Fundamentos dos Dados

SCC0652 – Visualização Computacional

Profa. Maria Cristina
cristina@icmc.usp.br

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)

VICG ICMC USP Grupo de Visualização, Imagens e Computação Gráfica

Sumário

Introdução

- Fontes de dados: **sensores**, **medições ou coletas**, **simulações ou computações**, **transações digitais**

Introdução

- Fontes de dados: **sensores**, **medições ou coletas**, **simulações ou computações**, **transações digitais**

- Dados **brutos** (*raw data*): não tratados, disponíveis como foram coletados

- Dados **processados** (*curated data*): passaram por processos de organização, limpeza, adequação (p.ex., suavização, normalização, interpolação)

Pipeline de Visualização

- Estágios do **pipeline de visualização**
  - Organização e seleção dos dados
  - Mapeamento visual
  - Definição dos parâmetros da cena
  - *Rendering* da visualização

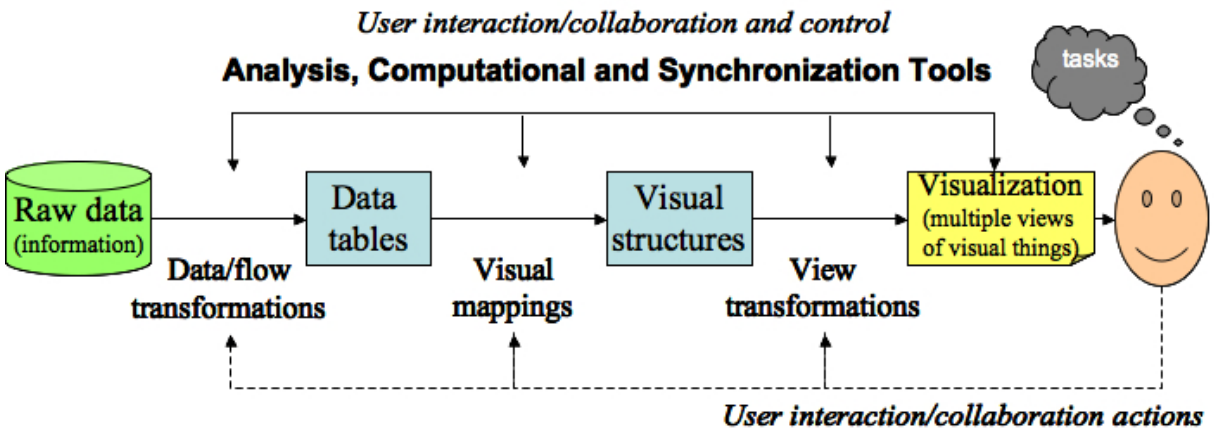

Figura: Pipeline de visualização.

Sumário

Organização dos Dados

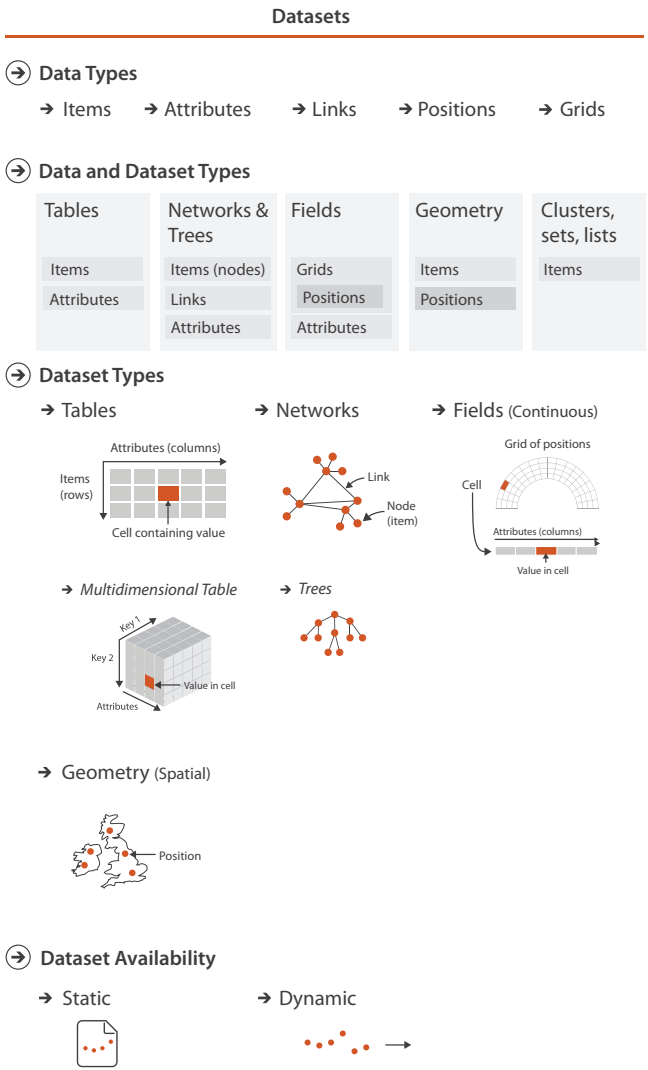

Figura: **O que** pode ser visualizado: tipos de dados e tipos de conjuntos de dados. Fonte: T. Munzner Visualization Analysis & Design (Fig. 2.2)

