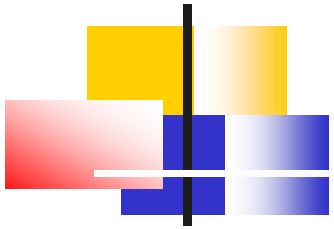
DADOS MULTIVARIADOS

CURSO: APRENDIZADO DE MÁQUINA

1º. SEMESTRE DE 2020

Profa. Roseli A.F Romero

Slides cedidos pelo Prof. André de Carvalho



Variância

- Medida mais utilizada para analisar espalhamento de valores

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - Denominador *n-1*: correção de Bessel, usada para uma melhor estimativa da variância verdadeira
    - Amostra (estimada) e população (verdadeira)
  - Desvio padrão: raiz quadrada da variância
  - Um dos momentos de uma distribuição de probabilidade

VARIÂNCIA

- "o quão longe" em geral os seus valores se encontram do valor esperado (média) da variável aleatória *X.*
- Desvio Padrão indica qual é o "erro" se quiséssemos substituir um dos valores coletados pelo valor da média.

Funcionários	Quantidade de peças produzidas por dia				
	Segunda	Terça	Quarta	Quinta	Sexta
A	10	9	11	12	8
B	15	12	16	10	11
C	11	10	8	11	12
D	8	12	15	9	11

Funcionários	Média Aritmética ($\bar{x}$)	
A	$\bar{X}_A = \frac{10 + 9 + 11 + 12 + 8}{5} = \frac{50}{5}$	$\bar{X}_A = 10,0$
B	$\bar{X}_B = \frac{15 + 12 + 16 + 10 + 11}{5} = \frac{64}{5}$	$\bar{X}_B = 12,8$
C	$\bar{X}_C = \frac{11 + 10 + 8 + 11 + 12}{5} = \frac{52}{5}$	$\bar{X}_C = 10,4$
D	$\bar{X}_D = \frac{8 + 12 + 15 + 9 + 11}{5} = \frac{55}{5}$	$\bar{X}_D = 11,0$

Variância → Funcionário A:

$$\text{var (A)} = \frac{(10 - 10)^2 + (9 - 10)^2 + (11 - 10)^2 + (12 - 10)^2 + (8 - 10)^2}{5}$$

$$\text{var (A)} = \frac{10}{5} = 2,0$$

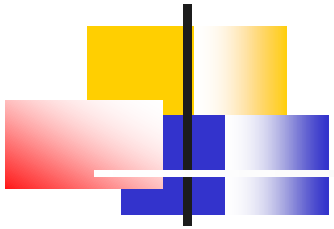
Var(B) = 5,36

Var(C) = 1,84

Var(D) = 6,0

Variância e Desvio Padrão

- **dp(A) ≈ 1,41**
- **dp(B) ≈ 2,32**
- **dp(C) ≈ 1,36**
- **dp(D) ≈ 2,45**
- **Funcionário A: 10,0 ± 1,41 peças por dia**
  Funcionário B: 12,8 ± 2,32 peças por dia
  Funcionário C: 10,4 ± 1,36 peças por dia
  Funcionário D: 11,0 ± 2,45 peças por dia



Dados multivariados

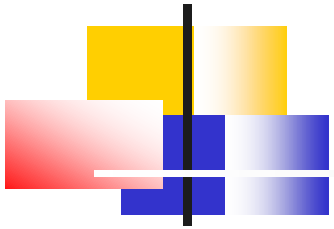
- Cálculo de cada elemento s_{ij} de uma matriz de covariância S para um conjunto de n objetos

$$s_{ij} = \text{covariância}\,(x_i, x_j) = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Onde:
$\overline{x_i}$: Valor médio do i-ésimo atributo
x_{ki} : Valor do i-ésimo atributo para o k-ésimo objeto

É de ordem nxn

- Obs: covariância (x_i, x_i) = variância (x_i)
  - Matriz de covariância tem em sua diagonal as variâncias dos atributos

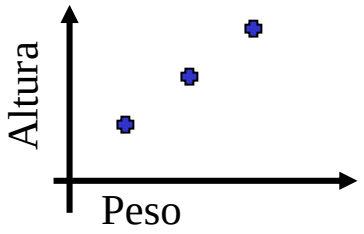


Dados multivariados

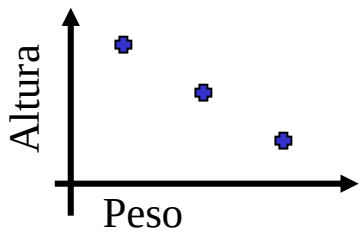
- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos (linearmente)
    - Valor próximo de 0:
      - Atributos não têm um relacionamento linear
    - Valor positivo:
      - Atributos diretamente relacionados
        - Quando o valor de um atributo aumenta, o do outro também aumenta
      - Valor negativo:
        - Atributos inversamente relacionados
    - Valor depende da magnitude dos atributos

