

# UNIDADE 1 - Paradigmas de Aprendizado Não-Supervisionado e Semi-Supervisionado

Marta Noronha

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
CURSO: CIÊNCIA DE DADOS  
APRENDIZADO DE MÁQUINA II



# SUMÁRIO

- 1.1 Definições básicas sobre agrupamento
- 1.2 Tipos de variáveis
- 1.3 Aprendizado Não Supervisionado (Agrupamento)
- 1.4 Aprendizado Semi-Supervisionado (Propagação de Rótulos)

# Origem

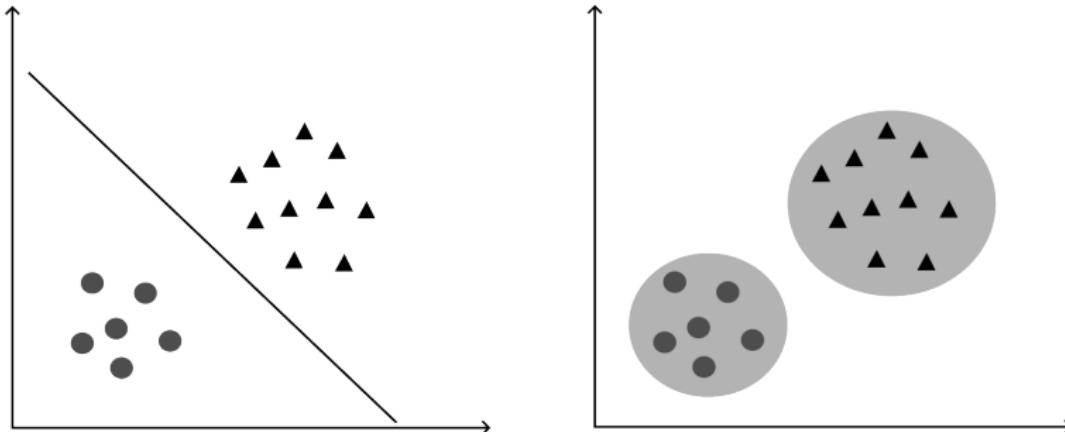
- Antropologia: *Quantitative Expression of Cultural Relationships* [Driver e Kroeber 1932]
  - Explora como as culturas podem ser quantificadas e comparadas.
- Psicologia: *A technique for measuring like-mindedness* [Zubin 1938]
  - Apresenta métodos para medir a semelhança de pensamentos entre indivíduos.
- Os trabalhos influenciaram no desenvolvimento de técnicas de agrupamento em diversas áreas, sendo uma delas na biologia onde as técnicas foram adotadas para a classificação e organização de espécies.
  - Agrupamento Hierárquico: Construção de árvores filogenéticas mostrando as relações evolutivas entre diferentes espécies.
  - Clusterização de Genes: Análise de expressão gênica para identificar grupos de genes que são co-expressos sob certas condições, ajudando a entender funções biológicas e processos celulares.

## Classificação vs. Clusterização

## Classificação

vs

## Clusterização (Agrupamento)



# Classificação vs. Clusterização

	<b>Classificação</b>	<b>Clusterização</b>
<b>Objetivo</b>	Atribuir rótulos de classes às instâncias	Agrupar instâncias avaliando a (dis)similaridade
<b>Propósito</b>	Predizer os rótulos de instâncias não vistas pelo modelo	Descobrir padrões ou estruturas subjacentes (ocultas) no conjunto de dados
<b>Aprendizado</b>	Supervisionado	Não supervisionado
<b>Rótulos</b>	Sim	Não
<b>Saída</b>	Instâncias com rótulos preditos	Grupos homogêneos por um critério
<b>Aplicação</b>	Determinar se um e-mail é ou não spam	Segmentação de clientes analisando suas preferências

## Requisitos típicos para clusterização

- Escalabilidade: agrupar muitos objetos (algumas centenas) pode levar ao enviesamento (um erro sistemático) dos dados. Em aplicações modernas, como análise de big data, é crucial que o algoritmo possa escalar bem conforme a quantidade de dados aumenta.
- Habilidade para tratar tipos diferentes de atributos (ordinais, categóricos, binários, numéricos). Dados reais frequentemente contêm uma mistura de diferentes tipos de atributos.
- Tratamento de dados com ruídos: algoritmos podem ser sensíveis a dados contendo outliers, ausentes, incorretos ou desconhecidos, os quais podem afetar a precisão da clusterização.

## Requisitos típicos para clusterização

- Descoberta de clusters com tamanhos arbitrários: algoritmos podem tender a encontrar clusters de formatos e densidades semelhantes.
- Quantidade de parâmetros: o algoritmo pode ser sensível aos parâmetros de entrada fornecidos pelo usuário. Quanto menor o número de parâmetros, mais fácil é o ajuste.
- Clusterização incremental e não sensível à ordem dos objetos de entrada: algoritmos podem ter dificuldade em incorporar objetos em clusters já existentes e produção de clusters diferentes dependendo da ordem de entrada dos objetos.
- Alta dimensionalidade: os algoritmos tendem a agrupar dados em poucas dimensões (2 a 3). Quanto mais dimensões existir no conjunto de dados, mais desafiador é o agrupamento.

