Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

Análise de Dados

Prof. Tiago A. Almeida

Dados

- Avanços recentes nas tecnologias de aquisição, transmissão e armazenamento de dados

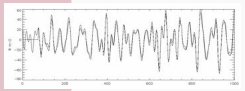
Bases de dados cada vez maiores

Dados

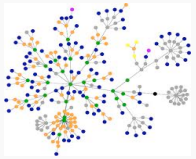
- Estima-se que a quantidade de dados em bases de dados mundiais dobra a cada 20 meses
- Crescimento tem ocorrido em várias áreas
  - Transações bancárias
  - Utilização de cartões de crédito
  - Dados governamentais
  - Medições ambientais
  - Dados clínicos
  - Projetos genoma
  - Informações disponíveis na web
  - etc.

Dados

- Podem ter diferentes formatos



Séries temporais



Grafos

Vídeos

Textos

Páginas web

Áudios

Imagens

Geralmente transformados para o formato atributo-valor

Formato atributo-valor

- Representação de conjunto de dados
  - Formados por objetos
    - Cada objeto corresponde a uma ocorrência dos dados

		Sintomas				
		temperatura	dor		pressão	doente
Objetos	$paciente_1$	38ºC	sim	...	12.7	Sim
	$paciente_2$	36ºC	não	...	12.7	Não
				⋮		
	$paciente_m$	40ºC	não	...	14	Sim

Tiago A. Almeida

Formato atributo-valor

- Cada objeto é descrito por um conjunto de atributos de entrada (Vetor de características)
  - Cada atributo está associado a uma propriedade do objeto

		Sintomas				
		temperatura	dor		pressão	doente
Dados	$paciente_1$	38ºC	sim	...	12.7	Sim
	$paciente_2$	36ºC	não	...	12.7	Não
				⋮		
	$paciente_m$	40ºC	não	...	14	Sim

Tiago A. Almeida

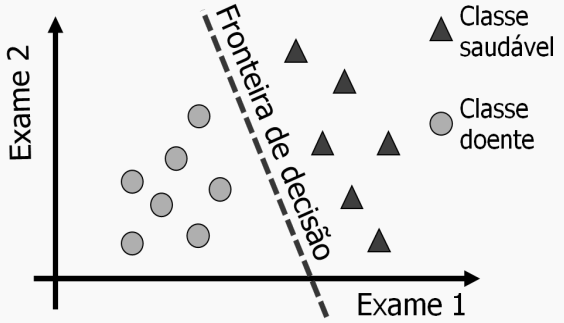
Conjunto de dados

- Pode ser representado por uma matriz de objetos $\mathbf{X}_{m \times n}$
  - m = número de amostras
  - n = número de atributos (excluindo atributo-meta)
    - Dimensionalidade dos objetos
      - Do espaço de objetos (de entradas/de atributos)
  - **Formalização**: amostra $x^{(i)}$ e atributo x_j
    - Elemento $x_j^{(i)}$ (ou x_{ij}) ⇨ valor do j-ésimo atributo para o objeto i

Tiago A. Almeida

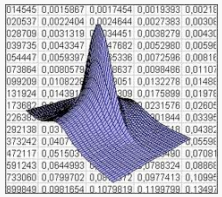
Conjunto de dados: visualização gráfica

- Supor conjunto de pacientes com dois exames



Análise de dados

- Análise das características de um conjunto de dados
  - Muitas podem ser obtidas por fórmulas estatísticas simples
    - Estatística descritiva
  - Análise visual também é importante



Análise de dados

- Caracterização de dados
  - Instâncias e Atributos
  - Tipos de dados
- Exploração de dados
  - Dados univariados
  - Medidas de localidade, espalhamento e distribuição
  - Dados multivariados
  - Visualização

Análise de dados

- Valores de atributos podem ser definidos por:
  - Tipo
    - Grau de quantização nos dados
  - Escala
    - Significância relativa dos valores

Conhecer o tipo/escala dos atributos auxilia a identificar a forma adequada de preparar os dados e posteriormente modelá-los