Exemplo

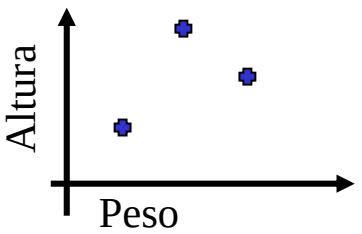
Peso	Altura
60	170
70	180
80	190

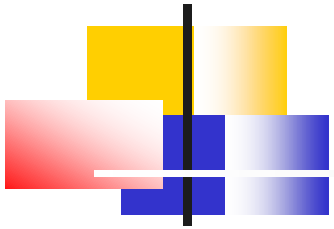


Peso	Altura
60	190
70	180
80	170



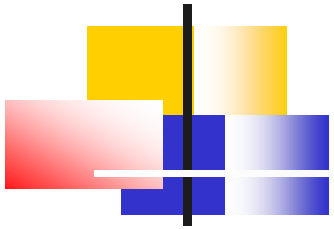
Peso	Altura
60	170
70	190
80	180





Dados multivariados

- Covariância de dois atributos
  - É difícil avaliar o relacionamento entre dois atributos olhando apenas a covariância
    - Sofre influência da faixa de valores dos atributos
    - Correlação linear entre dois atributos ilustra mais claramente a força da relação linear entre eles
      - Mais popular que covariância
      - Elimina influência da faixa de valores



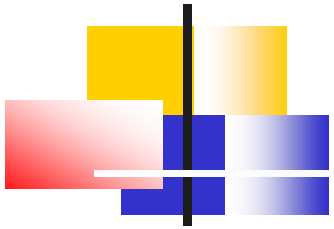
Dados multivariados

- Correlação linear
  - Indica força da relação linear entre dois atributos
  - Matriz de correlação R

$$r_{ij} = \text{correlação}(x_i, x_j) = \frac{\text{covariância}(x_i, x_j)}{s_i s_j}$$

Onde:
x_i: i-ésimo atributo
s_i: Desvio padrão do atributo x_i

  - Obs: correlação $(x_i, x_i) = 1$
    - Elementos da diagonal principal têm valor 1
    - Demais elementos têm valor entre –1 e +1

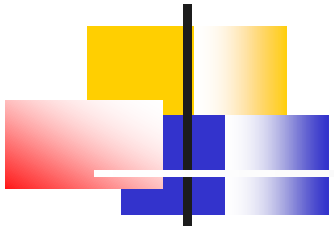


Exercício

- Calcular a matriz de covariância e a matriz de correlação para o conjunto de dados:

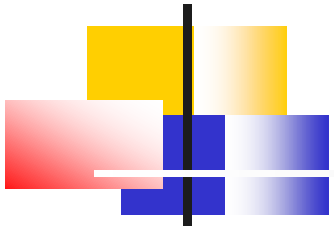
Peso	Altura	Temperatura
73	170	37
67	165	38
90	190	34
49	152	31

  - Ilustrar graficamente pares de atributos direta e inversamente correlacionados



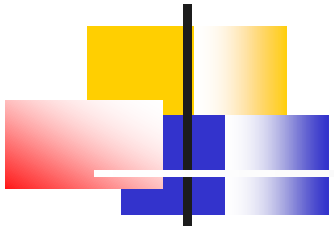
Limpeza

- Correção de problemas detectados nos dados deve lidar com:
  - Atributos com valores ausentes
  - Atributos e objetos redundantes
  - Atributos e objetos com valores inconsistentes
  - Atributos com ruídos
  - *Outliers*



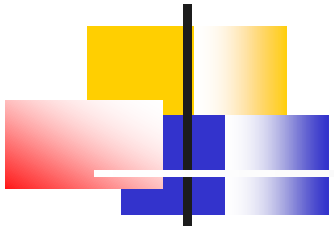
Valores ausentes

- Dados faltosos, faltantes, incompletos
- Várias técnicas de AM não foram projetadas para lidar com valores ausentes
  - Têm dificuldades ou não conseguem induzir um modelo