## Requisitos típicos para clusterização

- Cluster baseado em restrição: Em algumas aplicações, como planejamento urbano, pode ser necessário impor restrições nos clusters. Como exemplos de restrições nesta área têm-se:
  - Zona de Uso do Solo: Certos clusters devem respeitar as zonas de uso do solo, como a mistura de áreas residenciais com áreas industriais.
  - Infraestrutura: Considerando a acessibilidade, um cluster de bairros deve estar a uma certa distância de serviços essenciais, como hospitais e escolas.
  - Densidade Populacional: Pode haver uma restrição de densidade mínima ou máxima para garantir que nenhum cluster tenha uma população excessivamente alta ou baixa, o que pode impactar a qualidade de vida e a infraestrutura necessária.

## Requisitos típicos para clusterização

- Interpretabilidade e usabilidade: Resultados interpretáveis por humanos são essenciais para a aplicação prática dos algoritmos de clusterização.

Fonte: [Han e Kamber 2006]

## Exemplos em segmentação de mercado

- Identificação de Padrões de Comportamento: Agrupamento de clientes com base em características semelhantes, permitindo o desenvolvimento de produtos e serviços que atendam a esses padrões específicos.
- Personalização de Ações de Marketing: Personalização de campanhas para diferentes segmentos de consumidores, resultando em maior engajamento e conversão.
- Eficiência Operacional: Otimização de operações, direcionando recursos e esforços para segmentos de mercado mais lucrativos.

designed by freepik

## Exemplos em segmentação de mercado

- Desenvolvimento de Produtos: A clusterização pode informar a necessidade de desenvolver novos produtos ou adaptar produtos existentes para melhor atender às necessidades de segmentos específicos.
- Melhoria na Retenção de Clientes: Identificação de fatores que influenciam a retenção de clientes em diferentes segmentos, por meio do desenvolvimento de estratégias de retenção mais eficazes, como programas de fidelidade personalizados ou serviços de atendimento ao cliente ajustados para as necessidades de cada segmento.



## Exemplos em segmentação de mercado

- Exploração de Novos Mercados: Descoberta de segmentos de mercado inexplorados ou emergentes, possibilitando a expansão de operações e captura de novas fontes de receita.
- Aprimoramento da Experiência do Cliente: Ajuste de interações da empresa com os clientes para melhor atender às expectativas de diferentes segmentos, contribuindo com o aumento da lealdade e satisfação do cliente, tornando-os mais predispostos a recomendar a marca.



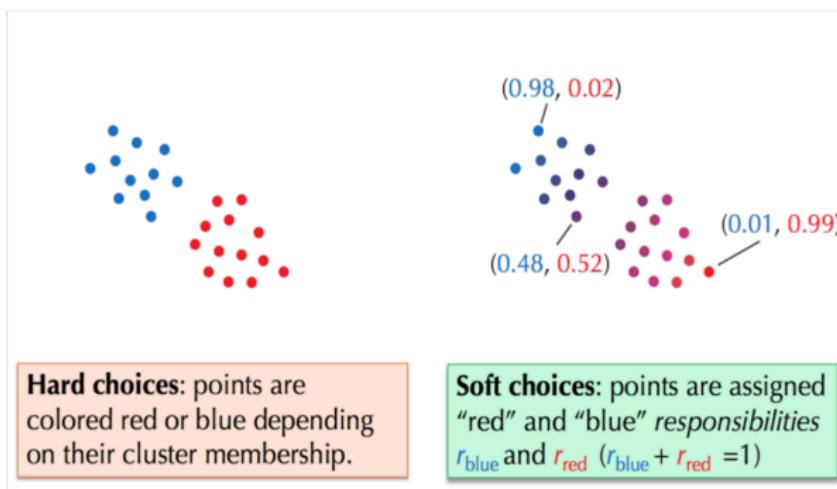
## Exemplo de categorias para segmentação

- Faixa etária.
- Renda.
- Gênero.
- Hábitos de consumo.
- Escolaridade e emprego.



## Hard e Soft Clustering

- Hard Clustering: K-Means aplicado em segmentação de clientes.
- Soft Clustering: Soft K-Means em análise genética onde um gene pode estar relacionado a múltiplas funções biológicas.



**Hard choices:** points are colored red or blue depending on their cluster membership.

**Soft choices:** points are assigned "red" and "blue" responsibilities  $r_{\text{blue}}$  and  $r_{\text{red}}$  ( $r_{\text{blue}} + r_{\text{red}} = 1$ )

FONTE: [https://www.cs.cmu.edu/~02251/recitations/recitation\\_soft\\_clustering.pdf](https://www.cs.cmu.edu/~02251/recitations/recitation_soft_clustering.pdf)

## Restrições ou permissões nos agrupamentos

- Agrupamento de particionamento estrito: divide os dados em clusters exclusivos, onde cada ponto de dados pertence exatamente a um cluster.
- Agrupamento de particionamento estrito com *outliers*: o agrupamento que permite a presença de outliers, ou seja, pontos de dados que não se ajustam bem a nenhum dos clusters.
- Sobreposição de cluster: os objetos podem pertencer a mais de um cluster simultaneamente.