Tipos de atributos

Quantitativo (numérico)	Qualitativo (simbólico ou categórico)
Representa quantidades	Representa qualidades
Valores podem ser ordenados e usados em operações aritméticas	Valores podem ser associados a categorias
Podem ser contínuos ou discretos	Alguns podem ser ordenados, mas operações aritméticas não são aplicáveis
Possuem unidade associada	*Ex. {pequeno, médio, grande}*

Tipos de atributos

Atributos Quantitativos

Contínuos

- Podem assumir um número infinito de valores
- Geralmente resultados de medidas
- Frequentemente representados por números reais
- *Ex. peso, distância*

Discretos

- Número finito ou infinito contável de valores
- Caso especial: atributos binários (booleanos)
- *Ex. {12, 23, 45}, {0, 1}*

Tipos de atributos

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Qualitativo **Quantitativo discreto** **Quantitativo contínuo**

Tipos de atributos

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Alguns atributos qualitativos são representados por números, mas não faz sentido a utilização de operadores aritméticos sobre seus valores

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
  - Nominais
  - Ordinais
  - Intervalares
  - Racionais

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
  - Nominais
  - Ordinais
  - Intervalares
  - Racionais

Qualitativos

Escala de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
  - Nominais
  - Ordinais
  - Intervalares
  - Racionais

Quantitativos

Escalas de atributos

Escala nominal

- Valores são nomes diferentes e carregam a menor quantidade de informação possível
- Não existe relação de ordem entre os valores
- Operações aplicáveis: =, ≠
- *Ex.: cores, sexo*

Escala ordinal

- Valores refletem ordem das categorias representadas
- Operações aplicáveis: =, ≠, <, >, ≤, ≥
- *Ex.: hierarquia militar, avaliações qualitativas de temperatura*

Escalas de atributos

Escala intervalar

- Números que variam em um intervalo
- É possível definir ordem e diferença em magnitude entre dois valores
- Origem da escala definida de maneira arbitrária
- Operações aplicáveis: =, ≠, <, >, ≤, ≥, +, –
- *Ex.: temperatura em ºC ou ºF, datas*

Escala racional

- Carregam mais informações
- Têm significado absoluto (existe 0 absoluto)
- Razão tem significado
- Operações aplicáveis: =, ≠, <, >, ≤, ≥, +, –, *, /
- *Ex.: tamanho, distância, salário, saldo em conta*

Escalas de atributos

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Nominal **Ordinal** **Intervalar** **Racional**

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: ?
  - Número de palavras de um texto: ?
  - Número de matrícula: ?
  - Data de nascimento: ?
  - Código postal: ?
  - Posição em uma corrida: ?

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: quantitativo racional
  - Número de palavras de um texto: ?
  - Número de matrícula: ?
  - Data de nascimento: ?
  - Código postal: ?
  - Posição em uma corrida: ?

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: quantitativo racional
  - Número de palavras de um texto: quantitativo racional
  - Número de matrícula: ?
  - Data de nascimento: ?
  - Código postal: ?
  - Posição em uma corrida: ?

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: quantitativo racional
  - Número de palavras de um texto: quantitativo racional
  - Número de matrícula: qualitativo nominal
  - Data de nascimento: ?
  - Código postal: ?
  - Posição em uma corrida: ?

Tiago A. Almeida

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: quantitativo racional
  - Número de palavras de um texto: quantitativo racional
  - Número de matrícula: qualitativo nominal
  - Data de nascimento: quantitativo intervalar
  - Código postal: ?
  - Posição em uma corrida: ?

Tiago A. Almeida

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: quantitativo racional
  - Número de palavras de um texto: quantitativo racional
  - Número de matrícula: qualitativo nominal
  - Data de nascimento: quantitativo intervalar
  - Código postal: qualitativo nominal
  - Posição em uma corrida: ?