Tabelas de Dados

- Uma **Tabela de Dados** consiste de n **itens** (instâncias, amostras, observações)

$$(r_1, r_2, r_3, \ldots, r_n)$$

- Cada item é descrito por m $(m \geq 1)$ **atributos**, ou **variáveis**

$$(v_1, v_2, v_3, \ldots, v_m)$$

Tabelas de Dados

- Uma **Tabela de Dados** consiste de n **itens** (instâncias, amostras, observações)

$$(r_1, r_2, r_3, \ldots, r_n)$$

- Cada item é descrito por m $(m \geq 1)$ **atributos**, ou **variáveis**

$$(v_1, v_2, v_3, \ldots, v_m)$$

- Matriz de dados $n \times m$: cada linha descreve um item, cada coluna descreve um atributo relativo a todos os itens

Tabelas de Dados

- Uma **Tabela de Dados** consiste de n **itens** (instâncias, amostras, observações)

$$(r_1, r_2, r_3, \ldots, r_n)$$

- Cada item é descrito por m $(m \geq 1)$ **atributos**, ou **variáveis**

$$(v_1, v_2, v_3, \ldots, v_m)$$

- Matriz de dados $n \times m$: cada linha descreve um item, cada coluna descreve um atributo relativo a todos os itens

- Em princípio, cada atributo pode ser descrito por um valor único (**número**, **símbolo** ou **cadeia de caracteres**), ou pode ser uma estrutura mais **complexa**

Tabelas de Dados

- Variáveis
  - **Independente** (iv_i): seu valor não é afetado ou controlado por outra variável.
  - **Dependente** (dv_j): seu valor é afetado/determinado por outras variáveis (uma ou mais)

Tabelas de Dados

- Variáveis
  - **Independente** (iv_i): seu valor não é afetado ou controlado por outra variável.
  - **Dependente** (dv_j): seu valor é afetado/determinado por outras variáveis (uma ou mais)

  - Exemplo: um sensor captura vocalizações de passáros ao longo do dia: o tempo é uma v.i., o som capturado é descrito por várias propriedades (intensidade, frequência) que seriam v.d.

Geometria e Grades

- Geometria: cada item de dado tem **coordenadas espaciais** associadas (atributos de geometria)

Geometria e Grades

- Geometria: cada item de dado tem **coordenadas espaciais** associadas (atributos de geometria)

- Atributos espaciais podem ser fornecidos de maneira implícita
  - Itens de dados posicionados em uma **grade** (*grid*)
  - Organização segundo uma estrutura regular, com geometria uniforme ou não uniforme
  - Cada item descrito por um ou mais atributos

Geometria e Grades

- Geometria: cada item de dado tem **coordenadas espaciais** associadas (atributos de geometria)

- Atributos espaciais podem ser fornecidos de maneira implícita
  - Itens de dados posicionados em uma **grade** (*grid*)
  - Organização segundo uma estrutura regular, com geometria uniforme ou não uniforme
  - Cada item descrito por um ou mais atributos

- Posições espaciais referenciadas em um **sistemas de coordenadas**
  - Cartesiano, esférico ou outro

Geometria vs. Topologia

- **Topologia**: define como os itens de dados são **conectados**
  - Caracterização da “vizinhança” de um item de dados
  - Vizinhança definida em **grade**, ou uma **hierarquia** ou um **grafo**
  - Importante para certos tipos de processamento, como **re-amostragem** e **interpolação**
  - Pode ser estabelecida **explicitamente** ou por meio da **estrutura dos dados**

Séries Temporais

- **Atributo Temporal** (*timestamp*): medidas tomadas de maneira uniforme ou não-uniforme
- Podem ocorrer em Tabelas, Geometrias, Grades, ...

Sumário

Tipos de Dados

Supondo cada atributo/variável descrito por um valor único, esse valor pode ser numérico ou nominal

Tipos de Dados

Supondo cada atributo/variável descrito por um valor único, esse valor pode ser numérico ou nominal

- Atributos numéricos
  - **binários**: valores 0 e 1
  - **discretos**: valores inteiros
  - **contínuos**: valores reais

Tipos de Dados

Supondo cada atributo/variável descrito por um valor único, esse valor pode ser numérico ou nominal

- Atributos numéricos
  - **binários**: valores 0 e 1
  - **discretos**: valores inteiros
  - **contínuos**: valores reais

- Atributos nominais
  - **categóricos**: valor em um conjunto finito de possibilidades (ex. vermelho, azul, verde)
  - **ordinais** (ranqueados): valores categóricos com uma ordem (ex. pequeno, médio e grande)
  - **arbitrários**: faixa infinita de valores, sem ordem (ex. endereço)