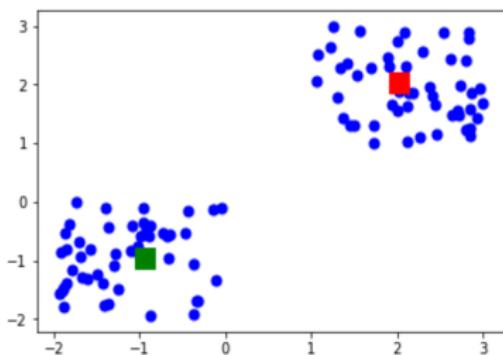
## Restrições ou permissões nos agrupamentos

- Agrupamento de subespaço: identifica clusters em subespaços dos dados, onde diferentes clusters podem ser definidos em diferentes subconjuntos de atributos.
- Agrupamento hierárquico: árvore de clusters (dendrograma) que pode ser cortada em diferentes níveis para obter diferentes números de clusters.

"O agrupamento de subespaço, por outro lado, define a qualidade do cluster com base em cada cluster individual incorporado em seu próprio subespaço." [Niu, Dy e Jordan 2013]

# Técnicas de clusterização

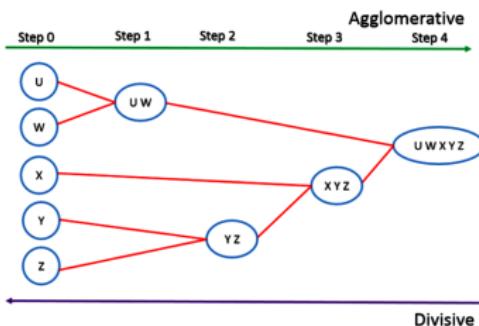
- Clusterização baseada em particionamento.
- Baseada na distância dos pontos ao centróide.
- Exemplo: K-Means.
- O número de grupos depende do valor do parâmetro K.



FONTE: <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

# Técnicas de clusterização

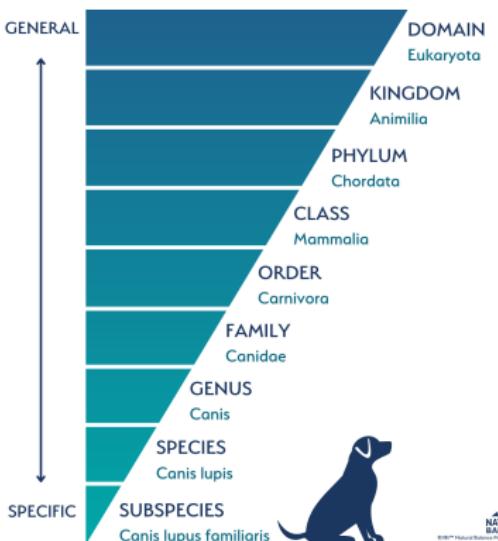
- Clusterização baseada em hierarquia.
- Matriz de distância baseada em (dis)similaridade.
- Aglomerativa ou divisiva.
- Agrupamento baseado em ligação: simples (ponto **mais próximo** de 2 clusters), completa (ponto **mais distante** de 2 clusters) ou média (**distância média** de cada ponto em um cluster com cada ponto em outro)



FONTE: <https://harshkarma1091996.medium.com/hierarchical-clustering-996745fe656b>

# Técnicas de clusterização

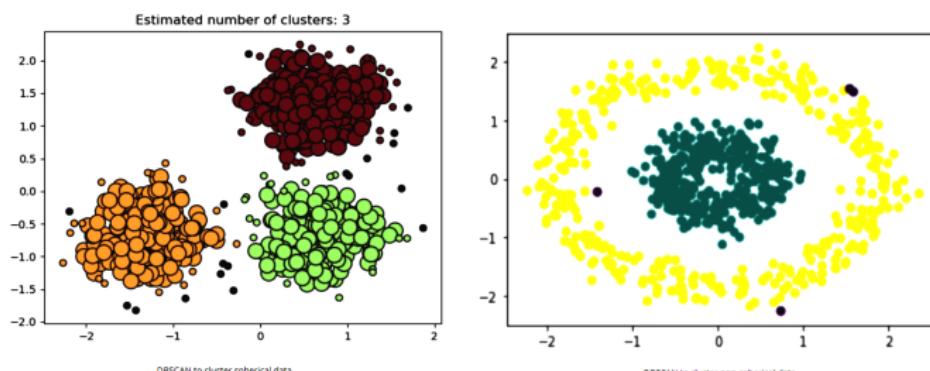
- Clusterização baseada em hierarquia.



NATURAL  
BALANCE  
©1997 Natural Balance Pet Foods, Inc.

