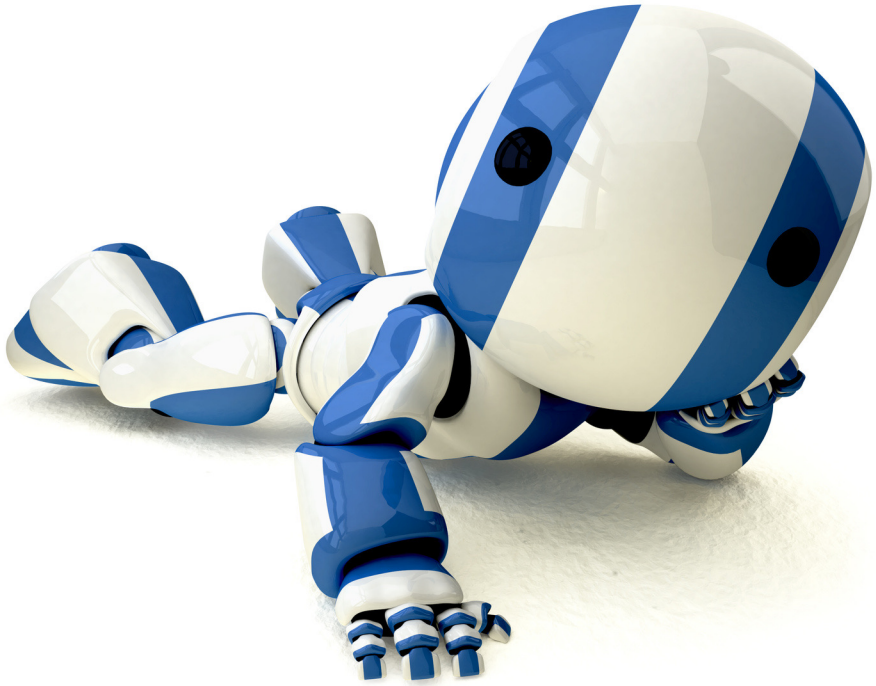
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA

Aprendizado de Máquina



Análise e Pré-Processamento de Dados

Prof. Me. Otávio Parraga

MALTA

Machine Learning Theory and Applications Lab

Aula de Hoje

- Caracterização de Dados
- Análise Exploratória de Dados
- Pré-Processamento de Dados

Aula de Hoje

- Caracterização de Dados
  - Instâncias e Atributos
  - Tipos de Atributos
- Análise Exploratória de Dados
- Pré-Processamento de Dados

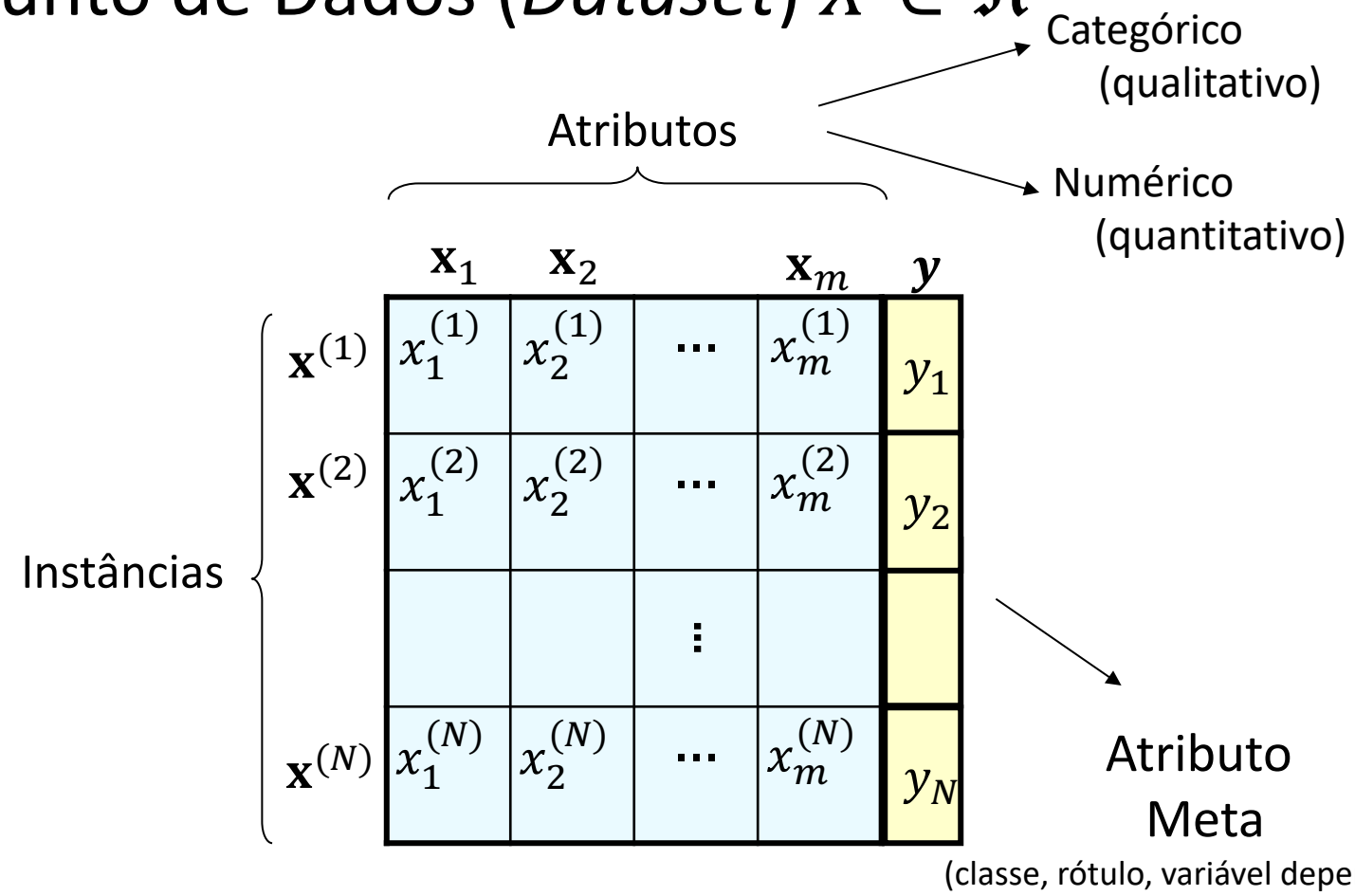
Dados

- Conjunto de Dados (*Dataset*) $X \in \Re^{N \times m}$
  - Linhas (N)
    - Instâncias (*instances*)
    - Objetos (*objects*)
    - Exemplos (*examples*)
    - Tuplas (*tuples*)
    - Amostras (*samples*)
    - Casos (*cases*)
    - Registros (*records*)
    - Vetores de características (*feature vectors*)
  - Colunas (m)
    - Atributos (*attributes*)
    - Características (*features*)
    - Campos (*fields*)
    - Variáveis (*variables*)
    - Dimensões (*dimensions*)

	$\mathbf{x}_1$	$\mathbf{x}_2$		$\mathbf{x}_m$
$\mathbf{x}^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	...	$x_m^{(1)}$
$\mathbf{x}^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	...	$x_m^{(2)}$
			...	
$\mathbf{x}^{(N)}$	$x_1^{(N)}$	$x_2^{(N)}$	...	$x_m^{(N)}$

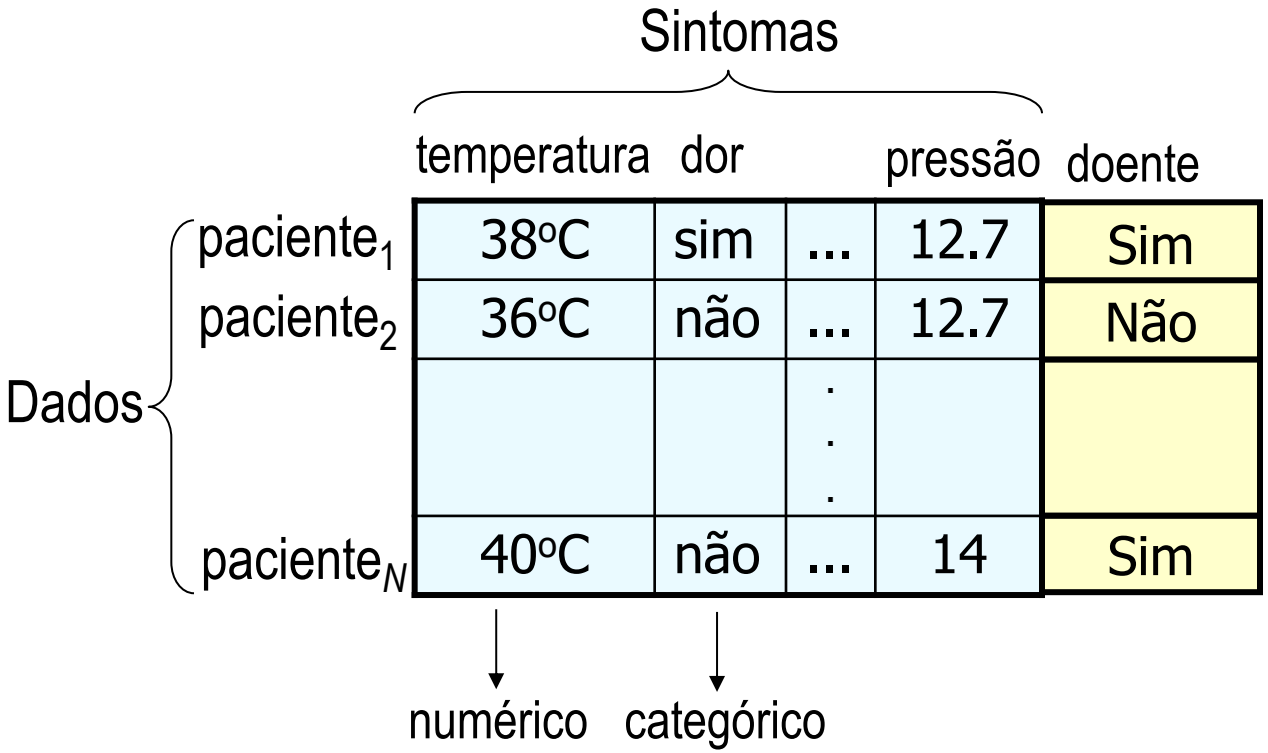
Dados

- Conjunto de Dados (*Dataset*) $X \in \Re^{N \times m}$



Dados

- Ex: Diagnóstico de uma doença

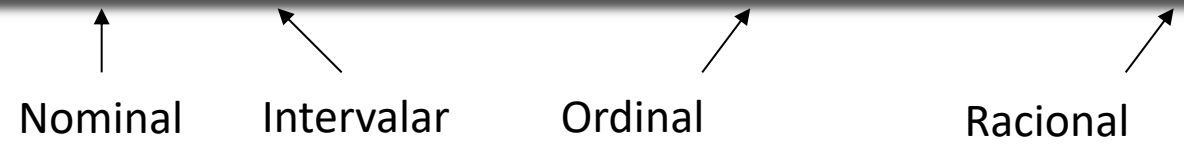


Tipos de Atributos

- Nominal (qualitativo, categórico)
  - Ex: cor, profissão, tipo sanguíneo
- Ordinal (qualitativo, categórico)
  - Ex: qualidade (ruim, médio, bom), dias da semana
- Intervalar (quantitativo, numérico)
  - Ex: data, temperatura em Celsius
- Racional (quantitativo, numérico)
  - Ex: peso, tamanho, idade, temperatura em Kelvin