Tipos de Dados

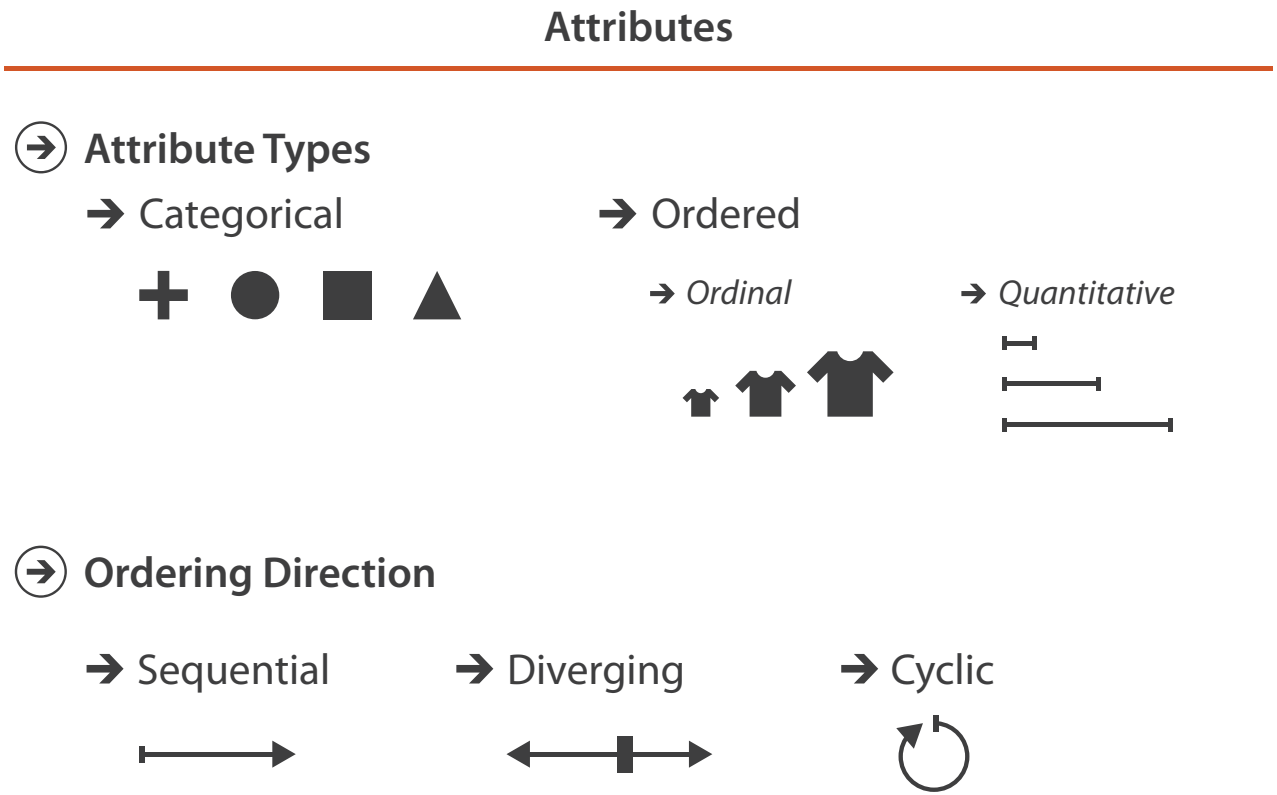

Figura: Tipos de atributos. Fonte: T. Munzner, Visualization Analysis & Design (Fig. 2.7)

Valores numéricos: Escalares, Vetores e Tensores

Escalar

- Um atributo descrito por um único valor numérico é chamado de **escalar**
  - Um item de dado pode ser descrito por múltiplos valores escalares: multidimensional, ou multivariado

Valores numéricos: Escalares, Vetores e Tensores

Escalar

- Um atributo descrito por um único valor numérico é chamado de **escalar**
  - Um item de dado pode ser descrito por múltiplos valores escalares: multidimensional, ou multivariado

Vetor

- Múltiplos valores escalares relacionados definem um **vetor**
  - Vetores tipicamente codificam valores compostos por **direção e magnitude**. P.ex. velocidade, aceleração e força

Escalares, Vetores e Tensores

Tensores

- Escalares e vetores são tipos simples de **tensores**
  - Um tensor é definido por um *rank* e uma dimensão
  - Dimensão determinada pela dimensionalidade do espaço no qual ele está definido
  - Representado como uma matriz, ou vetor

- Uma matriz 3×3 pode representar um tensor de *rank* 2 no espaço 3D

Exemplos de dados estruturados

- **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em grade 3D

Exemplos de dados estruturados

- **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em grade 3D
- **CFD (*computational fluid dynamics*)**: vetor tridimensional (deslocamento), com um atributo temporal e três atributos espaciais, conectividade em uma grade 3D (uniforme ou não-uniforme)

Exemplos de dados estruturados

- **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em grade 3D
- **CFD (*computational fluid dynamics*)**: vetor tridimensional (deslocamento), com um atributo temporal e três atributos espaciais, conectividade em uma grade 3D (uniforme ou não-uniforme)
- **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal

Exemplos de dados estruturados

- **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em grade 3D
- **CFD (*computational fluid dynamics*)**: vetor tridimensional (deslocamento), com um atributo temporal e três atributos espaciais, conectividade em uma grade 3D (uniforme ou não-uniforme)
- **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal
- **Sensoriamento remoto**: múltiplos canais, com dois ou três atributos espaciais, um atributo temporal, conectividade determinada pela grade