# Técnicas de clusterização

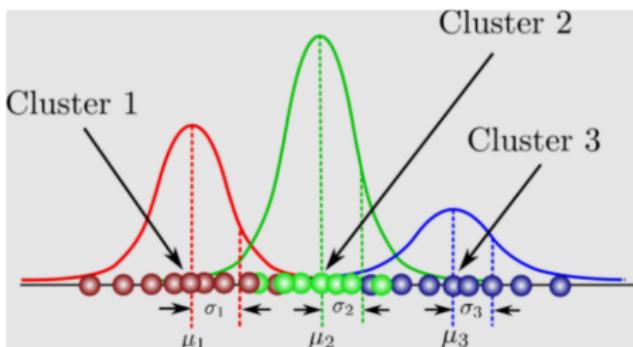
- Clusterização baseada em densidade.
- Mais robusto a ruídos do que métodos de particionamento e hierárquicos.
- Exemplo: *Density-Based Spatial Clustering Of Applications With Noise* (DBSCAN).



FONTE: <https://www.kdnuggets.com/2020/04/dbSCAN-clustering-machine-learning.html>

# Técnicas de clusterização

- Clusterização baseada em modelos.
- Otimizam o ajuste entre os dados e um modelo matemático.
- Alguns modelos podem assumir que o dado segue uma distribuição subjacente.
- Exemplo: *Expectation-Maximization*.



FONTE: <https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/>

## E agora?

### Qual método de clusterização escolher?

- Diferentes clusters atendem a diferentes propósitos.
- A técnica de clusterização deve ser avaliada de acordo com a concordância entre as instâncias de um agrupamento e a população deste agrupamento. [Hartigan 1985].

### Quais medidas de clusterização devo usar?

- Medidas de dissimilaridade (baseadas em distância)
- Medidas de similaridade (baseadas em correlação)
- A escolha depende também dos tipos de atributos (variáveis).

# Matriz de Dados

## Características da matriz de dados

- $M$  objetos com índices  $1 \leq i \leq m$ .
- $N$  variáveis com índices  $1 \leq j \leq n$ .
- Também é chamada de estrutura objeto por variável.

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{in} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & \dots & x_{mj} & \dots & x_{mn} \end{bmatrix}$$

# Matriz de Dissimilaridade

## Características da matriz de dissimilaridade

- A dissimilaridade é baseada em medidas de distância.
- Representa a dissimilaridade de  $M$  objetos com índices  $1 \leq i \leq m$ .
- A diagonal principal é igual a zero porque a distância do objeto  $i$  até ele mesmo é igual a zero.
- Também é chamada de estrutura objeto por objeto.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(m,1) & d(m,2) & \dots & 0 & \end{bmatrix}$$

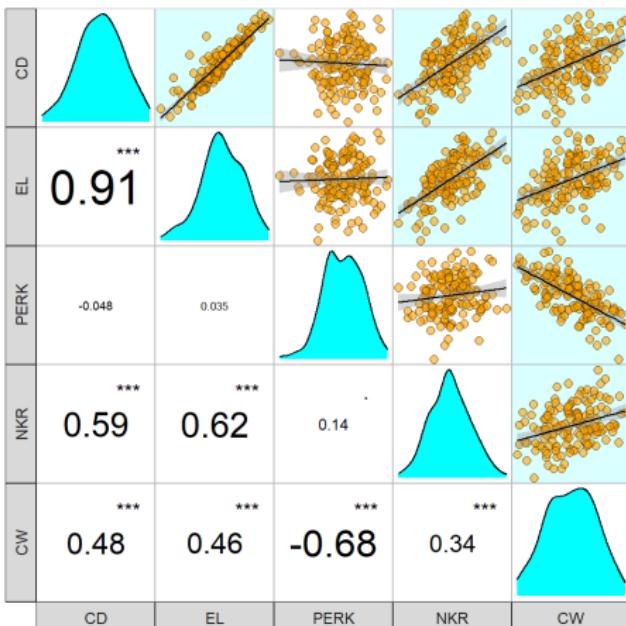
# Matriz de Similaridade

## Características da matriz de similaridade

- Mede o quanto próximos ou parecidos dois objetos são (0: menos parecidos, 1: mais parecidos).
- Pode ser baseada em medidas de correlação.
- Em dados binários, é dada por *matching*.
- A diagonal principal é igual a um porque um objeto é igual a ele mesmo.

$$\begin{bmatrix} 1 & & & \\ c(2, 1) & 1 & & \\ c(3, 1) & c(3, 2) & 1 & \\ \vdots & \vdots & \vdots & \\ c(m, 1) & c(m, 2) & \dots & 1 \end{bmatrix}$$

# Matriz de Similaridade



Fonte:

<https://tiagoolivoto.github.io/e-bookr/relations.html>

# Variáveis Numéricas

- Também chamadas de quantitativas.
- Podem ser discretas (Idade) ou contínuas (Peso).
- Suportam mais cálculos estatísticos.

## Escala

- 1 Intervalar:** O número variam dentro de um intervalo (Ex: Temperatura em Celsius ou Fahrenheit variam dentro de Kelvin). Possível definir a ordem e magnitude entre os dois valores.
- 2 Racional:** Números tem significado absoluto, ou seja, existe um zero absoluto associado a uma medida (Ex: Temperatura em Kelvin).