Exemplo

Nome	Temp	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável
José	39.0	não	pequena	sim	2000	doente
Ana	37.3	não	grande	sim	1800	saudável
Leila	37.7	não	grande	sim	900	doente



Tipos de Atributos

	Tipo de Atributo	Descrição	Exemplos
Categórico (Qualitativo)	**Nominal**	Valores são simplesmente nomes (símbolos) diferentes, i.e., atributos nominais provêm apenas informação suficiente para distinguir uma instância de outra: (=, ≠)	Sexo, Estado Civil, CEP, …
	Ordinal	Os valores de atributos ordinais provêm informação suficiente para distinguir e ordenar instâncias, i.e.: (=, ≠) e (<, >)	Grau de Educação, Números de Endereço, …
Numérico (Quantitativo)	**Intervalo**	Atributos para os quais a diferença entre valores faz sentido, i.e., existe uma unidade de medida com referência (zero) arbitrário. Suporta as operações anteriores e ainda (+, -)	Datas, Temperatura em Fahrenheit, ...
	Razão	Atributos para os quais não apenas a diferença entre valores faz sentido, mas também a razão entre valores (zero é absoluto). Suporta as ops. anteriores e ainda (*, /)	Contagens, Massa, Largura, Corrente Elétrica, Quantidades Monetárias, ...

Tipos de Atributos

- Uma taxonomia alternativa para atributos pode ser estabelecida pelo número de valores
  - Atributo Contínuo
    - Assume uma quantidade incontável de valores
  - Atributo Discreto
    - Assume um número contável de valores
      - Finito ou infinito

Atributos Contínuos

- Assumem valores que são números reais
  - Temperatura
  - Peso
  - Distância
  - …

Atributos Discretos

- Valores enumeráveis (finitos ou infinitos)
  - estações do ano, cores elementares, código postal
  - nº de filhos, nº de estrelas, nº de anos (quantidades de elementos)

- Caso especial: Atributos **Binários**
  - 0 ou 1
  - V ou F
  - S ou N
  - …

Atributos
Binários Assimétricos

- Caso particular de atributos discretos binários
  - Assume dois valores como todo atributo binário
  - Porém, apenas um deles é relevante
    - Indica que instância possui determinada característica
    - Ex: aluno matriculado ou não em cada disciplina
      - Se nº de disciplinas disponíveis for grande, alunos são todos similares com relação às disciplinas que não cursam
  - Identificar um atributo binário como assimétrico é importante para o projeto de sistemas de AM!
    - Cenário clássico: text mining

Aula de Hoje

- Caracterização de Dados
- Análise Exploratória de Dados
  - Dados univariados
  - Medidas de localidade, espalhamento e distribuição
  - Dados multivariados
- Pré-Processamento de Dados

Motivação

- Exploração preliminar dos dados
  - Facilita entendimento de suas características
  - Ajuda a selecionar melhor técnica de pré-processamento ou aprendizado
  - Faz uso principalmente de:
    - Estatística descritiva
    - Visualização

Estatística Descritiva

- Permite capturar
  - Frequência dos dados
  - Localização ou tendência central
  - Dispersão ou espalhamento
  - Distribuição ou formato

Frequência

- Proporção de vezes que um atributo assume um dado valor
  - Frequentemente utilizada para análise de atributos categóricos
  - Ex: em um *dataset* de dados médicos, 37% dos pacientes têm febre

Exemplo de Frequência

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

66% das manchas são grandes

50% dos pacientes são doentes (saudáveis)

Medidas de Localidade

- Dados categóricos
  - Moda
- Dados numéricos
  - Média
  - Mediana
  - Percentil

Exemplo de Moda

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente

Moda para o atributo mancha: grande

Média

- Pode ser calculada facilmente:

$$\overline{x_j} = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}$$

- Problema: sensível a *outliers*

Mediana

- Menos sensível a *outliers* que a média
- Necessário ordenar valores
  - Complexidade maior que a média

$$mediana(x) = \begin{cases} x_{(r+1)} & \text{se n é ímpar} \\ \frac{1}{2}(x_r + x_{(r+1)}) & \text{se n é par} \end{cases}$$

$$r = \frac{n}{2} \longrightarrow \text{Divisão inteira}$$

Média e Mediana

- Média é bom indicador do centro do conjunto de valores quando estes estão distribuídos simetricamente
- Mediana indica melhor centro
  - Se distribuição é oblíqua (assimétrica, *skewed*)
  - Se existem *outliers*

```
idade = np.array([23, 20, 22, 21, 110])

# Calcula a média do array.
media = idade.mean()

# Calcula a mediana do array.
mediana = np.median(idade)

print (f'A média das idades é {media:.2f} e a mediana é {mediana:.2f}')
```

```
A média das idades é 39.20 e a mediana é 22.00
```

Média Podada

- *Trimmed mean*
- Minimiza problema da média descartando valores dos extremos
  - Dados são ordenados
  - Porcentagem p de valores são eliminados de cada extremidade

```
from scipy.stats import trim_mean

x = np.array([1, 2, 3, 4, 5])

# Eliminar 20% (0.2) dos valores em cada extremidade
media_podada = trim_mean(x, 0.2)

print(media_podada)
```

```
3.0
```

Quartis e Percentis

- Mediana divide os dados ao meio
- Outras medidas usam pontos de divisão distintos
  - Quartis (divisão em quartos)
    - Primeiro quartil (Q_1) é o valor da amostra onde 25% dos valores são inferiores a ele
      - Também conhecido por 25º percentil (P25)
    - Segundo quartil Q_2 = mediana

Percentil

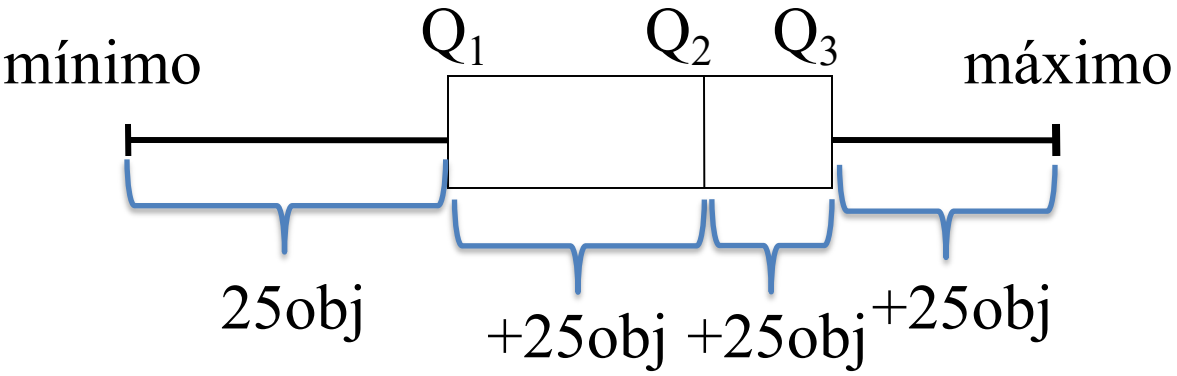
- Seja $\mathbf{x}$ um atributo contínuo/ordinal e $p = [1,100]$
  - O pº percentil é um valor de $\mathbf{x}$ tal que $p\%$ dos valores observados são menores do que aquele valor
    - Ex: 40º percentil de $\mathbf{x}$ é o valor $\mathbf{x}_{40\%}$ tal que 40% dos valores de $\mathbf{x}$ são menores que $\mathbf{x}_{40\%}$

```
x = np.array([1, 2, 3, 4, 5])

# Percentil 50% (mediana)
p50 = np.percentile(x,50)

print (f'P50% = {p50}')
```

```
P50% = 3.0
```

Boxplot

- Resumo das informações dos quartis pode ser apresentado em um gráfico chamado boxplot
  - Quanto menor a área, mais objetos estão na faixa delimitada
  - Se supormos 100 instâncias (Ex. 100 idades)



Medidas de Espalhamento

- Medem a dispersão (ou grau de espalhamento) de um conjunto de valores
- Indicam se um atributo está
  - Amplamente espalhado
  - Relativamente concentrado em torno de um ponto (ex. média)
- Medidas comuns
  - Intervalo
  - Variância
  - Desvio padrão

Intervalo

- Medida simples
  - Calcula o espalhamento máximo
  - Sejam $\{\mathbf{x}_j^{(1)}, \ldots, \mathbf{x}_j^{(\mathrm{N})}\}$ os valores de $\mathbf{x}_j$ para N objetos:

$$r(\mathbf{x}_j) = \max(\mathbf{x}_j) - \min(\mathbf{x}_j)$$

  - Atenção: pode ser má ideia!
    - Valores concentrados ao redor de um ponto, porém com alguns poucos valores extremos