Exemplos de dados estruturados

- **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em grade 3D
- **CFD (*computational fluid dynamics*)**: vetor tridimensional (deslocamento), com um atributo temporal e três atributos espaciais, conectividade em uma grade 3D (uniforme ou não-uniforme)
- **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal
- **Sensoriamento remoto**: múltiplos canais, com dois ou três atributos espaciais, um atributo temporal, conectividade determinada pela grade
- **Censo**: Tabela, múltiplos atributos de vários tipos, atributos espaciais e temporais, conectividade determinada pela similaridade dos atributos

Exemplos de dados estruturados

- **MRI (*magnetic resonance imagery*)**: densidade (escalar), com três atributos espaciais, conectividade em grade 3D
- **CFD (*computational fluid dynamics*)**: vetor tridimensional (deslocamento), com um atributo temporal e três atributos espaciais, conectividade em uma grade 3D (uniforme ou não-uniforme)
- **Financeiros**: sem estrutura geométrica, n componentes possivelmente independentes, nominal e ordinal, com atributo temporal
- **Sensoriamento remoto**: múltiplos canais, com dois ou três atributos espaciais, um atributo temporal, conectividade determinada pela grade
- **Censo**: Tabela, múltiplos atributos de vários tipos, atributos espaciais e temporais, conectividade determinada pela similaridade dos atributos
- **Redes sociais**: campos de todos tipos, com vários atributos que podem ser espaciais, temporais e conectividade informada (grafo)

Sumário

Processamento dos Dados

- Em algumas situações, como em aplicações médicas, pode ser preferível visualizar dados “brutos”
- Mas, em geral, é **necessário** algum tipo de preparação, ou **pré-processamento**
- Importante conhecer os dados (distribuição e estatísticas descritivas)
- Processamentos típicos: tratamento de erros, remoção de ruído, normalização, identificação de correlações entre variáveis, identificação de *outliers*, amostragem, interpolação, etc.

Processamento dos dados

- **Metadados**: dados sobre os dados

- Apoiam a interpretação **provendo informação** semântica
  - unidades e resolução das medidas
  - símbolo indicativo de valores faltantes
  - descrição do formato
  - etc.

- Ex. artigo científico: o “dado” é o texto, os metadados são título, autores, data, filiação, páginas, editora, etc.
- Ex. imagem: o “dado” são os valores nos pixels, os metadados são a resolução, número de canais, profundidade de cada canal, etc.

Processamento dos Dados

- **Estatísticas descritivas** , numéricas ou gráficas, podem ser usadas para sumarizar e caracterizar os dados

- Perguntas típicas:
  - como estão distribuídos os valores?
  - ocorrem valores **espúrios**, anômalos, faltantes? (erros de medida/coleta, *outliers*)
  - há **agrupamentos**? (ocorrências de valores similares)
  - existem atributos **correlacionados** (talvez redundantes)? (análise de correlação)
  - ...

Sumarização

Estatísticas de sumarização típicas (para atributos numéricos) são a **média** e **desvio padrão**.

Considerando o j-ésimo atributo da i-ésima instância:

- Média

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n}(x_{ij})$$

- Desvio padrão

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}$$

Sumarização

- Exemplo: distribuição de medidas do peso de pacotes de açucar com (idealmente) 1Kg.

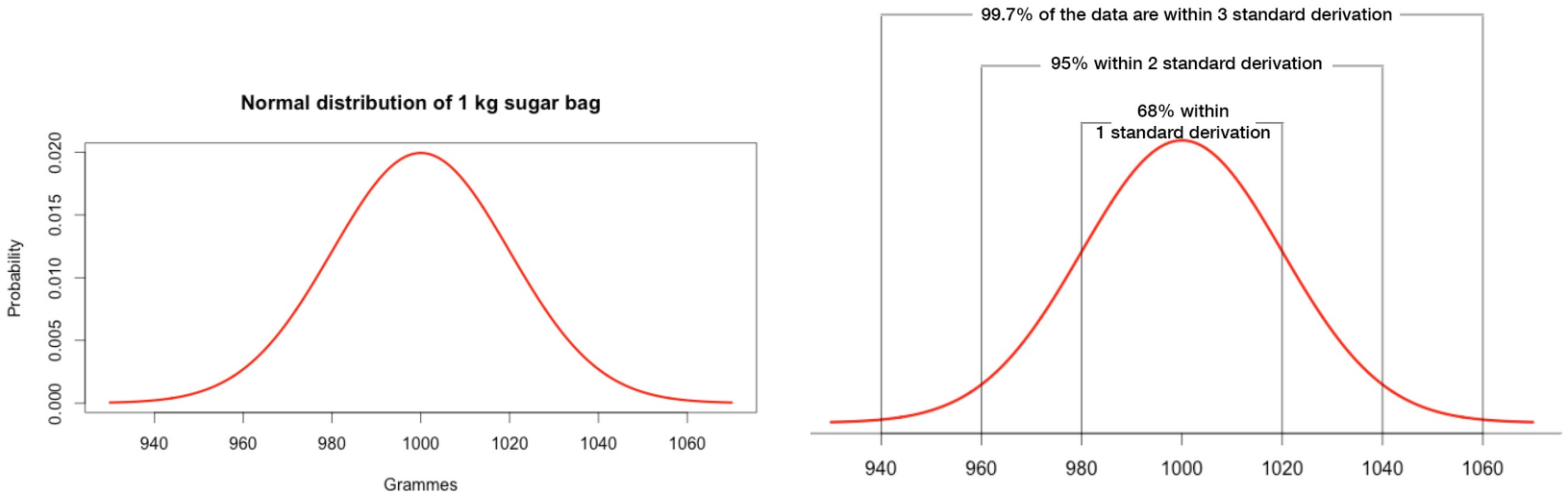

Figura: Distribuição dos pesos observados.