Conversão de	para	Fórmula
Celsius	→ Fahrenheit	$^{\circ}\text{F} = ^{\circ}\text{C} \times 1,8 + 32$
Fahrenheit	→ Celsius	$^{\circ}\text{C} = (^{\circ}\text{F} - 32) / 1,8$
Celsius	→ Kelvin	$\text{K} = ^{\circ}\text{C} + 273,15$
Kelvin	→ Celsius	$^{\circ}\text{C} = \text{K} - 273,15$

# Variáveis numéricas

Como atributos numéricos podem impactar na clusterização?

- escala;
- distribuição, e;
- correlação.

## Desafios:

- Discretização de atributos
- *Outliers*

## Usos:

- Análise estatística (médias, dispersão)
- Modelagem para predição
- Modelagem para regressão
- Clusterização

# Variáveis Categóricas

- Também chamadas de qualitativas ou simbólicas.

## Escala

- ① Ordinal
- ② Nominal



## Tratamentos de variáveis categóricas

- ① Mapeamento de variáveis ordinais (*Label encoding*)
- ② Codificação de rótulos de classe
- ③ *One-hot encoding* em variáveis nominais

## Variáveis Categóricas



# Variáveis Categóricas

## Mapeamento de variáveis ordinais

Consiste em atribuir valores numéricos que expressam a ordem dos objetos em relação ao atributo analisado.

Tamanho	Tamanho
P	1
M	2
G	3
GG	4

- Prós: Simplicidade e não aumenta a dimensionalidade.
- Contras: Não lida com atributos nominais e pode não refletir adequadamente o relacionamento de ordem de atributos categóricos ordinais.

# Variáveis Categóricas

## Codificação de rótulos de classe

- Rótulos de classe não são ordinais.
- Codificação é realizada quando os algoritmos não conseguem lidar com rótulos não inteiros.
- É uma boa prática codificar os rótulos para inteiros mesmo que o algoritmo tenha capacidade de lidar com o mesmo para evitar erros técnicos.

TAMANHO	SAÍDA
P	POUCO
M	MUITO
G	MUITO
GG	POUCO



TAMANHO	SAÍDA
P	0
M	1
G	1
GG	0

# Variáveis Categóricas

## One-hot encoding em variáveis nominais

- Não existe relação de ordem em variáveis nominais.
- Mapeamento é realizado para transformar as *strings* em inteiros.
- Aumenta a dimensionalidade do conjunto de dados.
- Introduz esparsidade no conjunto de dados.



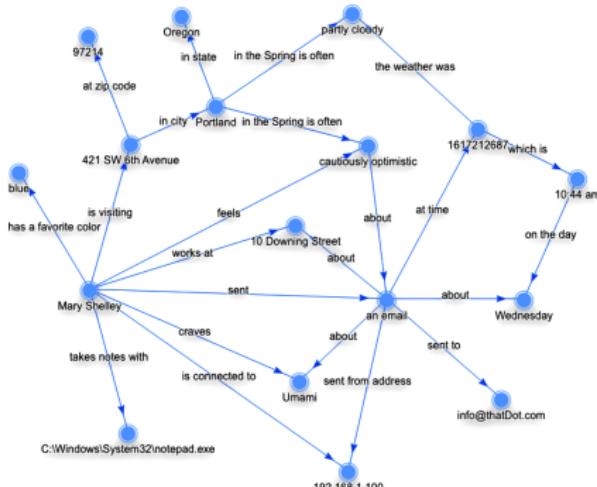
COR	COR_Vermelho	COR_Verde	COR_Azul	COR_Verde_claro
Vermelho	1	0	0	0
Verde	0	1	0	0
Azul	0	0	1	0
Verde_claro	0	0	0	1

- Prós: Simplicidade de interpretação e evita erros de ordenação.
- Contras: Esparsidade e aumento da dimensionalidade.

# Variáveis Categóricas

## Alternativas

- Ignorar.
- Contar.
- Transformar.
- Conectar.



Fonte: [www.thatdot.com/blog/what-is-categorical-data](http://www.thatdot.com/blog/what-is-categorical-data)

# Variáveis Categóricas

## Desafios

- Cardinalidade do atributo: Quantidade de valores que pode ser assumido pelo atributo.
- Viés ao atribuir valor numérico: representação enganosa pode levar a previsões imprecisas do modelo.
- Categorias não vistas em um atributo não presentes em modelos de aprendizado.

## Usos

- Segmentação de mercado
- Análise descritiva
- Análise qualitativa

# Variáveis Categóricas Binárias

Classificadas em:

- **Simétricas:** As variáveis possuem o mesmo peso (sexo biológico).
- **Assimétricas:** As variáveis possuem diferentes pesos sendo que o positivo é o único contabilizado no cálculo da dissimilaridade (Diagnóstico positivo para uma doença).