Tiago A. Almeida

Exercício

- Definir o tipo e escala dos seguintes atributos:
  - Renda mensal: quantitativo racional
  - Número de palavras de um texto: quantitativo racional
  - Número de matrícula: qualitativo nominal
  - Data de nascimento: quantitativo intervalar
  - Código postal: qualitativo nominal
  - Posição em uma corrida: qualitativo ordinal

Exploração de dados

- Estatística descritiva: resumo quantitativo das principais características de um conjunto de dados
  - Muitas medidas podem ser calculadas rapidamente
  - Captura de informações como:
    - Frequência
    - Localização ou tendência central
    - Dispersão ou espalhamento
    - Distribuição ou formato

Informações obtidas podem ajudar na seleção de técnicas apropriadas de pré-processamento e aprendizado

Exploração de dados

Frequência

• Proporção de vezes que um atributo assume um dado valor

• Aplicável a valores numéricos e simbólicos

• *Ex.: 40% dos pacientes têm febre*

Localização, dispersão e distribuição

• Diferem para dados univariados e multivariados
  • *Maioria dos dados em AM é multivariado, mas análises em cada atributo podem fornecer informações valiosas*

• Geralmente aplicados a valores numéricos

Frequência

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Frequência: 25% das manchas são médias

Dados univariados

- Objetos com apenas um atributo
  - Conjunto com m objetos $\mathbf{x} = \{x^1, x^2, ..., x^m\}$

Observação: termo conjunto não tem o mesmo significado do usado em teoria dos conjuntos
Em um conjunto de dados, o mesmo valor pode aparecer mais de uma vez em um atributo

Dados univariados: medidas de localidade

- Definem pontos de referência nos dados
  - Valor "típico", resume os dados

Valores numéricos	Valores simbólicos
• Média • Mediana • Percentil	• Moda: valor mais frequente

Moda

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Moda: Grandes

Média

- Equação:

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x^i$$

Problema: sensível a *outliers*

Bom indicador apenas se valores são distribuídos simetricamente

Mediana

- Passos:
  - Ordenar os valores de forma crescente
  - Calcular a equação:

$$mediana(\mathbf{x}) = \begin{cases} \frac{1}{2}(x^r + x^{r+1}) & \text{se } m \text{ for par } (m = 2r) \\ x^{r+1} & \text{se } m \text{ for ímpar } (m = 2r + 1) \end{cases}$$

Facilita observar se distribuição é assimétrica ou se existem *outliers*

Mediana

- Exemplos:
  - {17, 4, 8, 21, 4}
    - Ordenando: 4, 4, 8, 17, 21
    - Número ímpar de elementos ⇒ mediana = 8
      - Valor do meio na ordenação
  - {17, 4, 8, 21, 4, 15, 13, 9}
    - Ordenando: 4, 4, 8, 9, 13, 15, 17, 21
    - Número par de elementos ⇒ mediana = (9+13)/2 = 11
      - Média dos dois valores do meio na ordenação

Média e mediana

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 26,1
Mediana: 21,5

Média e mediana

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 5
Mediana: 2,5

Média truncada

- Descarta elementos extremos da sequência ordenada de valores
  - Minimizar problemas da média
  - Necessário definir porcentagem
- Passos:
  - Definir porcentagem p
  - Ordenar valores
  - Descartar (p/2)% de valores de cada extremo
  - Calcular a média dos exemplos restantes

Média truncada

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 26,1
Mediana: 21,5
Média truncada (p = 25%): 23,7

Tiago A. Almeida

Média truncada

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 5
Mediana: 2,5
Média truncada (p = 25%): 3,2

Tiago A. Almeida

Exercícios

- Dado o conjunto de dados {1, 2, 3, 4, 5, 80}, calcular:
  - Média
  - Mediana
  - Média truncada com p = 33%

Tiago A. Almeida

Exercícios

- Dado o conjunto de dados {1, 2, 3, 4, 5, 80}, calcular:
  - Média: (1+2+3+4+5+80)/6 = 15,8
  - Mediana: 3+4 / 2 = 3,5
  - Média truncada com p = 33%: (2+3+4+5)/4 = 3,5

Quartis e percentis

- Mediana divide dados ordenados ao meio
  - Quartis e percentis usam pontos de divisão diferentes