Sumarização

- **Atenção**: média e desvio padrão nem sempre são estatísticas adequadas para sumarizar os dados!

- P.ex., suponha um atributo que descreve os preços, em US$, de 11 produtos: $\{1, 1, 1.5, 0.5, 1, 1, 1, 1, 1, 1, 20\}$

Sumarização

- **Atenção**: média e desvio padrão nem sempre são estatísticas adequadas para sumarizar os dados!
- P.ex., suponha um atributo que descreve os preços, em US$, de 11 produtos: $\{1, 1, 1.5, 0.5, 1, 1, 1, 1, 1, 1, 20\}$
- O preço médio é US$2.73 e o desvio padrão é US$5.46

Sumarização

- **Atenção**: média e desvio padrão nem sempre são estatísticas adequadas para sumarizar os dados!

- P.ex., suponha um atributo que descreve os preços, em US$, de 11 produtos: $\{1, 1, 1.5, 0.5, 1, 1, 1, 1, 1, 1, 20\}$

- O preço médio é US$$2.73$ e o desvio padrão é US$$5.46$
- Entretanto, nenhum dos produtos custa perto desse valor! Tampouco os preços variam entre US$$(2.73 - 5.46)$ e US$$(2.73 + 5.46)$!

Sumarização

- **Atenção**: média e desvio padrão nem sempre são estatísticas adequadas para sumarizar os dados!

- P.ex., suponha um atributo que descreve os preços, em US$, de 11 produtos: $\{1, 1, 1.5, 0.5, 1, 1, 1, 1, 1, 1, 20\}$

- O preço médio é US$2.73 e o desvio padrão é US$5.46
- Entretanto, nenhum dos produtos custa perto desse valor! Tampouco os preços variam entre US$(2.73 − 5.46) e US$(2.73 + 5.46) !

- Neste caso seria mais adequado considerar a mediana e os quartis

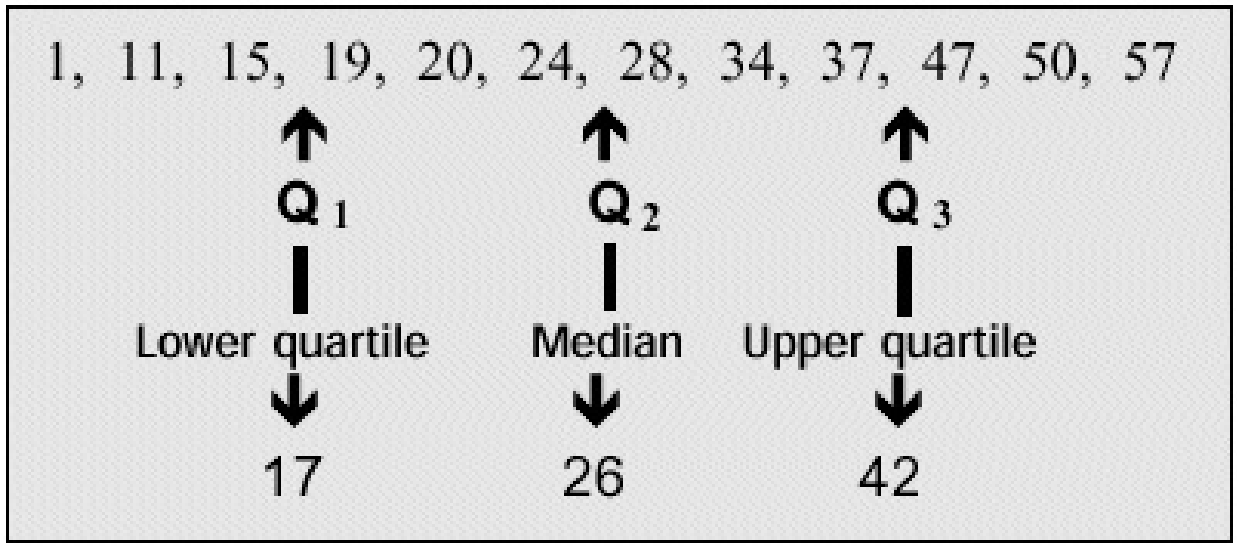

Figura: Mediana e quartis.

`http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214890-eng.htm`

Sumarização

- Histogramas e box plots também são úteis para observar a distribuição dos dados
- `https://asq.org/quality-resources/histogram`
- `https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/box-plot/`

..