		Observado		Total
		Sim	Não	
Previsto	Sim	q (acertos)	r (alarme falsos)	q+r
	Não	s (erros)	t (rejeições corretas)	s+t
	Total	q+s	r+t	M= q+r+s+t

Tabela: Tabela de Contigência

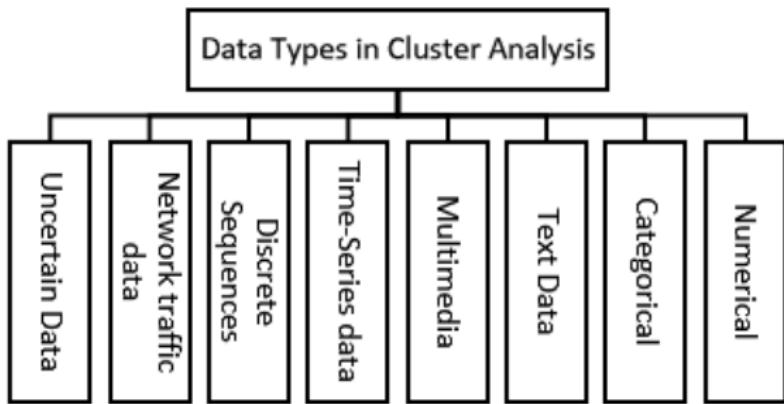
## Classifique os tipos dos atributos:

- ① Nome
- ② Idade
- ③ Peso
- ④ Patentes militares
- ⑤ Temperatura K
- ⑥ Temperatura °F
- ⑦ Estado civil
- ⑧ Regime político
- ⑨ Faixas em esporte (Karatê)
- ⑩ Raça de um cachorro
- ⑪ Salário
- ⑫ Distância
- ⑬ Nacionalidade
- ⑭ *Ranking* em esporte

# Discussão

- Quando codificar por categorias de atributo (*label encoding*) e quando usar *one-hot encoding*?

## Outros tipos de dados:



Fonte: [Mahdi, Hosny e El-henawy 2021]

# Aprendizado não supervisionado

- Não possui rótulos de saída para guiar o aprendizado.
- Redução de dimensionalidade (seleção de características, PCA e Análise fatorial).
- Descoberta de regras de associação (Mineração de padrões frequentes).
- Descoberta de padrões subjacentes (clusters) na estrutura de dados por meio de clusterização.

## Exemplos de uso:

- Segmentação;
- Descoberta de *outliers*;
- Engenharia de características (ex: imputar valores ausentes por meio da análise de grupos);
- Criação de filtros automáticos (ex: filtros de spam não supervisionados contendo mensagens semelhantes suspeitas).

# Aprendizado não supervisionado

- Seleção de características por meio de clusterização.

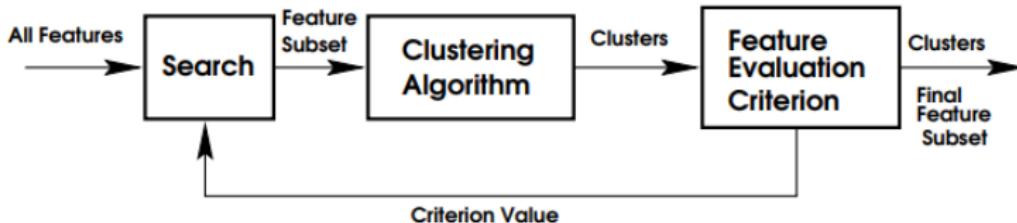


Figure 4: Wrapper approach for unsupervised learning.

Fonte: [Dy e Brodley 2004]

# Aprendizado não supervisionado

- Mineração de itens frequentes
- Algoritmo Apriori e FP-growth (*Frequent Pattern Growth*)

$$\text{Rule: } X \Rightarrow Y$$
$$\text{Support} = \frac{\text{frq}(X, Y)}{N}$$
$$\text{Confidence} = \frac{\text{frq}(X, Y)}{\text{frq}(X)}$$
$$\text{Lift} = \frac{\text{Support}}{\text{Supp}(X) \times \text{Supp}(Y)}$$

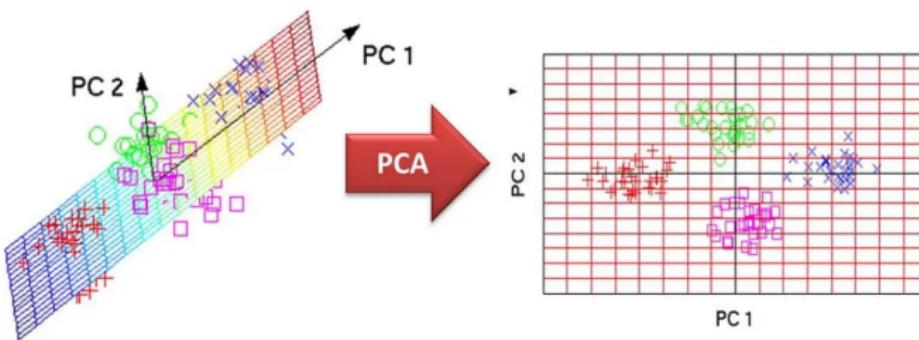


Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Fonte: [www.analyticsvidhya.com/blog/2021/10/end-to-end-introduction-to-market-basket-analysis-in-r/](http://www.analyticsvidhya.com/blog/2021/10/end-to-end-introduction-to-market-basket-analysis-in-r/)

# Aprendizado não supervisionado

- Abordagens de redução baseadas em análises de PCA.