Variância (σ^2)

- Medida preferida para análise de espalhamento

$$\sigma^2(\mathbf{x}_j) = \frac{1}{(N-1)} \sum_{i=1}^{N} \left(x_j^{(i)} - \bar{x}_j \right)^2$$

- Denominador ($N-1$): correção de Bessel, usada para uma melhor estimativa da verdadeira variância
- Segundo momento central
- Desvio padrão (σ): raiz quadrada da variância

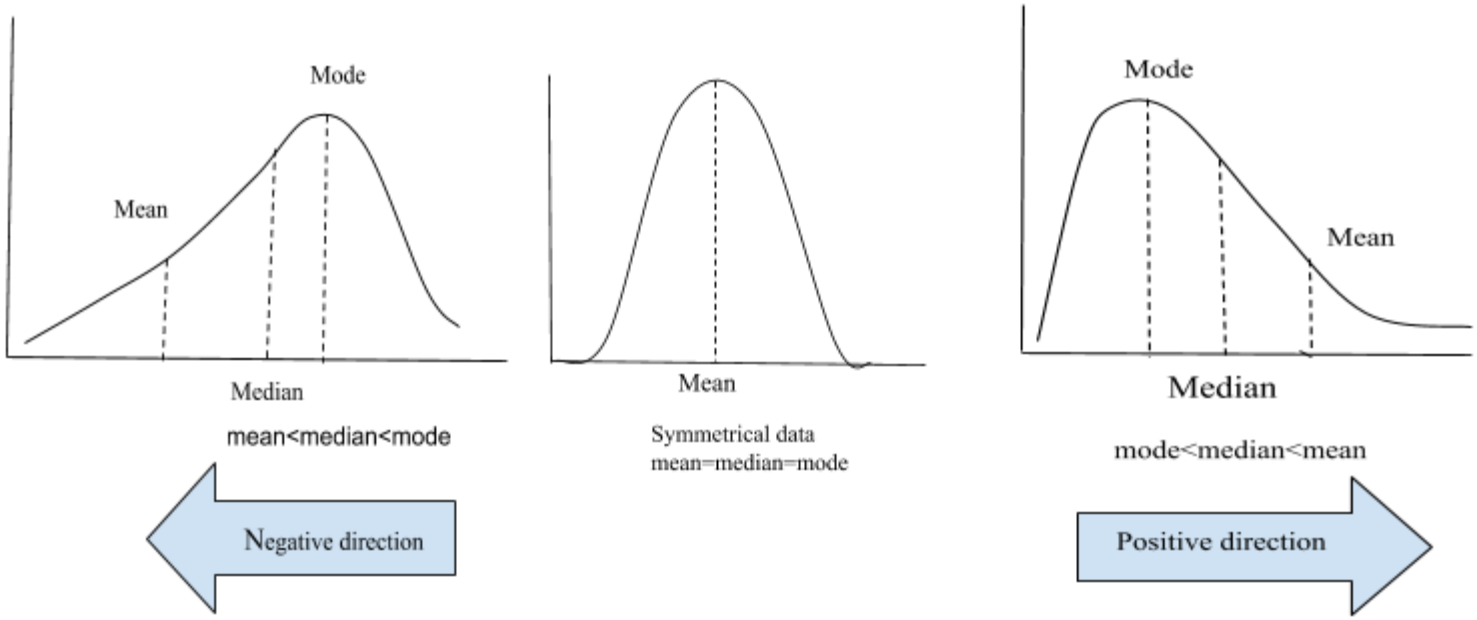
Obliquidade (γ)

- *Skewness*
  - Mede a simetria da distribuição em torno da média

$$\gamma(\mathbf{x}_j) = \frac{\frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_j^{(i)} - \overline{x_j}\right)^3}{\sigma^3}$$

  - Divisão por σ^3 torna medida independente de escala
  - Terceiro momento central padronizado

Obliquidade (γ)



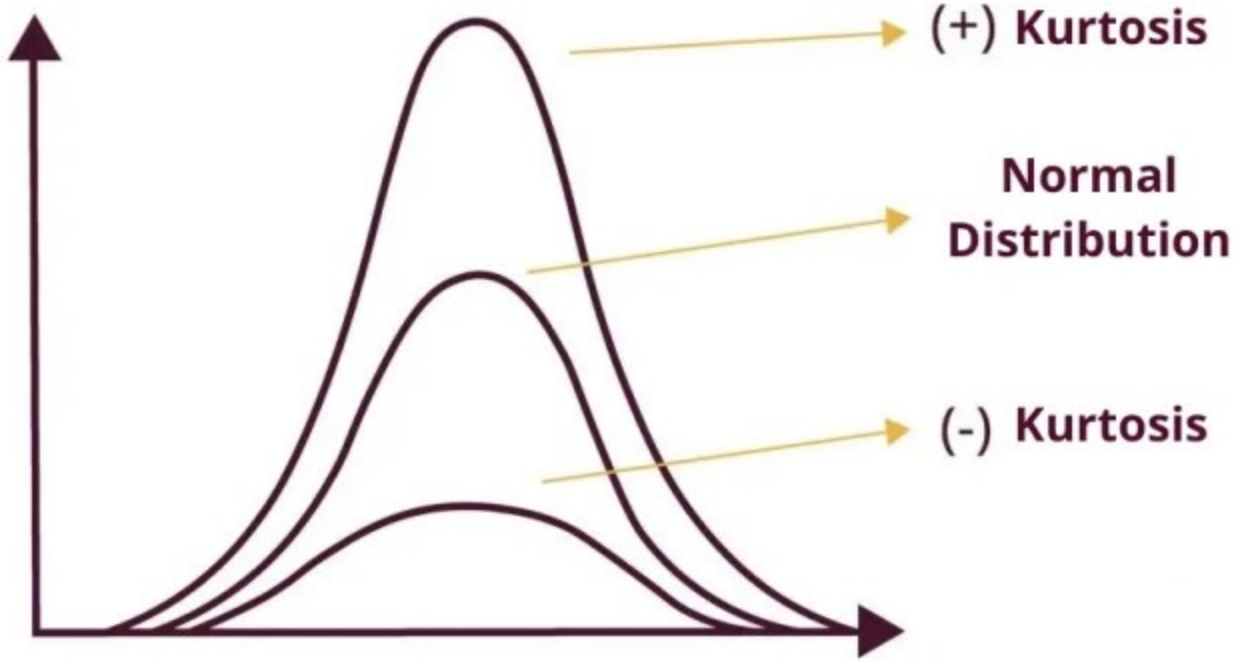
Curtose (β)

- Kurtosis
  - Mede o formato (achatamento) da distribuição

$$\beta(\mathbf{x}_j) = \frac{\frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_j^{(i)} - \overline{x_j}\right)^4}{\sigma^4}$$

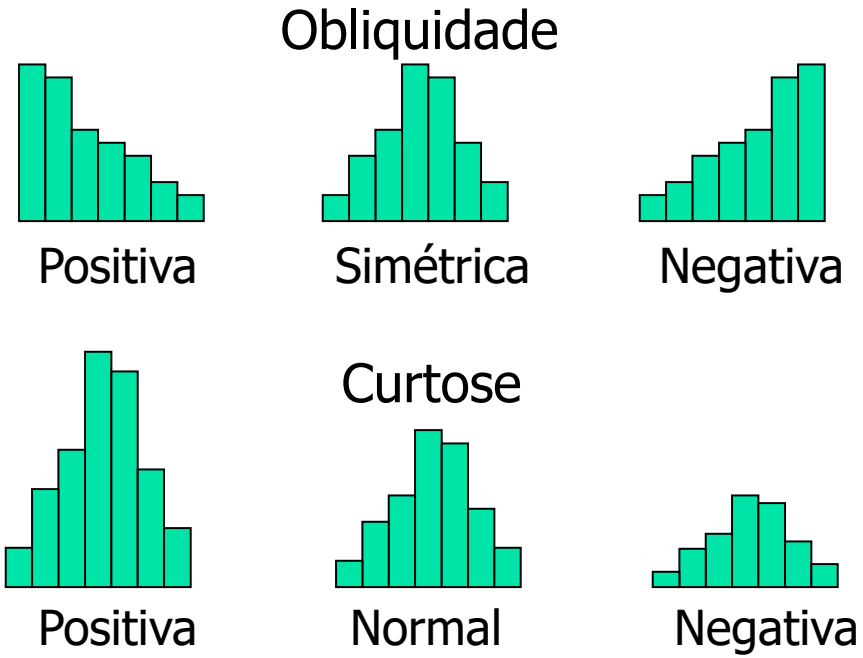
  - Quarto momento central padronizado
  - Valor de $\beta(x) = 3$ para distribuição normal padrão
  - Portanto, é comum subtrair $\beta(x)$ de 3 para que a distribuição normal padrão tenha $\beta(x) = 0$

Curtose (β)



Histograma

- Para visualizar graficamente curtose e obliquidade, utilizar histograma dos dados



Dados Multivariados

- Análise sobre vários atributos
- Medidas de localização
  - Obtidas por cada atributo separadamente
    - Depois agregá-las (embora nem sempre faça sentido)
  - Média dos objetos de um *dataset* com m atributos

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)^T$$

Dados Multivariados

- Medidas de espalhamento
  - Calculadas por atributo, de maneira independente
    - Qualquer uma das medidas
  - Variáveis contínuas
    - Espalhamento do *dataset* pode ser capturado por uma matriz de covariância
      - Cada elemento é a covariância do par de atributos