Valores Ausentes

Em conjuntos de dados “reais” é comum existirem valores **ausentes** (ou errados)

Valores Ausentes

Em conjuntos de dados “reais” é comum existirem valores **ausentes** (ou errados)

- Estratégias para lidar com o problema da ausência de dados
  - **Descartar** as instâncias **incompletas**: pode implicar em grande perda de informação

Valores Ausentes

Em conjuntos de dados “reais” é comum existirem valores **ausentes** (ou errados)

- Estratégias para lidar com o problema da ausência de dados
  - **Descartar** as instâncias **incompletas**: pode implicar em grande perda de informação
  - Assinalar um valor **sentinela**: cuidado para não incluir a sentinela nos cálculos

Valores Ausentes

Em conjuntos de dados “reais” é comum existirem valores **ausentes** (ou errados)

- Estratégias para lidar com o problema da ausência de dados
  - **Descartar** as instâncias **incompletas**: pode implicar em grande perda de informação
  - Assinalar um valor **sentinela**: cuidado para não incluir a sentinela nos cálculos
  - Calcular um valor **substituto**: atribuição, ou *“data imputation”*

Atribuição de dados

- Dois métodos simples de *"data imputation"* são
  - Atribuir o valor **médio**: desvantagem é que pode mascarar *outliers*

Atribuição de dados

- Dois métodos simples de *"data imputation"* são
  - Atribuir o valor **médio**: desvantagem é que pode mascarar *outliers*
  - Atribuir o valor do **vizinho** mais próximo: requer uma definição de vizinhança

Atribuição de dados

- Dois métodos simples de *"data imputation"* são
  - Atribuir o valor **médio**: desvantagem é que pode mascarar *outliers*
  - Atribuir o valor do **vizinho** mais próximo: requer uma definição de vizinhança
  - Outras estratégias de interpolação

Normalização

Problema

- Em algoritmos que comparam instâncias de dados, duas situações podem distorcer os resultados:
  - Os vetores que representam as instâncias têm normas Euclideanas (magnitudes) muito diferentes (ordens de grandeza)
  - Os atributos assumirem valores em escalas muito distintas

Normalização

Problema

- Em algoritmos que comparam instâncias de dados, duas situações podem distorcer os resultados:
  - Os vetores que representam as instâncias têm normas Euclideanas (magnitudes) muito diferentes (ordens de grandeza)
  - Os atributos assumirem valores em escalas muito distintas

- Usual aplicar um processo de **Normalização**
  - Transformar os valores de modo que satisfaçam alguma **propriedade estatística**

Normalização

- Para evitar o primeiro cenário, pode-se transformar os vetores para que tenham magnitude igual a 1 (norma unitária)
  - $x'_{ij} = x_{ij}/||\mathbf{x}_i||$ para $1 \leq j \leq m$

Normalização

- Outro processo de normalização consiste em transformar os valores para que fiquem dentro do intervalo $[0, 1]$

- Sejam x_{max} e x_{min} os valores máximo e mínimo do intervalo de variação original do $j-$ésimo atributo

- Um valor x_{ij} é transformado da seguinte maneira

$$x'_{ij} = (x_{ij} - x_{min})/(x_{max} - x_{min})$$

Normalização

- Outro processo de normalização consiste em transformar os valores para que fiquem dentro do intervalo $[0, 1]$

- Sejam x_{max} e x_{min} os valores máximo e mínimo do intervalo de variação original do $j-$ésimo atributo

- Um valor x_{ij} é transformado da seguinte maneira

$$x'_{ij} = (x_{ij} - x_{min})/(x_{max} - x_{min})$$

Essa **normalização** pode **distorcer** os dados na presença de valores espúrios

Normalização

- A normalização conhecida como **z-score** ou “**standardization**” consiste em transformar os valores de um atributo de modo que este tenha média 0 e desvio padrão unitário

- `https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9`

Normalização

- A normalização conhecida como **z-score** ou **"standardization"** consiste em transformar os valores de um atributo de modo que este tenha média 0 e desvio padrão unitário

- `https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9`

Dados a média μ_j e o desvio padrão σ_j do $j-$ésimo atributo:

$$x'_{ij} = (x_{ij} - \mu_j)/\sigma_j$$

..

Correlação entre atributos

Atributos correlacionados podem indicar redundância da informação

- Dados dois atributos x_i e x_j, uma medida de correlação é calculada como

$$cor(x_i, x_j) = \frac{cov(x_i, x_j)}{var(x_i) var(x_j)}$$

em que $cov(x_i, x_j)$ indica a covariância entre x_i e x_j

$$cov(x_i, x_j) = \frac{1}{m-1} \sum_{k=1}^{m} (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

e $var(x_i)$ indica a variância de x_i

$$var(x_i) = \sigma_i^2$$

- Valores de $|cor(x_i, x_j)|$ próximos de 1 indicam correlação alta entre x_i e x_j, sugerindo que talvez seja possível descartar um deles

Correlação entre atributos

- Outros coeficientes de correlação
  - Correlação de Pearson: relação linear, positiva ou negativa ($[-1, +1]$)
  - Correlação de Spearman (ρ): não presume relação linear entre as variáveis, é aplicável a variáveis discretas ($[-1, +1]$)
- v. `https://operdata.com.br/blog/coeficientes-de-correlacao`