Fonte: [Scholz 2006]

# Aprendizado não supervisionado

- Abordagens de redução baseadas em análises de PCA.

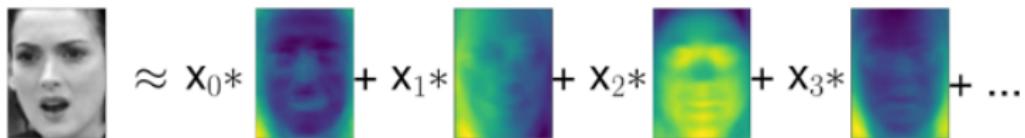
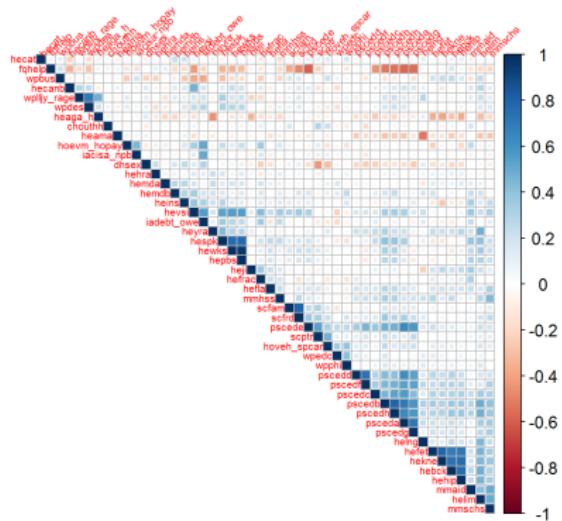


Figure 3-10. Schematic view of PCA as decomposing an image into a weighted sum of components

Fonte: [Müller e Guido 2016]

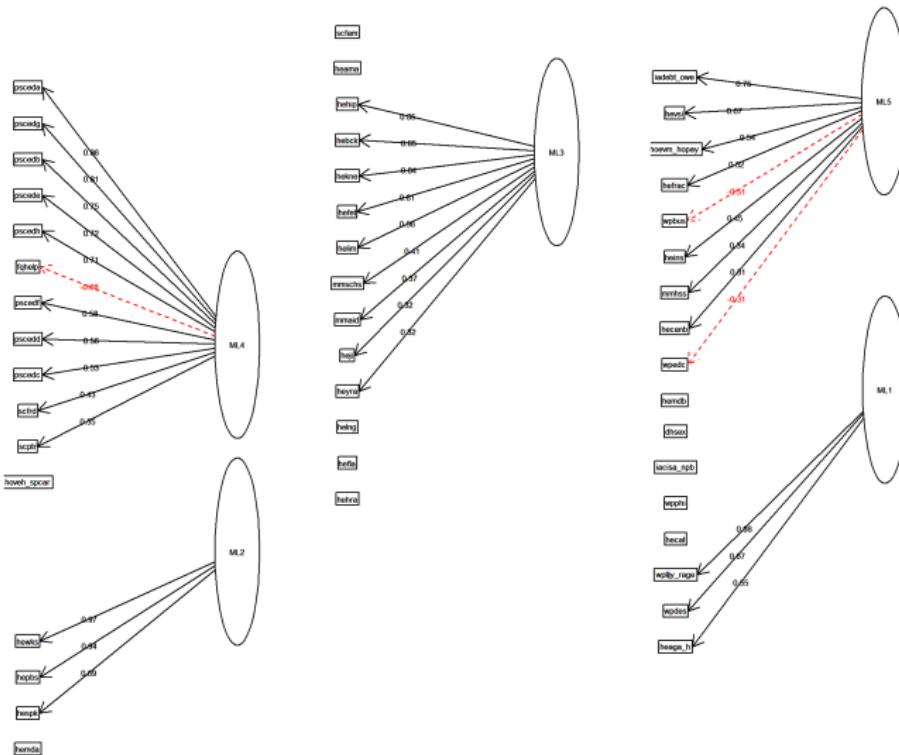
## Aprendizado não supervisionado

- Abordagens de redução baseadas em Análise fatorial (AF).
  - Usa a matriz de correlação.
  - Deve-se usar os testes Kaiser-Meyer-Olkin (KMO) e Bartlett para verificar a adequacidade da aplicação da AF.