Quartis

- Divide em quartos
- 1º quartil (*Q1*) ⇨ valor que tem 25% dos demais valores abaixo dele
- 2º quartil = mediana

Percentil

- Para p entre 0 e 100
- *p*º percentil = *Pp* ⇨ x_i tal que *p*% dos valores observados são menores do que x_i
- *P25* = *Q1*
- *P50* = *Q2* = mediana

Percentil

Algoritmo para cálculo do percentil

Entrada: m valores e percentil p

Saída: valor do percentil

- Ordenar os m valores de maneira crescente
- Calcular k = m * p
- Se k não for inteiro então
  - Arredondar para o próximo inteiro
  - Retornar o valor dessa posição na sequência
- Senão
  - Retornar média entre os valores nas posições k e k+1

Quartil e percentil

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média: 26,1
Mediana: 21,5
Média truncada (p= 25%): 23,7
Q1: 18,5; *Q2*: 21,5; *Q3*: 31
*P*40: 21

Quartil e percentil

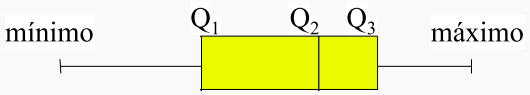
- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Média:5
Mediana: 2,5
Média truncada (p= 25%): 3,2
Q1: 2; *Q2*: 2,5; *Q3*: 5
*P*40: 2

Boxplots

- Também chamados diagramas de Box e Whisker
- Forma gráfica de visualizar quartis
  - Usa quartis e valores máximo e mínimo

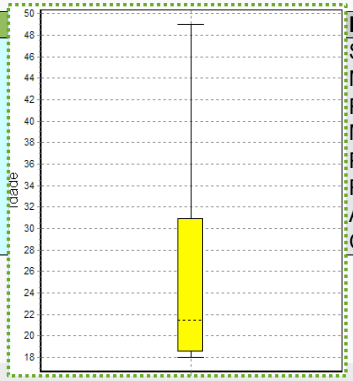


Boxplot modificado: limite superior/inferior vai até maior/menor valor apenas se esse valor não for muito distante do 3º/1º quartil (até 1,5 * intervalo entre quartis Q3 e Q1)
Valores acima/abaixo são considerados *outliers*

Boxplot

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34



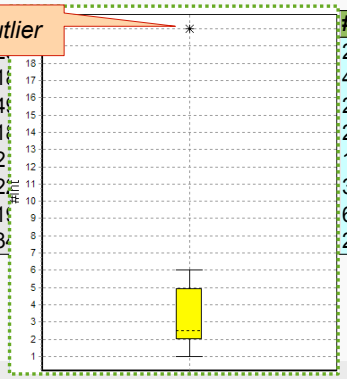
Est.	Diagnóstico
SP	Doente
MG	Doente
RS	Saudável
MG	Doente
PE	Saudável
RJ	Doente
AM	Doente
GO	Saudável

Boxplot modificado

- Ex. conjunto de dados `hospital`

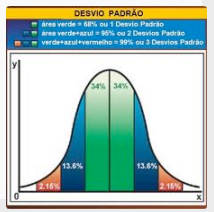
Id.	Nom	# Int.	Est.	Diagnóstico
4201	Joã	2	SP	Doente
3217	Maria	4	MG	Doente
4039	Luiz	2	RS	Saudável
1920	José	20	MG	Doente
4340	Cláudia	1	PE	Saudável
2301	Ana	3	RJ	Doente
1322	Marta	6	AM	Doente
3027	Paulo	2	GO	Saudável

Outlier



Dados univariados: medidas de espalhamento

- Medem dispersão ou espalhamento de um conjunto de valores
  - Permitem observar se valores estão:
    - Espalhados
    - Concentrados em torno de um valor (ex. da média)
  - Medidas mais comuns:
    - Intervalo
    - Variância
    - Desvio padrão



Intervalo

- Mostra espalhamento máximo entre valores
  - Medida mais simples

$$\text{intervalo}(\mathbf{x}) = \max_{i=1,\ldots,m}(x^i) - \min_{i=1,\ldots,m}(x^i)$$