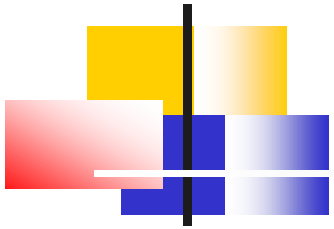
Valores ausentes

- Não é raro um objeto não ter valores para um ou mais atributos
- Possíveis causas:
  - Atributo não foi considerado quando os primeiros dados foram coletados
  - Desconhecimento do valor do atributo por ocasião do preenchimento
  - Distração, mal entendido ou declinação na hora do preenchimento
  - Problema com dispositivo / processo de coleta de valores para o atributo



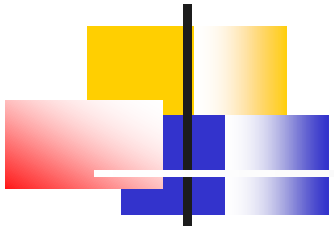
Exemplo de valores ausentes

Nome	Febre	Enjoo	Batimentos	Dor	Salário	Diagnóstico
João	sim	sim	baixo	sim	1000	doente
	não	não	baixo	não	1100	saudável
Maria	sim	sim		não	600	saudável
José	sim	não	baixo	sim		doente
Sérgio	não	não	baixo	não	1100	saudável
Ana	sim	não	alto	sim	1800	saudável
Leila		não	alto		900	doente
Marta	sim	não	baixo	sim	2000	doente



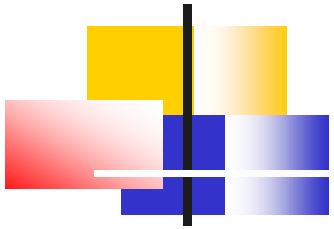
Lidar com valores ausentes

- Agir como se não houvessem valores ausentes
  - Utilizar apenas os valores que estão presentes
    - Ex.: Menos atributos no cálculo da distância entre objetos
  - Modificar algoritmo de AM para lidar com valores ausentes
- Descartar objetos com atributos sem valores
- Preencher valores ausentes



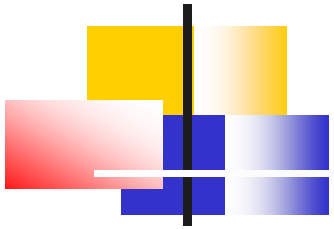
Descarte de objetos

- Geralmente empregado quando:
  - Um dos atributos ausentes é o atributo classe
  - Objeto tem muitos valores ausentes
- Não é indicado quando:
  - Ocorre com poucos atributos do objeto
  - Há risco de perder dados importantes



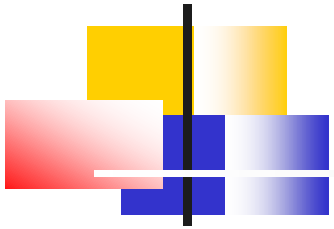
Preenchimento de valor

- Criação de um novo valor que significa ausência
  - Para valores nominais (sem ordem)
- Criação de um novo atributo preditivo
  - Marcando objetos em que um dado atributo tinha valor ausente
- Estimativa de um valor para suprir a ausência



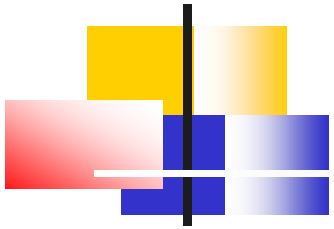
Estimativa do valor

- Usar medida de localidade
  - Média (mediana, moda) dos valores do atributo
    - Todos os valores
    - Dos objetos mais próximos e/ou da mesma classe
  - Para série temporais, medida de localidade entre valores anterior e posterior



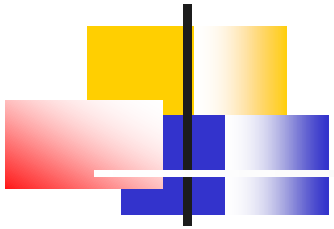
Estimativa do valor

- Induzir valor induzido por algum estimador
  - Valor presente em objetos semelhantes
  - Utilizar algoritmo de AM
  - Alternativa mais eficiente



Valores ausentes

- Observações
  - Em alguns casos, a ausência de valor é uma informação importante sobre o objeto
  - Existem situações em que o valor pode ou precisa estar ausente
    - Ex.: Atributo número do apartamento para quem mora em uma casa
    - Ao invés de ausente, é um valor inexistente
    - Difícil tratar de forma automática
      - Criação de um novo atributo



Exercício

- Tratar dos valores ausentes da tabela abaixo

Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	adimplente
Lia		Superior	200	174	7000	inadimplente
Maria	Advogado	Médio		180	600	adimplente
José	Médico	Superior	100		2000	inadimplente
Sérgio	Bancário		82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	36	2000	inadimplente
José	Médico	Médio	340		800	