.

Métodos tradicionais de imputação de valores ausentes:

- Avestruz: descarta o objeto que possui atributo ausente.
- Manual: escolher manual de forma empírica um valor a ser imputado para cada valor ausente.
- Constante: substitui todo valor ausente por uma constante.
- Hot-deck: substitui o valor ausente por um valor mais similar a ele.
- Last observation carried forward: considera que a representação é uma medida contínua, para isto ordena todos os atributos, substituindo os valores ausentes por seus antecessores.
- Medidas centrais: usar a média ou a moda para substituir valores ausentes.
- Medidas centrais para classe: usar a média ou a moda da classe para substituir valores ausentes da mesma.
- Modelo preditivos: utiliza modelo preditivos para imputar os valores ausentes. Nesse caso, o atributo com valores ausentes é utilizado como atributo dependente, ao passo que os outros atributos são usados como independentes para se criar o modelo preditivo. Portanto, o modelo preditivo é usado para estimar os valores ausentes.

Métodos de Redução de dados

- Redução de dimensionalidade: seleção de atributos
- Compressão de atributos: também efetua uma redução da dimensionalidade, mas empregando algoritmos de codificação ou transformação de dados (atributos), em vez de seleção. Exemplo é a Análise de Componentes Principais (Principal Component Analysis – PCA), que é um procedimento estatístico que converte um conjunto de objetos com atributos possivelmente correlacionados em um conjunto de objetos com atributos linearmente descorrelacionados, chamados de componentes principais. O número de componentes principais é menor ou igual ao número de atributos da base, e a transformação é definida de forma que o primeiro componente principal possua a maior variância (ou seja, represente a maior variabilidade dos dados), o segundo componente principal
- Redução de número de dados: realiza um corte temporal das instância, podendo ser combinada com a redução de dimensionalidade.

- Discretização: os valores de atributos são substituídos por intervalos ou níveis conceituais mais elevados, reduzindo a quantidade final de atributos.

Transformação dos dados
Padronização: escala e unidades em bases compatíveis.
Normalização

- Máximo pelo minímo: A normalização Max-Min realiza uma transformação linear nos dados originais. Assuma que max_a e min_a são,respectivamente, os valores máximo e mínimo de determinado atributo a. A normalização max-min mapeia um valor a em um valor a' no domínio $[novo_min'_a, novo_max'_a]$, de acordo com a Equação abaixo. A aplicação mais frequente dessa normalização é colocar todos os atributos de uma base de dados sob um mesmo intervalo de valores, por exemplo no intervalo $[0,1]$.

$$a' = \frac{a - min_a}{max_a - min_a} \tag{1}$$

- Escore-Z (Escore Padronizado): Útil quando se desconhece a amplitude dos dados ou há outliers, faz parte das medidas de posição relativa

$$a' = \frac{a - \bar{a}}{\delta_a} \tag{2}$$

- Escalonamento decimal: Estabelecido pelo escalonamento decimal move a casa decimal dos valores do atributo a. O número de casas decimais movidas depende do valor máximo absoluto do atributo a. A Equação abaixo, na qual j é o menor inteiro tal que $max(|a'|) < 1$, ilustra o cálculo do valor normalizado.

$$a' = \frac{a}{10^j} \tag{3}$$

- Range interquatil: Participa das medidas de posição relativa.

$$IQR = Q_3 - Q_1 \tag{4}$$

- Trivial:

$$a' = \frac{a}{max_a} \tag{5}$$

3.3 Análise descritiva de dados

O processo de análise descritiva de dados, incluindo distribuições de frequência, técnicas de visualização de dados e medidas resumo

Descrever e encontrar o que há nos dados. Ao passo que no futuro pode ser implementar algoritmos de mineração que buscam conclusões que extrapolam os dados e permitem inferir

predições. Portanto, analise descritiva descreve as características dos dado e a mineração geralmente usada em análise mais abrangentes visando a predição. Entretanto, precisa ficar atento com falsas correlações e predições dos dados.

Análise descritiva permite descrever a distribuição e a correlação dos atributos, utilzando medidas estatísticas, como distribuição de frequência, tendência central e visualização gráfica, sendo para atributos univariada e para bivariada relações entre atributos.

Processo

Distribução de frequência

Técnica de visualização

- Histograma
- Polígonos de frequências
- Ogiva
- Gráfico de Pareto
- Gráfico de setores
- Gráfico de dispersão - scatterplots

Medidas de tendência central, variação e associação

Medida de centro	Definição	Existência	Considera todos os valores?	Afetada por valores extremos?	Vantagens e desvantagens
Média	$\Sigma x/N$	Sempre	Sim	Sim	Mais comum
Mediana	Valor do meio	Sempre	Sim	Não	Quando há valores extremos
Ponto médio	(maior+menor)/2	Sempre	Não	Sim	Sensível a extremos
Moda	Valor mais frequente	Pode não existir ou pode haver múltiplos	Não	Não	Dados nominais