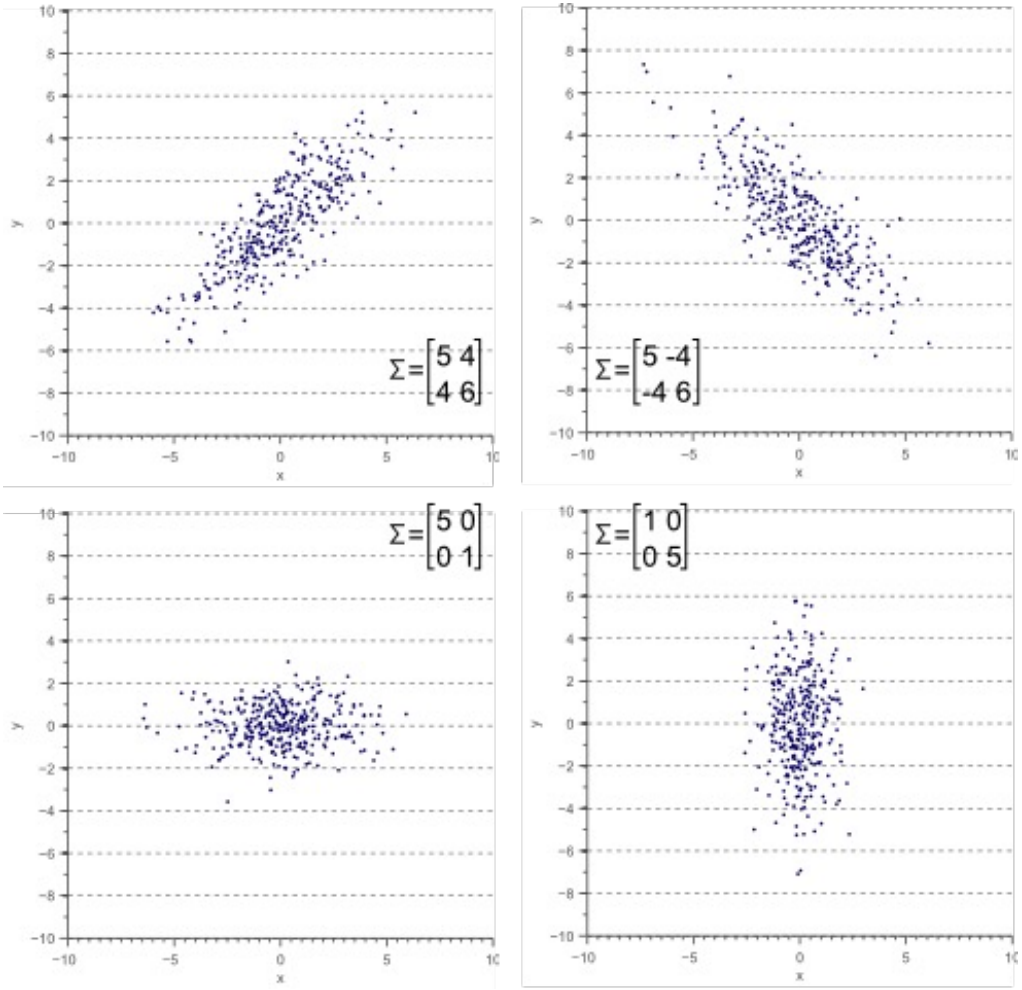
Dados Multivariados

- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos

$$cov(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_j^{(i)} - \bar{x}_j)(x_k^{(i)} - \bar{x}_k)$$

Dados Multivariados

- Matriz de covariância



Dados Multivariados

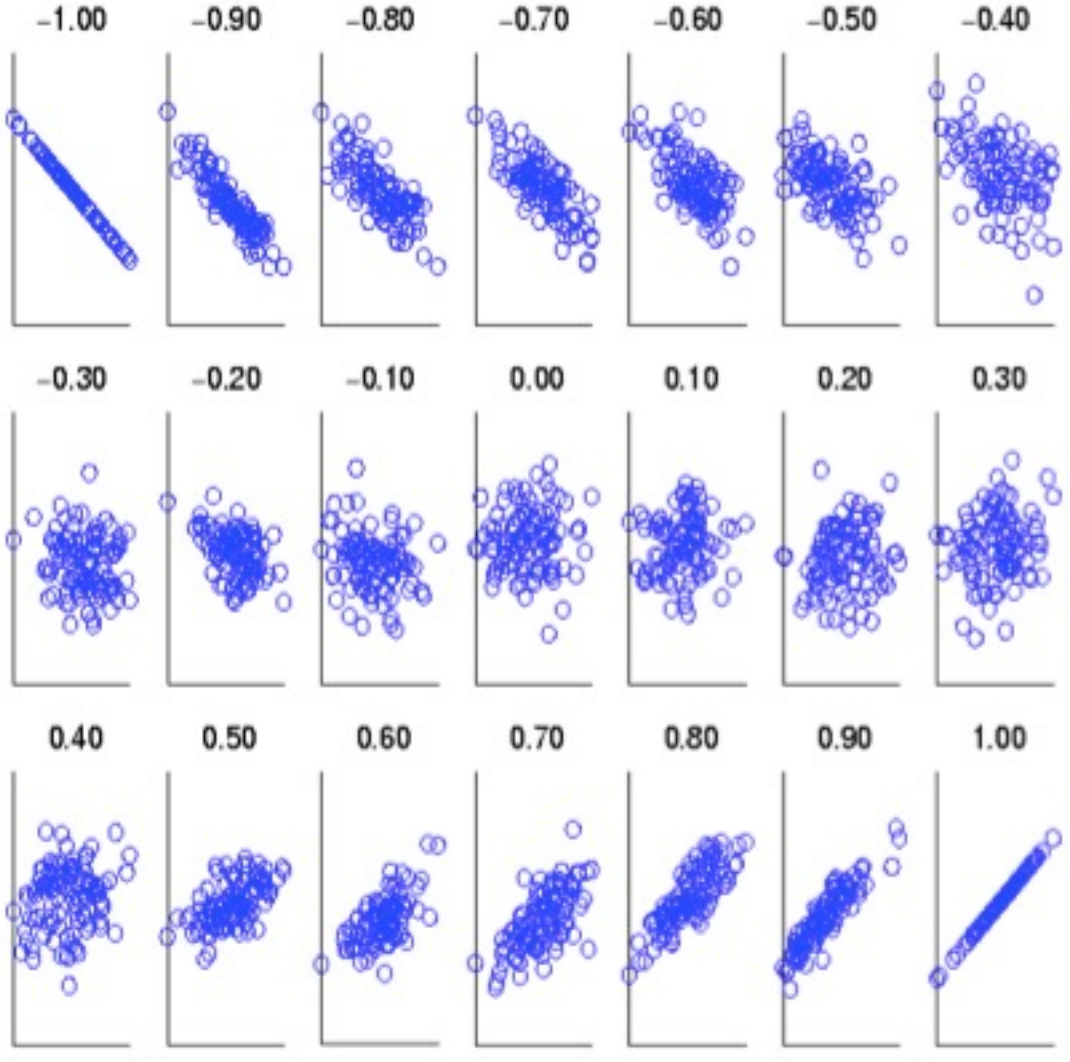
- Covariância não indica com clareza o relacionamento entre os pares de atributos
  - Sugestão é utilizar correlação!
    - Indica relacionamento linear entre 2 variáveis
    - É preferível para explorar dados do que covariância

$$\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = \frac{cov(\mathbf{x}_j, \mathbf{x}_k)}{\sigma_{\mathbf{x}_j}\sigma_{\mathbf{x}_k}}$$

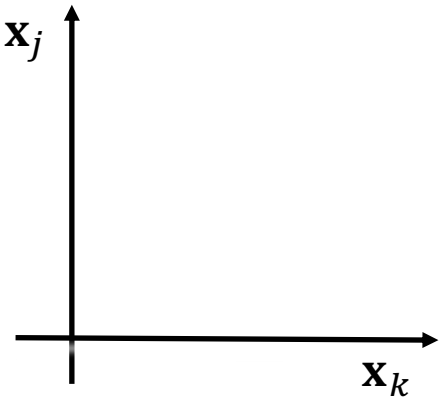
Dados Multivariados

- Correlação (de Pearson)
  - Varia entre -1 e 1
  - Magnitudes dos vetores são desprezadas ao se normalizar pelos desvios
    - Despreza média e variabilidade
  - Correlação de 1 (-1) significa que $\mathbf{x}_j$ e $\mathbf{x}_k$ tem um relacionamento linear positivo (negativo) perfeito
    - Correlação de 0 significa que não há relacionamento linear entre as variáveis (mas pode ser que haja relacionamento não-linear!)

Análise Visual da Correlação



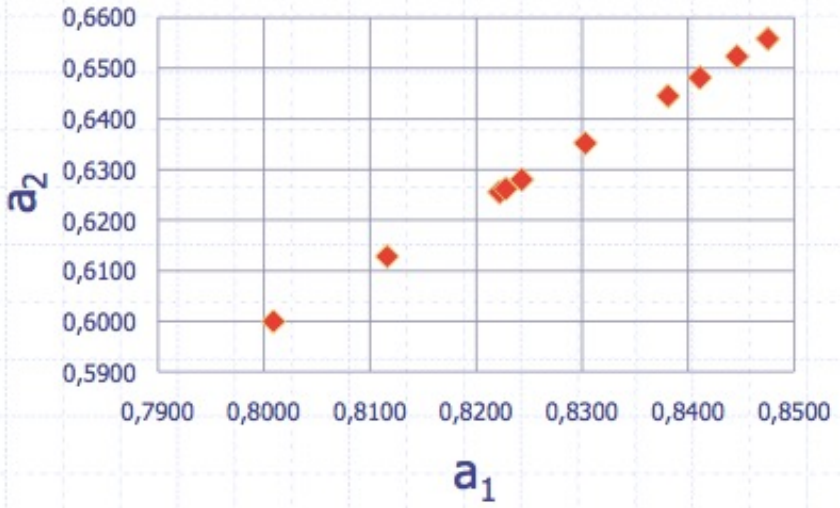
par de atributos $\mathbf{x}_j$ e $\mathbf{x}_k$
30 instâncias