Redução de Dimensionalidade

- Dados multidimensionais: **dimensionalidade** do dado é maior do que 2 ou 3
- Muitas visualizações implicam em um mapeamento das instâncias de dados na tela (bidimensional)
- Dimensionalidade do dado excede a capacidade da técnica de visualização: redução de dimensionalidade antes de criar a representação visual
- Esse processo deve tentar **preservar**, o máximo possível, a **informação** original

Redução de Dimensionalidade

- Pode ser feita por seleção explícita dos atributos (**feature selection** em Aprendizado de Máquina), ou por meio de alguma **técnica** de redução de dimensionalidade, como
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - Self-Organizing Maps (SOM)
  - Técnicas de Projeção Multidimensional

Redução de Dimensionalidade

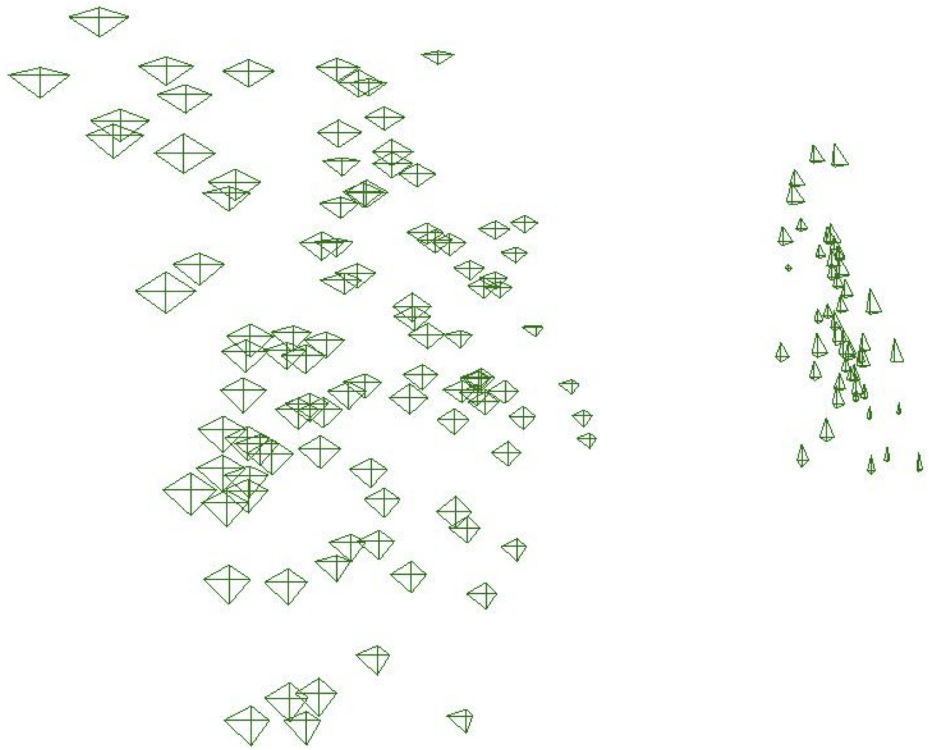

Figura: Mapeamento visual do conjunto de dados Iris usando PCA. Os elementos gráficos (*glyphs*) representam as 4 variáveis originais – cada linha emanando do centro é proporcional ao valor de um dos atributos.

Mapeando Atributos Nominais para Números

- “encoding categorical variables”
- No caso de valores **categóricos ordenáveis** o mapeamento é **direto**, mas o problema é mais **complexo** no caso de valores **não ranqueados**
- Requer estratégias específicas, p.ex., “one-hot encoding”
- ver `https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/`

Observação

- O analista deve saber se (e como) os dados foram processados

Sumário

Distâncias

- A distância ($\delta(\mathbf{x}_i, \mathbf{x}_j)$) entre instâncias de dados multidimensionais $x_i, x_j \in \mathbf{X}$ desempenha papel central em muitas técnicas
- Uma medida da dissimilaridade entre um par de instâncias
- Pode ser calculada de muitas maneiras diferentes

- Distância de *Minkowski* – família de métricas de distância denominadas normas L_p

$$L_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p\right)^{\frac{1}{p}} \tag{1}$$

- Com $p = 1$ obtém-se a distância *Manhattan* (*City Block*)
- Com $p = 2$ tem-se a distância Euclideana
- Com $p = \infty$ obtém-se a distância do infinito ($L_\infty(\mathbf{x}_i, \mathbf{x}_j) = \max_{k=1}^{m} |x_{ik} - x_{jk}|$)

Distâncias

Propriedades de uma Métrica (Distância)

1. **Não-Negatividade:** $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X},\ \delta(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
2. **Identidade:** $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X},\ \mathbf{x}_i = \mathbf{x}_j \Leftrightarrow \delta(\mathbf{x}_i, \mathbf{x}_j) = 0$
3. **Simetria:** $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X},\ \delta(\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_j, \mathbf{x}_i)$
4. **Desigualdade Triangular:** $\forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X},$ $\delta(\mathbf{x}_i, \mathbf{x}_k) \leq \delta(\mathbf{x}_i, \mathbf{x}_j) + \delta(\mathbf{x}_j, \mathbf{x}_k)$