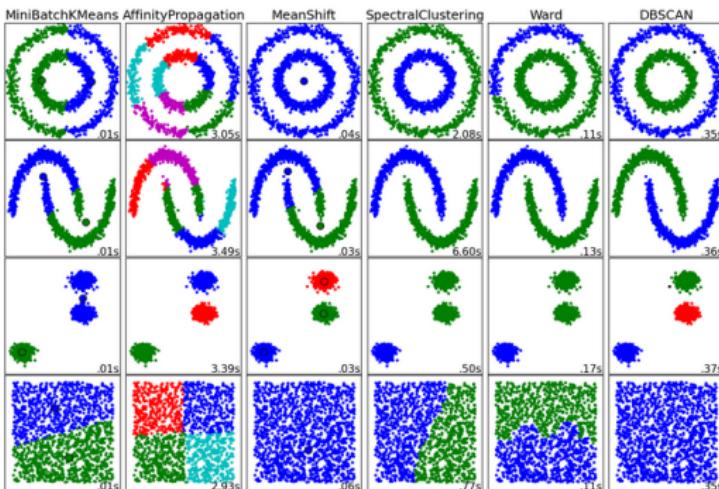
## Aprendizado não supervisionado

## Factor Analysis

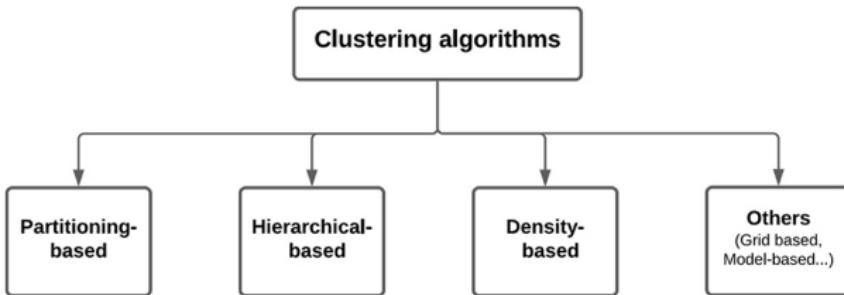


# Aprendizado não supervisionado

- Clusterização
- Pode ser usada como pré-processamento anterior ao uso de algoritmos de aprendizado supervisionado



# Aprendizado não supervisionado



Fonte: [http://www.acsu.buffalo.edu/~danet/Sp18/MTH448/class23/class23\\_files/community\\_detection.html](http://www.acsu.buffalo.edu/~danet/Sp18/MTH448/class23/class23_files/community_detection.html)

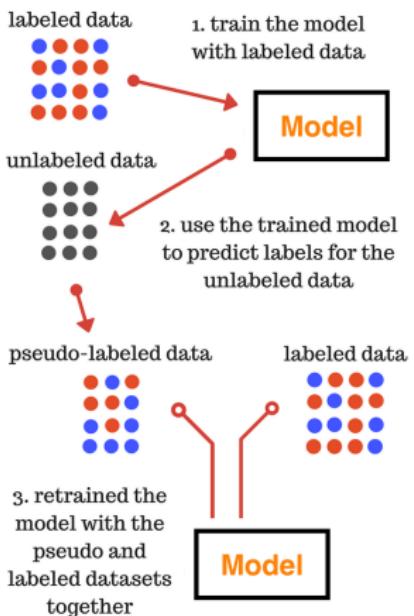
## Aprendizado semi-supervisionado (ASS)

- O conjunto possui dados rotulados e uma grande quantidade de dados não rotulados.
- Utilizam-se os **rótulos** dos dados rotulados para inferir **pseudorótulos** aos demais.
- Seleção do melhor modelo construído com base nos rótulos.

### Uso:

- Reconhecimento de fala;
- Classificação de conteúdo na web.
- Classificação de documentos de texto.

# Aprendizado semi-supervisionado (ASS)



Fonte: <https://www.kaggle.com/code/xiejialun/prepare-pseudo-label>

## Aprendizado semi-supervisionado (ASS)

- Prós: Com uma porção mínima de dados representativos, o aprendizado pode ter resultados aceitáveis. Bom para lidar com clusterização e detecção de anomalias.
- Contras: Falta de representatividade dos dados rotulados pode prejudicar o aprendizado reduzindo a acurácia, fazendo com que seja necessário ter mais dados rotulados para melhorar o aprendizado. Se muitos dados já forem rotulados, o ideal é partir para algoritmos de classificação porque o custo envolvido em ASS é alto.

# REFERÊNCIAS

-  DRIVER, H.; KROEBER, A. *Quantitative Expression of Cultural Relationships....* University Press, 1932. (University of California publications in American archaeology and ethnology). Disponível em: <<https://books.google.com.br/books?id=4u9szQEACAAJ>>.
-  DY, J. G.; BRODLEY, C. E. Feature selection for unsupervised learning. *Journal of machine learning research*, v. 5, n. Aug, p. 845–889, 2004.
-  HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 2rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2006. ISBN 1-55860-901-6.
-  HARTIGAN, J. A. Statistical theory in clustering. *Journal of classification*, Springer, v. 2, p. 63–76, 1985.
-  MAHDI, M.; HOSNY, K.; EL-HENAWY, I. Scalable clustering algorithms for big data: A review. *IEEE Access*, PP, p. 1–1, 05 2021.
-  MÜLLER, A.; GUIDO, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016. Disponível em: <<https://books.google.com.br/books?id=vbQIDQAAQBAJ>>. ISBN 9781449369897.
-  NIU, D.; DY, J. G.; JORDAN, M. I. Iterative discovery of multiple alternativeclustering views. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 36, n. 7, p. 1340–1353, 2013.
-  SCHOLZ, M. *Approaches to analyse and interpret biological profile data*. Tese (Doutorado) — Universität Potsdam.
-  ZUBIN, J. A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, American Psychological Association, v. 33, n. 4, p. 508, 1938.