Problema: não é boa medida se maioria dos valores está próxima de um ponto, com um pequeno número de valores extremos

Intervalo

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 31

Intervalo

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 19

Variância e desvio padrão

- Mais utilizadas

$$\text{variância}(\mathbf{x}) = \frac{1}{m\text{-}1} \sum_{i=1}^{m} (x^i - \bar{x})^2$$

$$\text{desvio padrão}(\mathbf{x}) = \sqrt{\text{variância}(\mathbf{x})}$$

Problema: também são distorcidas pela presença de *outliers*

Desvio padrão

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 31
Desvio padrão: 10,8

Tiago A. Almeida

Desvio padrão

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 19
Desvio padrão: 6,3

Tiago A. Almeida

Outras medidas de espalhamento

- Desvio médio absoluto

$$DMA(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} |x^i - \bar{x}|$$

- Desvio mediano absoluto

$$DMedA(\mathbf{x}) = mediana(\{|x^1 - \bar{x}|, ..., |x^m - \bar{x}|\})$$

- Intervalo interquartil

$$IQ(\mathbf{x}) = P75 - P25$$

Tiago A. Almeida

Outras medidas de espalhamento

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 31
Desvio padrão: 10,8
DMA: 8,2
DMedA: 3,5
IQ: 12,5

Outras medidas de espalhamento

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Intervalo: 19
Desvio padrão: 6,3
DMA: 4
DmedA: 1
IQ: 3

Histograma

- Forma gráfica para visualizar distribuição: histograma
  - Divide valores em cestas
    - Valores categóricos: cada valor é uma cesta
    - Valores numéricos: divisão em intervalos contíguos de mesmo tamanho e cada intervalo é uma cesta
  - Para cada cesta, desenha uma barra com altura proporcional ao número de elementos na cesta

Histograma

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade
4201	João	28
3217	Maria	18
4039	Luiz	49
1920	José	18
4340	Cláudia	21
2301	Ana	22
1322	Marta	19
3027	Paulo	34

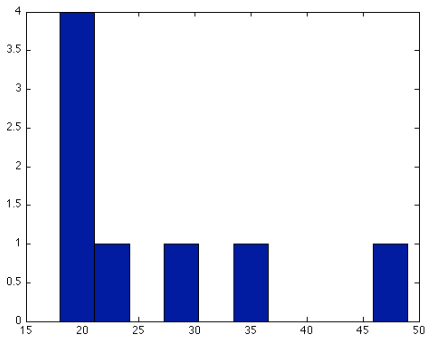


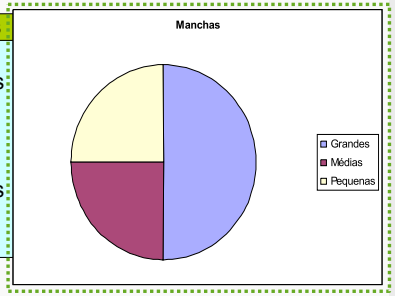
Gráfico de pizza

- Outra forma gráfica de visualizar distribuição de um conjunto de valores
  - Indicado para valores qualitativos
    - Para quantitativos, deve agrupar valores em cestas
  - Cada valor ocupa fatia com área proporcional ao número de vezes que aparece no conjunto de dados

Gráfico de pizza

- Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas
4201	João	28	M	79	Grandes
3217	Maria	18	F	67	Pequenas
4039	Luiz	49	M	92	Grandes
1920	José	18	M	43	Grandes
4340	Cláudia	21	F	52	Médias
2301	Ana	22	F	72	Pequenas
1322	Marta	19	F	87	Grandes
3027	Paulo	34	M	67	Médias



Dados multivariados

- Possuem mais de um atributo de entrada
  - Ex. conjuntos de dados `hospital`
  - Medidas de localidade e espalhamento podem ser calculadas para cada atributo separadamente
    - Ex. média

$$\bar{\mathbf{x}} = (\bar{x}^1, ..., \bar{x}^m)$$

Dados multivariados

- Permitem análises da relação entre dois ou mais atributos
  - Para variáveis contínuas, espalhamento é melhor capturado por uma matriz de covariância
    - Cada elemento é covariância entre dois atributos

$$\text{covariância}(\mathbf{x}^i, \mathbf{x}^j) = \frac{1}{m-1} \sum_{k=1}^{n} (x_k^i - \bar{x}^i)(x_k^j - \bar{x}^j)$$