Figura 6 – Tendência Central

- Moda
- Mediana
- Ponto Médio

$$\frac{maior - menor}{2} \tag{6}$$

- Média amostral

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

- Média populacional

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

- Média de distribuição de frequências

$$\bar{x} = \frac{\sum_{i=n}^{n} f_i * x_i}{\sum_{i=n}^{n} f_i} \tag{9}$$

- Média ponderada

$$\bar{x} = \frac{\sum_{i=n}^{n} (w_i * x_i)}{\sum_{i=n}^{n} w_i} \tag{10}$$

- Média geométrica

$$\bar{x} = (\Pi_{i=1}^{n} x_i)^{\frac{1}{n}} \tag{11}$$

- Média harmônica

$$\bar{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{12}$$

- Medidas de dispersão

- Amplitude

$$amplitude = maior - menor \tag{13}$$

- Desvio Padrão

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \sigma &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned} \tag{14}$$

- Coeficiente de variação (CV)

$$CV = \frac{\sigma}{\mu} * 100\% \tag{15}$$

- Medidas de forma: a assimetria (Skewness) distribuição dos dados, pode ser nula(No Skewness) conhecido pela nomeclatura de curva sino, podendo receber descolamento positivo da assimetria (Positive Skewness) ou negativa (Negative Skewness). Assimetria é calculada da seguinte forma:

$$\gamma = \frac{E(x - \bar{x})^3}{\sigma^3} \tag{16}$$

- Curtose Kurtosis é uma medida de dispersão que caracteriza o pico ou achatamento da curva da função de distribuição normal.

$$\beta = \frac{E(x - \bar{x})^4}{\sigma^4} - 3 \tag{17}$$

Tabela 1 – Coeficiente de correlação de Pearson

Size of Correlation	Interpretation
$\mid 0.90 - 1.00 \mid$	Very high positive (negative) correlation
$\mid 0.70 - 0.90 \mid$	High positive (negative) correlation
$\mid 0.50 - 0.70 \mid$	Moderate positive (negative) correlation
$\mid 0.30 - 0.50 \mid$	Low positive (negative) correlation
$\mid 0.00 - 0.30 \mid$	Negligible correlation

- Medidas de posição relativa
- Quartils e boxplot
  - Medida de associação
  - Covariância

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x-\bar{x})(y-\bar{y}) \tag{18}$$

- Coeficiente de correlação de Pearson: mede a dependência linear entre os atributos de forma linear [37].

$$\rho(x,y) = \frac{cov(x,y)}{\sigma(x) * \sigma(y)} \tag{19}$$

Visualização dos dados

- Medidas de resumo
- Medidas de tendência central
- Medida de dispersão
- Medida de forma distribuição

3.4 Análise de grupos

Grupos naturais descrito por Carmichael que grupos são aqueles que satisfazem duas condições particulares:

1. Existência de regiões contínuas do espaço, relativamente densamente populadas por objetos;
2. Tais regiões estão rodeadas por regiões relativamente vazias.

3.4.1 Medidas de similaridade

Matriz de confusão (contingência)

[37] https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Tabela 2 – Medidas de dissimilaridade para variáveis contínuas

Medida	Fórmula
S1: Coeficiente de Matching	$S_{ij} = \frac{a+d}{a+b+c+d}$
S2: Coeficiente de Jaccard	$S_{ij} = \frac{a}{a+b+c}$
S3: Rogers & Tanimoto	$S_{ij} = \frac{a+d}{a+2(b+c)+d}$
S4: Sokal & Sneath	$S_{ij} = \frac{a}{a+2(b+c)}$
S5: Gower & Legendre	$S_{ij} = \frac{a+d}{a+0.5(b+c)+d}$
S6: Gower & Legendre 2	$S_{ij} = \frac{a}{a+0.5(b+c)}$

Dados binários

Distância Hamming

3.4.2 Medidas de dissimilaridade para variáveis contínuas

Medidade distância

Família de distância Minkowski

Distância de Canberra

Medidas tipos de correlação

Correlação de Pearson $[-1,1]$

Medidda do Cosseno $[-1,1]$

3.4.3 Métodos de agrupamento

- Hierárquicos
- Particionais

Avaliação

- Compactação
- Separação

Medidas internas, seguindo o índice de (p. 241):

- Dunn (DUNN, 1973)
- Davies-Bouldi [0; infinito] (DAVIES; BOULDIN, 1979)

- Bezdek-Pal (BEZDEK; PAL, 1998)
- Silhueta (ROUSSEEUW, 1987)

Medidas externas

- Entropia: define homogeneidade dos grupos encontrados. Portanto, o valor de baixa entropia indica mais homogeneidade
- Pureza
- índice FBCubed (AMIGÓ et al., 2009)

Referências

AMIGÓ, E. et al. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, Springer, v. 12, n. 4, p. 461–486, 2009.

BEZDEK, J. C.; PAL, N. R. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 28, n. 3, p. 301–315, 1998. Disponível em: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.850.9929&rep=rep1&type=pdf>.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Taylor & Francis, 1973.

FAWCETT, T.; PROVOST, F. *Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. [S.l.]: Alta Books Editora, 2018.

FERRARI, D. G.; SILVA, L. N. D. C. *Introdução a mineração de dados*. [S.l.]: Saraiva Educação SA, 2017.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, 2020. Disponível em: <http://www.rizbicki.ufscar.br/ame/>.

MCKINNEY, W. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. [S.l.]: Novatec Editora, 2019.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.