Distâncias

- Nem toda dissimilaridade é uma distância (métrica), i.e., não necessariamente satisfaz as propriedades métricas

- Uma dissimilaridade pode ser definida como o inverso de uma similaridade
- Exemplo: dissimilaridade do cosseno

$$1 - cos(\mathbf{x}_i, \mathbf{x}_j)$$

Distâncias

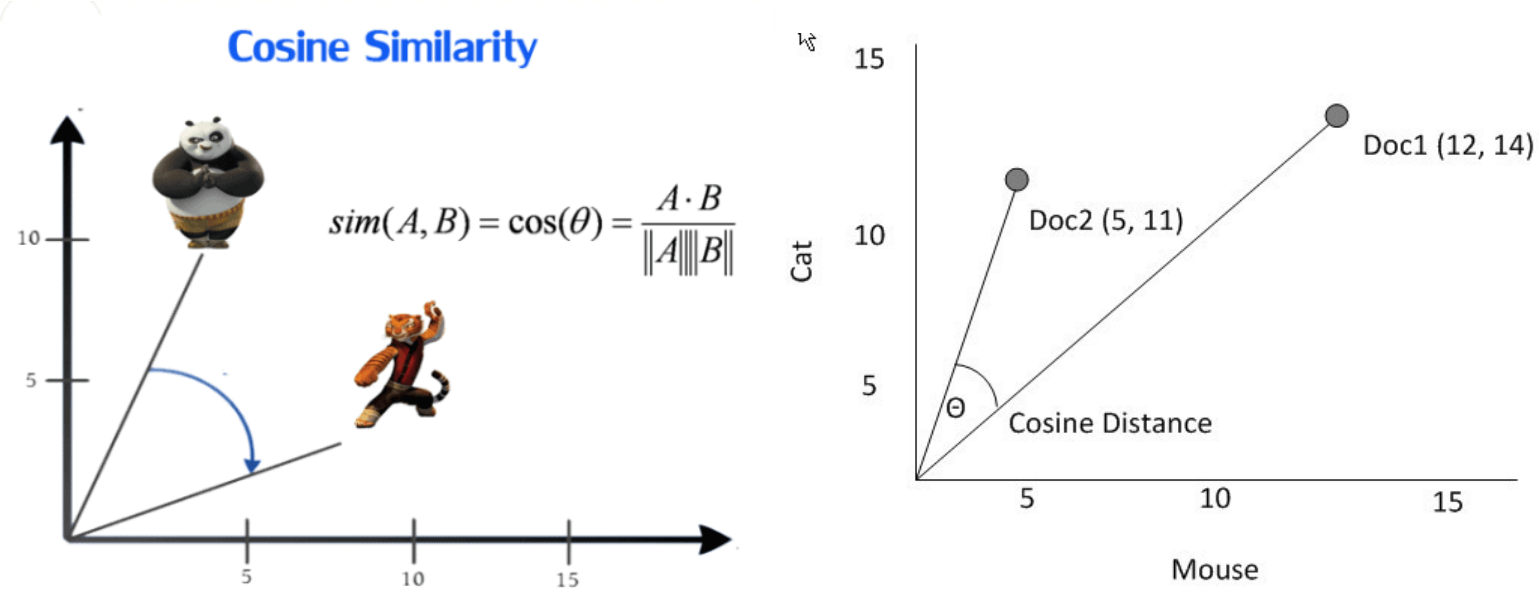

Figura: Distâncias entre pares de instâncias de dados bidimensionais.

Distâncias Binárias

- Dados binários: métricas específicas

- `http://people.revoledu.com/kardi/tutorial/Similarity/BinaryVariables.html`

Distâncias Binárias

- p: número de variáveis c/ valor positivo em ambas as instâncias
- q: número de variáveis c/ valor positivo em uma instância e negativo em outra
- r: número de variáveis c/ valor negativo em uma instância e positivo em outra
- s: número de variáreis c/ valor negativo em ambas as instâncias
- $t = p + q + r + s$: número total de instâncias

- $d_{ij} = (q + r)/t$ (simple matching)
- $d_{ij} = (q + r)/(p + q + r)$ (Jaccard's distance)
- $d_{ij} = (q + r)$ (Hamming distance)
- $d_{ij} = (p + s)/t$ (simple matching coefficient)
- $d_{ij} = p/t$
- $d_{ij} = p/(p + q + r)$ (Jaccard's corfficient)
- $d_{ij} = 2p/(2p + q + r)$
- $d_{ij} = 2(p + s)/(2(p + s) + q + r)$
- $d_{ij} = p/(q + r)$
- $d_{ij} = (p + s)/(q + r)$

Referências

- Ward, M., Grinstein, G. G., Keim, D. **Interactive data visualization foundations, techniques, and applications**. Natick, Mass., A K Peters, 2010, Cap. 2.

- **[Munzner, 2015]** Tamara Munzner, Visualization Analysis & Design, CRC Press, Cap. 2.