Observação: covariância($\mathbf{x}^i$, $\mathbf{x}^i$) = variância($\mathbf{x}^i$)

Covariância

- Covariância entre dois atributos mede grau com que variam juntos

Valores de covariância entre dois atributos x^i e x^j:

- **Próximo de 0**: atributos não têm um relacionamento linear
- **> 0 (positiva)**: atributos são diretamente relacionados
- **< 0 (negativa)**: atributos são inversamente relacionados

  - Valor depende da magnitude dos atributos
    - Não é possível avaliar relacionamento de atributos apenas por covariância

Correlação

- Indicação mais clara da força da relação linear entre dois atributos
  - Matriz de correlação: correlação entre todos pares de atributos

$$\text{correlação}(\mathbf{x}^i, \mathbf{x}^j) = \frac{\text{covariância}(\mathbf{x}^i, \mathbf{x}^j)}{\text{desv\_pad}(\mathbf{x}^i) * \text{desv\_pad}(\mathbf{x}^j)}$$

Observação: valores variam de -1 (correlação negativa máxima) a +1 (correlação positiva máxima) e correlação($\mathbf{x}^i$, $\mathbf{x}^i$) = 1

Covariância e correlação

- Ex. conjunto de dados `iris`
  - Matriz de covariância:

	Tamanho_sépala	Largura_sépala	Tamanho_pétala	Largura_pétala
Tamanho_sépala	0,68569	-0,03927	1,27368	0,51690
Largura_sépala	-0,03927	0,18800	-0,32171	-0,11798
Tamanho_pétala	1,27368	-0,32171	3,11318	1,29639
Largura_pétala	0,51690	-0,11798	1,29639	0,58241

  - Matriz de correlação:

	Tamanho_sépala	Largura_sépala	Tamanho_pétala	Largura_pétala
Tamanho_sépala	1,00000	-0,10937	0,87175	0,81795
Largura_sépala	-0,10937	1,00000	-0,42052	-0,35654
Tamanho_pétala	0,87175	-0,42052	1,00000	0,96276
Largura_pétala	0,81795	-0,35654	0,96276	1,00000

Referências

- Ilustrações utilizadas:
  - http://neowayinfo.blogspot.com/2011/05/como-gerenciar-um-grande-volume-de.html
  - http://www.icess.ucsb.edu/gem/filtragem1.htm
  - http://brainstormdeti.wordpress.com/2010/11/06/prova-todo-grafo-completo-e-conexo/
  - http://entomologia.rediris.es/iberodorcadion/Fotos/textos.html
  - http://www.adrformacion.com/cursos/front/leccion1/tutorial3.html
  - http://clipart.usscouts.org/library/
  - http://www.clker.com/clipart-video-camera.html
  - http://www.clker.com/clipart-audio-speaker-1.html
  - http://www.canalexecutivo.com/t533.htm
  - http://intrometendo.com/hierarquia-militar-no-brasil/
  - http://www.sortimentos.com/gente/espaco-profissional-pagamento-13-salario.htm
  - http://fisioterapiahumberto.blogspot.com/2009/12/desvio-padrao-afinal-de-contas-para-que.html
  - http://www.alaska-in-pictures.com/wild-iris-picture-alaskan-summer-8865-pictures.htm
  - http://www.fs.fed.us/wildflowers/beauty/iris/blueflag/iris_virginica.shtml
  - http://www.floweringflowers.net/2010/04/iris/iris-versicolor/

Referências

- Softwares utilizados:
  - Fast Statistics 2.0.4
  - Weka
  - http://www.shodor.org/interactivate/activities/
- Alguns slides são baseados em apresentações de:
  - Prof Dr André C. P. L. F. Carvalho, ICMC-USP