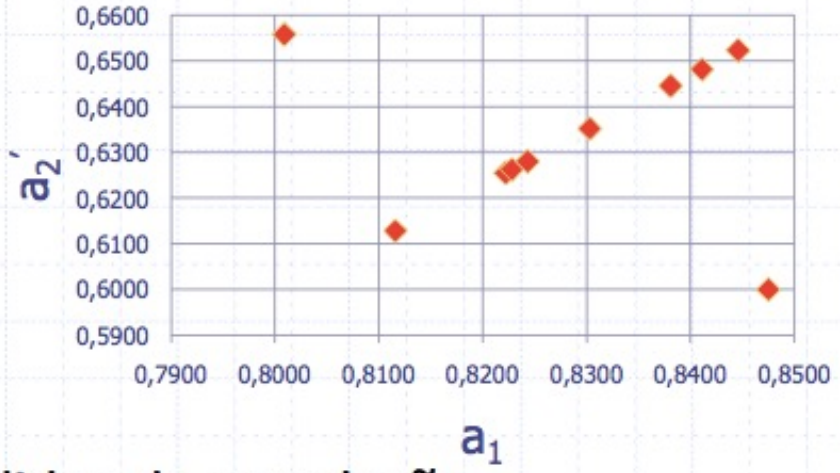


Pearson possui seus problemas; por exemplo...

a_1	a_2	a_2'
0.8009	0.6000	0.6558
0.8116	0.6128	0.6128
0.8222	0.6255	0.6255
0.8228	0.6262	0.6262
0.8243	0.6280	0.6280
0.8303	0.6352	0.6352
0.8381	0.6446	0.6446
0.8411	0.6482	0.6482
0.8446	0.6523	0.6523
0.8475	0.6558	0.6000

	(a_1, a_2)	(a_1, a_2')
Pearson	1.000	-0.080

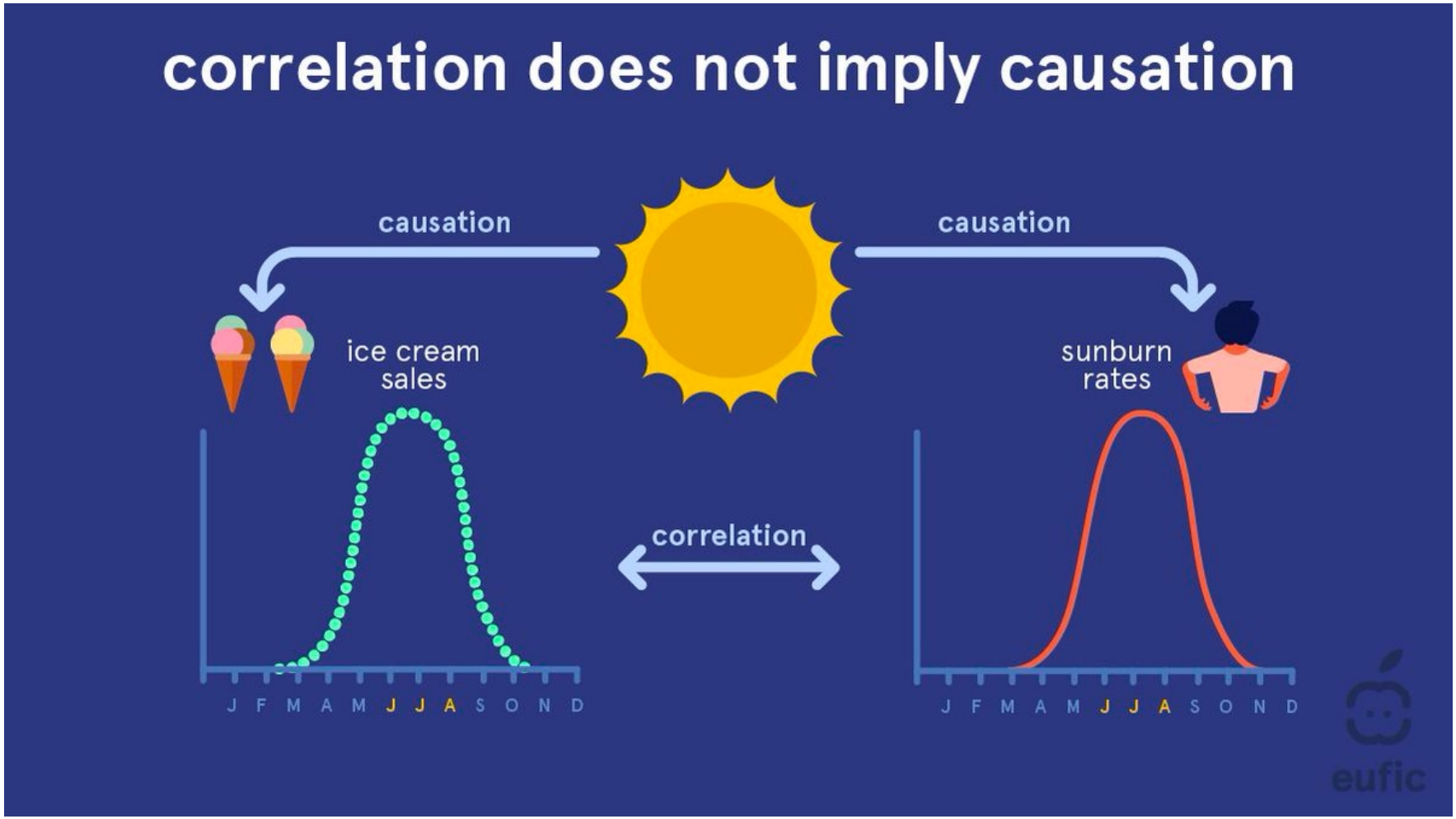




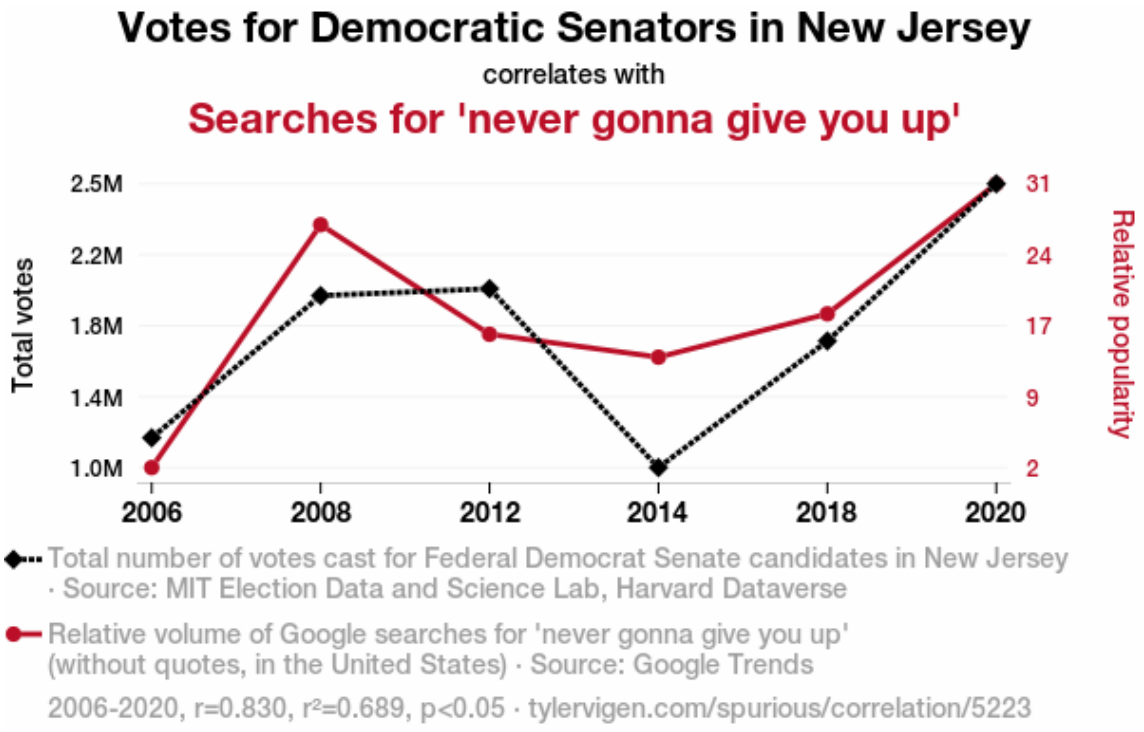
... mas existem outras medidas de correlação...

Campello, R. J. G. B., Hruschka, E. R., *On Comparing Two Sequences of Numbers and Its Applications to Clustering Analysis*, **Information Sciences**, 2009

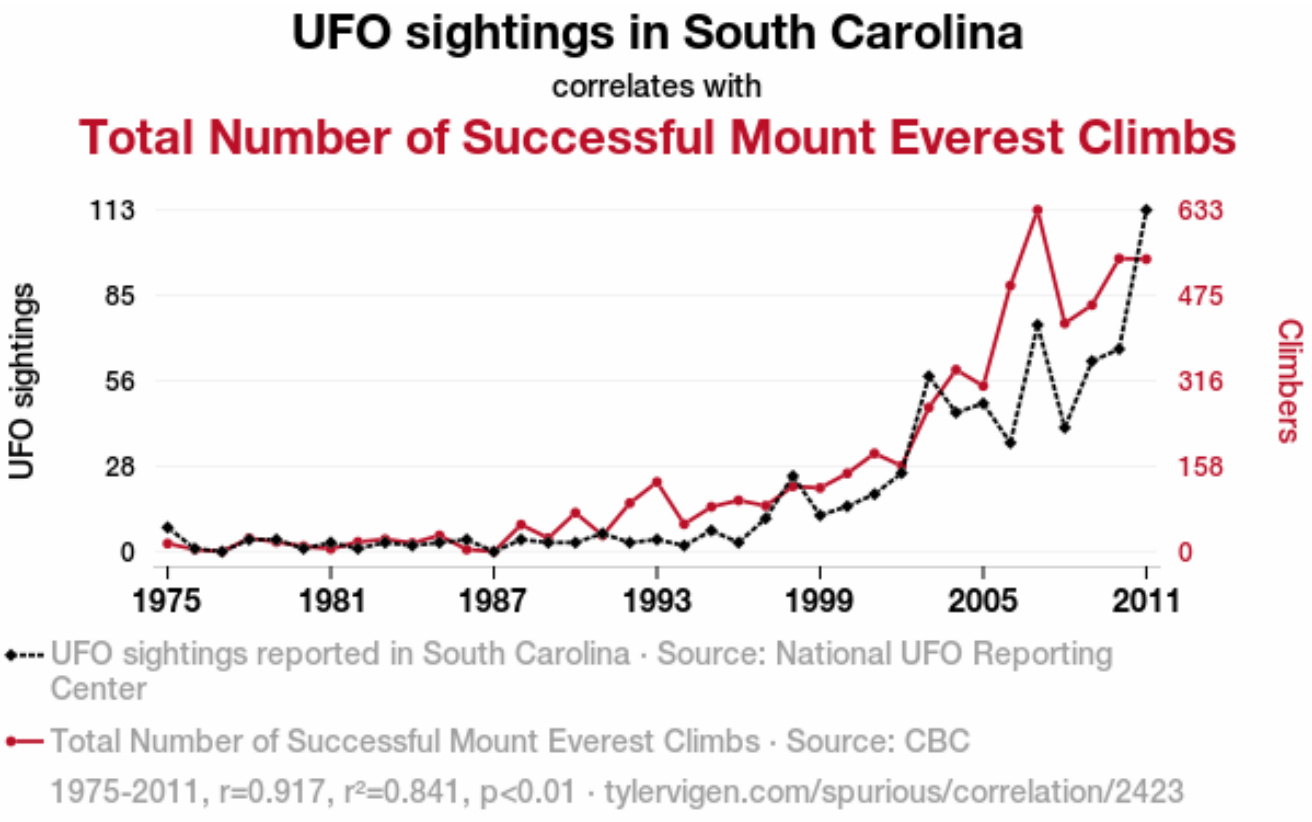
Correlação não Indica Causalidade



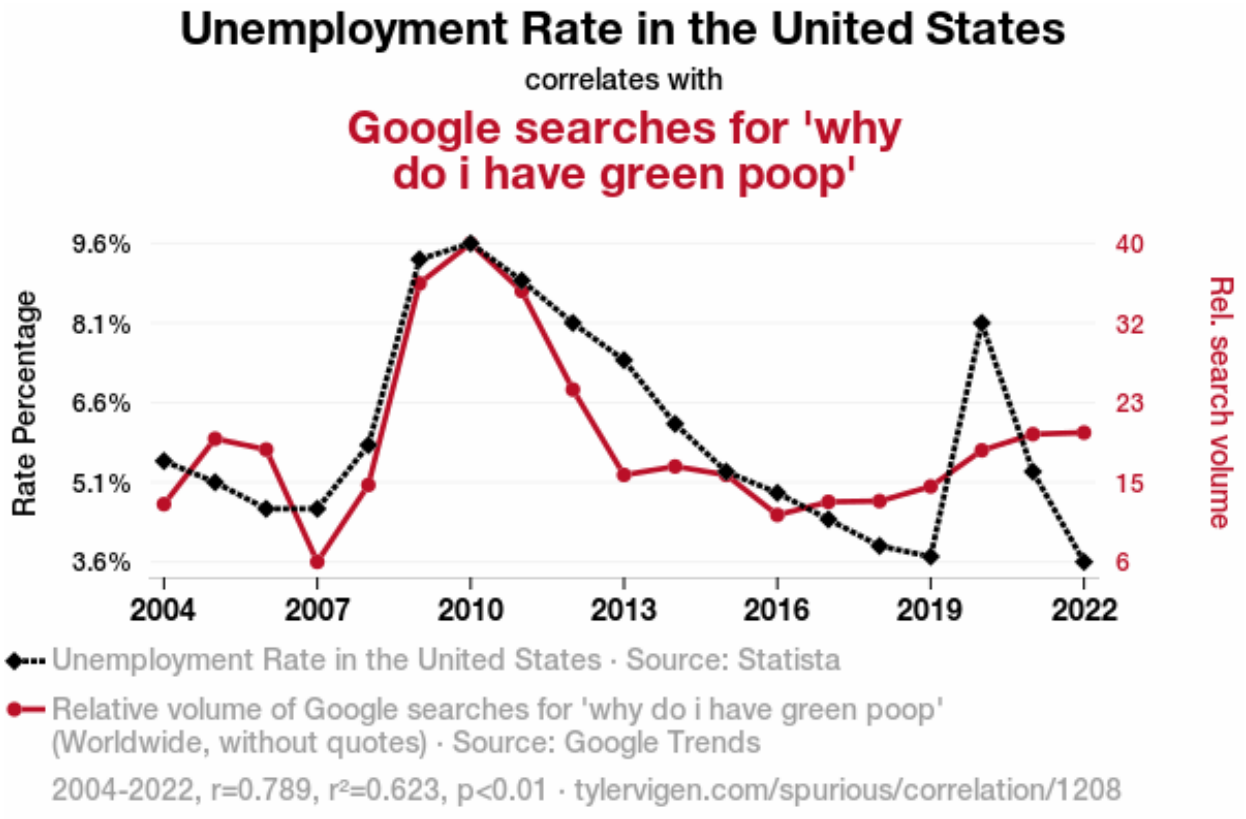
Correlação não Indica Causalidade



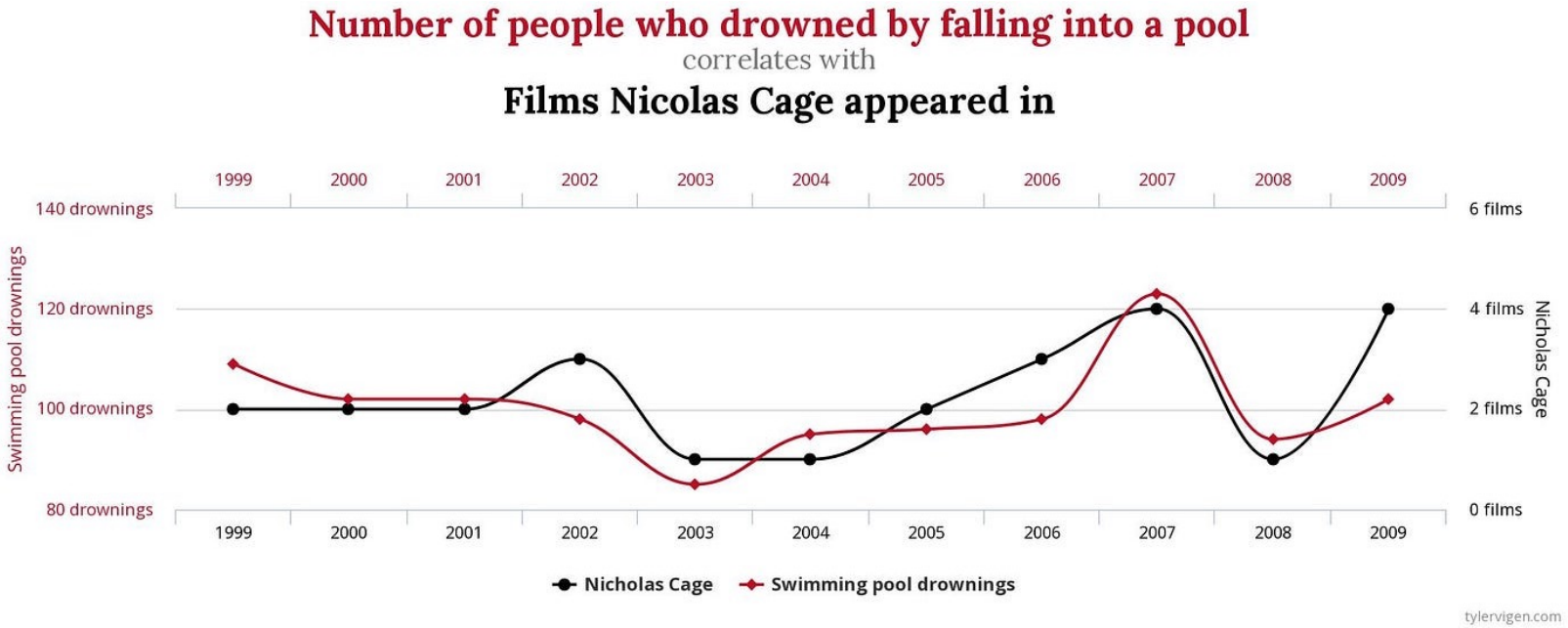
Correlação não Indica Causalidade



Correlação não Indica Causalidade



Correlação não Indica Causalidade



Aula de Hoje

- Caracterização de Dados
- Análise Exploratória de Dados
- Pré-Processamento de Dados
  - Transformações de Dados
    - Conversões
    - Discretizações
    - Normalizações

Transformações de Dados

- Conversão de valores simbólicos para numéricos
- Conversão de valores numéricos para simbólicos
- Normalização de valores numéricos

Conversão de Valores Categóricos

- Muitos algoritmos de AM trabalham apenas com variáveis numéricas
  - Redes Neurais, SVMs, etc.
  - Variáveis categóricas precisam ser convertidas
- Conversão depende da existência de ordem
  - Variáveis são nominais ou ordinais?

Conversão de Valores Ordinais

- Para variáveis ordinais, a ordem dos valores deve ser mantida de alguma maneira
  - Estratégia comum: associar valores inteiros crescentes
    - Ex: {frio, morno, quente} = {1, 2, 3}
  - Tal estratégia pode inserir distorções relativas entre os conceitos (qualquer política de pesos também insere!)
    - Diferenças entre símbolos são subjetivas

Conversão de Valores Nominais

- Atributos nominais
  - Conversão é feita por binarização
  - Codificações usuais
    - Codificação inteira-binária
    - Codificação 1-de-k (canônica, one-hot)

Conversão de Valores Nominais

- Codificação 1-de-k (*one-hot encoding*)
  - Um atributo binário associado a cada valor nominal
  - Exemplo:
    - Codificar {amarelo, vermelho, verde, azul, laranja, branco}

100000 - amarelo
010000 - vermelho
001000 - verde
000100 - azul
000010 - laranja
000001 - branco

Conversão de Valores Nominais

- Codificação 1-de-k
  - Pode gerar um número enorme de atributos!
    - Maldição da Dimensionalidade
  - Porém, possui várias propriedades interessantes:
    - Mantém equidistantes quaisquer dois vetores binários
    - Atributos binários são descorrelacionados
    - Atributos binários são assimétricos (importante no caso de alguns algoritmos de AM)
    - Moda do atributo nominal corresponde ao atributo binário com maior número de 1s

Atributos Nominais com Muitos Valores

- Codificação 1-de-k pode levar a dados muito esparsos quando k é grande
- Solução pode estar no uso de conhecimento de domínio do problema em questão
- Ex: atributo = nome de país
  - Existem 193 países membros da ONU
    - Codificação 1-de-k demandaria 192 atributos adicionais
    - Dados esparsos / maldição da dimensionalidade!

Atributos Nominais com Muitos Valores

- Ex: atributo = nome de país
  - Possível solução:
    - Utilizar 1 atributo nominal com apenas 7 valores (continentes)
    - Tentar discriminar entre os países com um conjunto menor de pseudo-atributos numéricos
      - PIB, população, IDH, temperatura média, …
    - Funcionamento satisfatório depende da aplicação
    - Não existe abordagem de AM sempre melhor ou pior
      - No free lunch!

Discretização

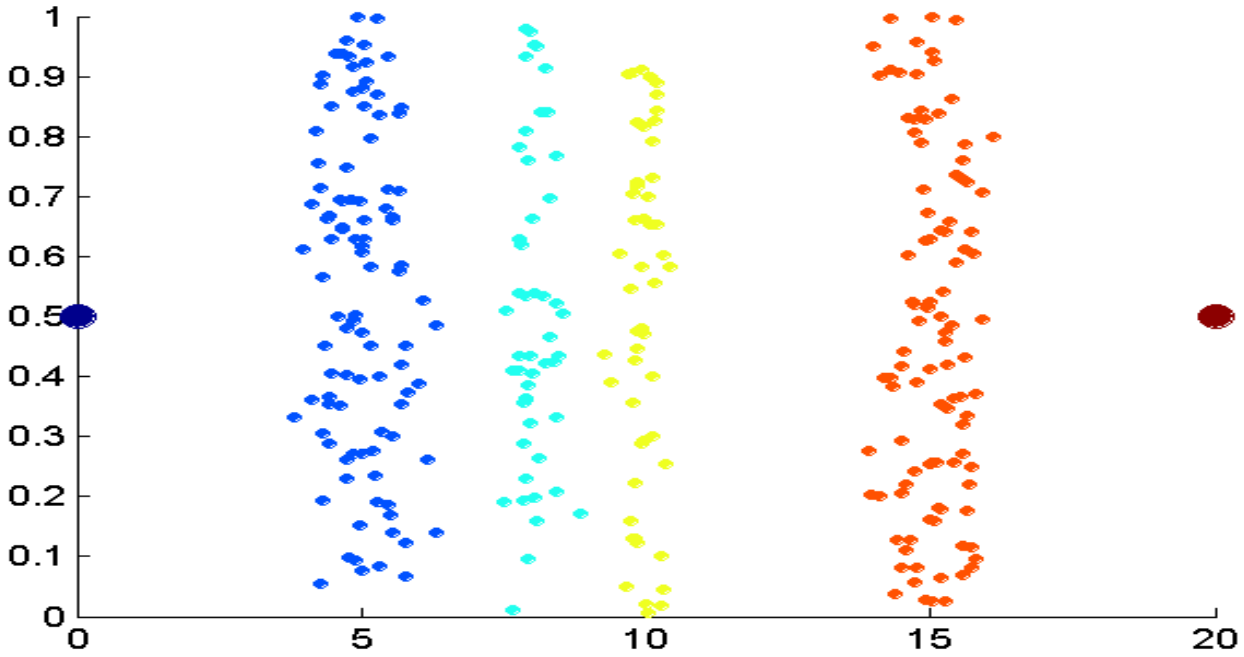
- Alguns algoritmos de AM aceitam apenas valores categóricos
  - Valor numérico precisa ser discretizado em intervalos
- Melhor discretização depende de:
  - Algoritmo que utilizará os valores discretizados
  - Demais atributos
  - …
- Em geral é realizada *a priori*, como pré-processamento

Discretização

- Transformar valores contínuos em intervalos
  - Atributo se transforma em categórico ordinal
- Passos necessários
  - Definição do número de intervalos (categorias)
    - Geralmente ad-hoc (feito pelo usuário)
  - Definição de como mapear os valores contínuos para as novas categorias
    - Definir limites/tamanho dos intervalos
    - Geralmente feito pelo algoritmo

Discretização Não-Supervisionada

- Exemplo: discretizar eixo x
  - Eixo y apenas para melhorar visualização

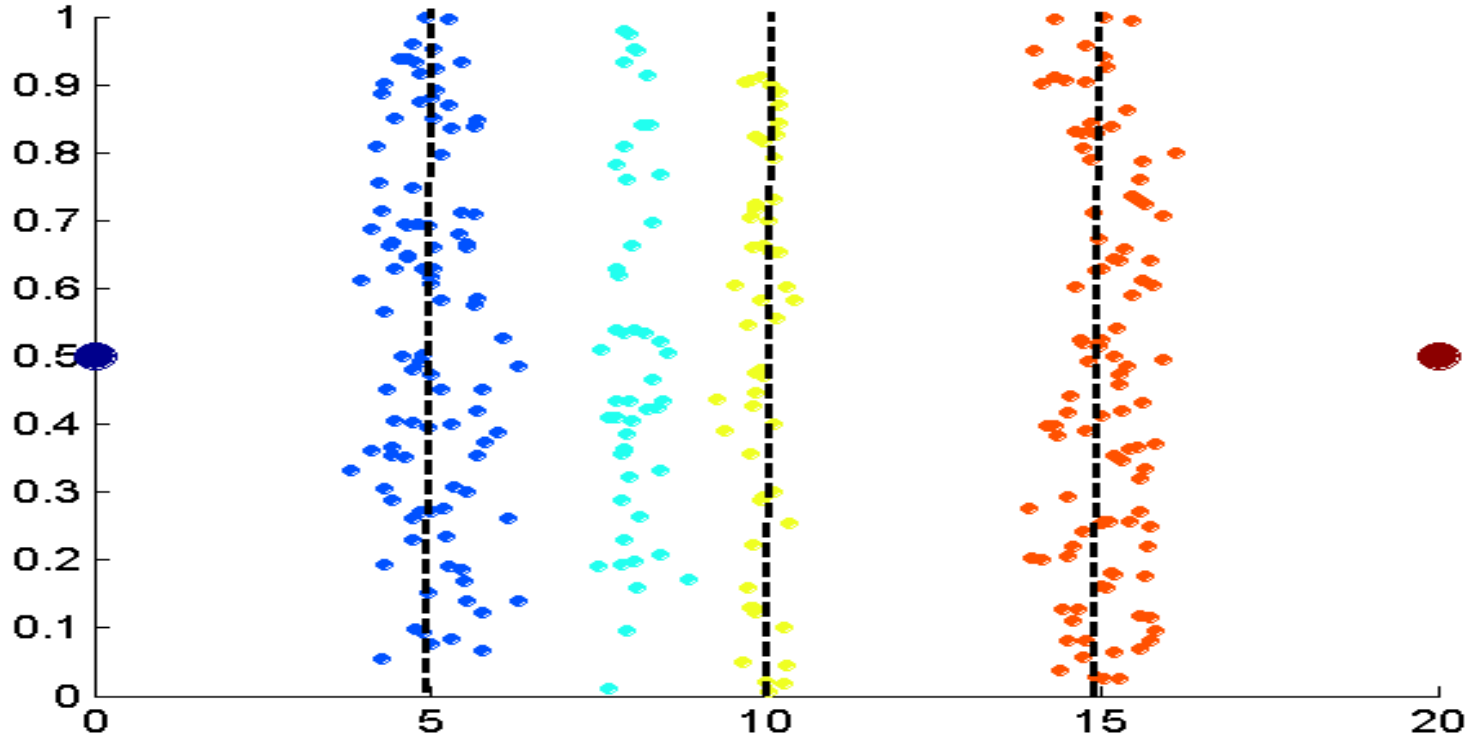


Discretização Não-Supervisionada

- Algoritmos Simples
  - Larguras Iguais
    - Divide intervalo original de valores em k sub-intervalos com mesma largura
    - Simples de implementar, porém:
      - Assume que valores possuem distribuição uniforme
      - Muito ineficaz em distribuições não uniformes
      - Muito sensível à presença de outliers

Discretização Não-Supervisionada

- Larguras iguais ($k = 4$)

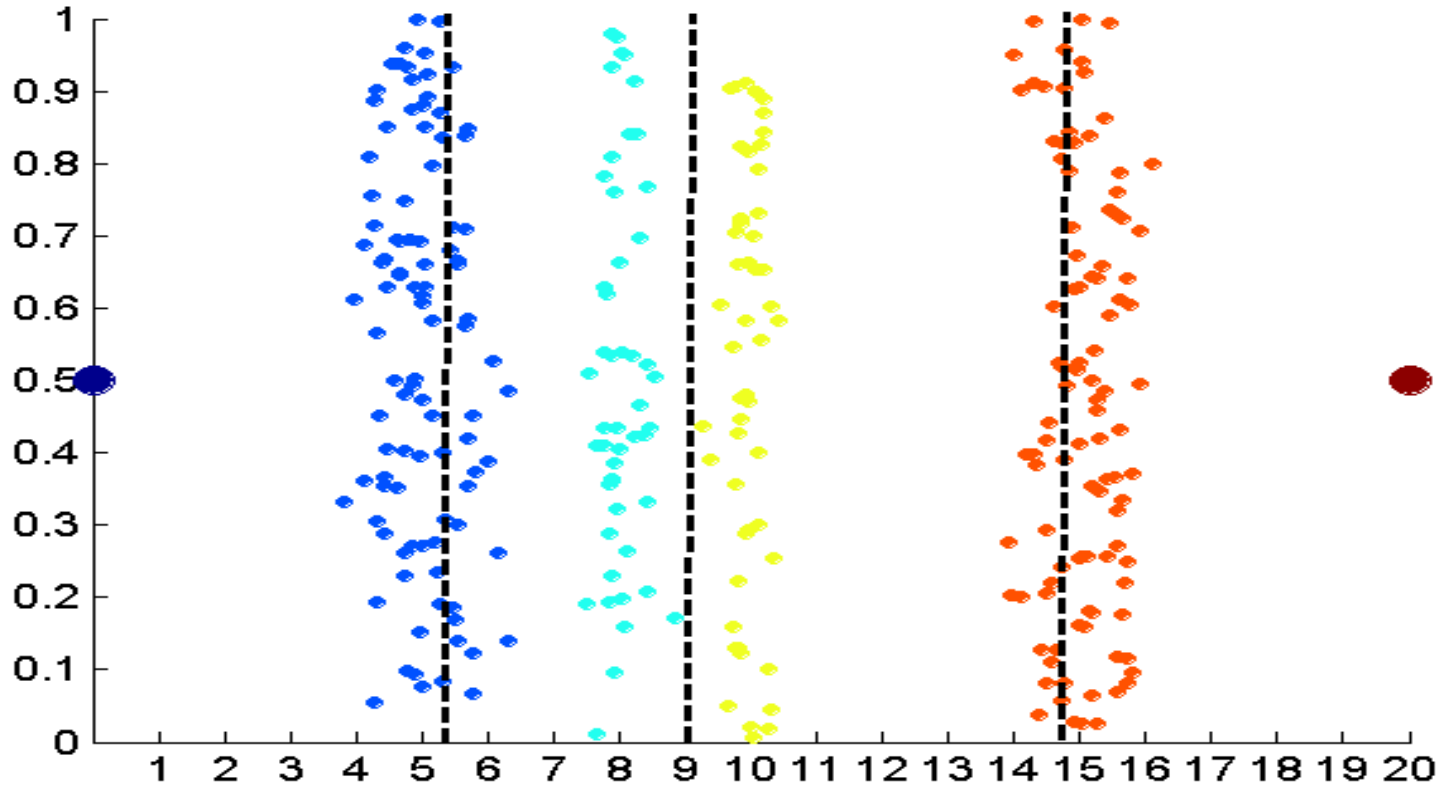


Discretização Não-Supervisionada

- Algoritmos Simples
  - Frequências Iguais
    - Atribui o mesmo número de instâncias por intervalo
    - Simples de implementar, porém:
      - Assume que valores estão em grupos balanceados
      - Muito ineficaz em distribuições desbalanceadas

Discretização Não-Supervisionada

- Frequências iguais ($k = 4$)



Discretização Não-Supervisionada

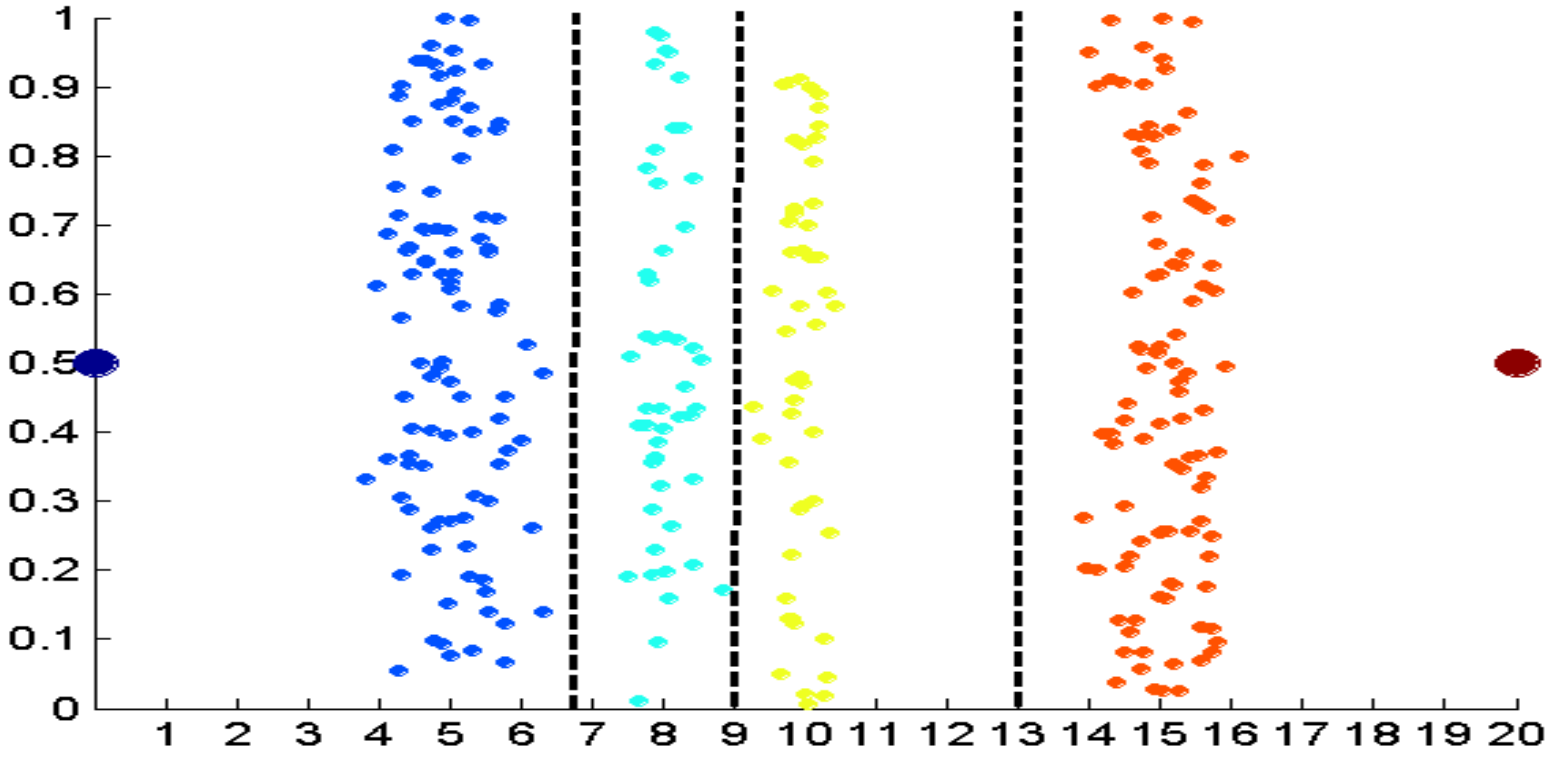
- Algoritmos Simples
  - Inspeção Visual
    - Observar gráfico com os valores do atributo e determinar visualmente os intervalos de acordo com a distribuição natural dos dados
    - Simples, eficaz e permite determinação eficiente de k
    - Porém:
      - Pré-processamento manual (*time-consuming*)

Discretização Não-Supervisionada

- Algoritmo de *Clustering*
  - Aplicar algum algoritmo de agrupamento de dados, como os que veremos posteriormente no curso (ex: k-means) para descobrir automaticamente a distribuição natural dos dados
    - Relativamente simples, eficaz e eficiente
    - Permite determinação automática do número de intervalos (grupos) através de critérios de validação de agrupamento

Discretização Não-Supervisionada

- Discretização de x com k-*means* ($k = 4$)



Normalização

- Transformação aplicada aos dados de forma com que estes exibam propriedades em comum
- Normalizações mais usuais em AM são as lineares
  - Re-escalar
  - Padronizar

Re-Escalar

- Re-escalar os valores de um atributo:
  - Adicionar ou subtrair uma constante
  - Multiplicar ou dividir por uma constante
- Utilizada para mudar unidade de medida dos dados
- Uso mais comum é para converter valores de atributos para o intervalo [0, 1] ou [-1, +1]
  - Ex: converter valores para intervalo $[novoMin, novoMax]$

$$x' = \frac{(x - \min(x)) \times (R)}{(\max(x) - \min(x))} + novoMin$$

$R = (novoMax - novoMin)$

Re-Escalar

- Muito utilizada em algoritmos baseados em otimização
  - Ex: redes neurais e SVM
  - Principalmente para evitar problemas numéricos oriundos da estratégia de otimização
- Problema: extremamente influenciada por *outliers*
  - Deve-se evitar re-escalar em aplicações sujeitas a outliers e/ou ruídos

Padronizar

- Padronizar os valores de um atributo
  - Adicionar/subtrair uma medida de localização
  - Multiplicar/dividir por uma medida de escala
- Para atributos com distribuição Gaussiana
  - Subtrair cada valor da média (μ)
  - Dividir pelo desvio padrão (σ)
  - Resultado: distribuição normal padrão: $\mathcal{N}(0,1)$
    Chamada de normalização z-score

Padronizar

- Normalização **z-score**:

$$x' = \frac{(x - \mu_x)}{(\sigma_x)}$$

- Muito utilizado para pré-processar dados de algoritmos de agrupamento (e demais tarefas que exijam cálculos de distância)

Exemplo

- Considere um *dataset* com atributos idade e salário
  - Diferenças em salário são bem maiores que diferenças em idade
    - Influencia algoritmos de AM que se utilizam de informações sobre diferenças (ex: distância Euclidiana)
    - Tal influência serve como "peso"
      - Salário está tendo mais importância que idade
      - Se isso não é desejável, padronizar!

Leitura Recomendada

- Capítulo 2 (Tan et al. 2006)
  - Amostragem de Dados
  - Qualidade de Dados
  - Redução de Dimensionalidade
    - Seleção de Atributos
- Capítulos 2 e 3 (Faceli et al. 2011)

Créditos e Referências

- Slides adaptados dos originais gentilmente cedidos pelos Profs. André Carvalho, Eduardo Hruschka, Ricardo Campello do ICMC-USP, Pang-Ning Tan, da MSU, e Rodrigo Coelho Barros, da PUCRS

- TAN, P. N. STEINBACH, M. KUMAR, V. **Introduction to Data Mining.** Addison-Wesley, 2005. 769 p.

- Faceli et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. LTC, 2011. 378 p.