Sistemas Inteligentes Aplicados: Pré-processamento dos dados – Escalas

Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Escala dos dados

	igualdade	ordem	soma / subtração	multiplicação / divisão
nominal	✓	✗	✗	✗
ordinal	✓	✓	✗	✗
intervalar	✓	✓	✓	✗
razão	✓	✓	✓	✓

Escala dos dados (2)

	igualdade	ordem	soma / subtração	multiplicação / divisão
nominal	✓	✗	✗	✗
ordinal	✓	✓	✗	✗
intervalar	✓	✓	✓	✗
razão	✓	✓	✓	✓

- **Nominal** (ou categórica): cidade, profissão, cor favorita, etc
- Podemos comparar a **igualdade**: a profissão de Maria é a mesma de João?
- Não podemos **ordenar**: Bombeiro vem antes ou depois de professor? Vermelho < amarelo?

Escala dos dados (3)

	igualdade	ordem	soma / subtração	multiplicação / divisão
nominal	✓	✗	✗	✗
ordinal	✓	✓	✗	✗
intervalar	✓	✓	✓	✗
razão	✓	✓	✓	✓

- **Ordinal**: notas (A, B, C, D, E), dia da semana, tamanho (pequeno, médio e grande), etc
- Podemos comparar a **igualdade** e **ordenar**: a nota de Mateus é maior que a de Antônio?
- Não podemos **somar**: A + B? E – E?

Escala dos dados (4)

	igualdade	ordem	soma / subtração	multiplicação / divisão
nominal	✓	✗	✗	✗
ordinal	✓	✓	✗	✗
intervalar	✓	✓	✓	✗
razão	✓	✓	✓	✓

- **Intervalar**: temperatura (em Celsius), século (XIX, XX, XXI, etc)
  - Obs: temperatura se enquadra em outras categorias
- Podemos comparar a **igualdade**, **ordenar** e **somar**: 20 graus + 10 graus = 30 graus
- Não podemos **multiplicar**: o zero é relativo

Escala dos dados (5)

	igualdade	ordem	soma / subtração	multiplicação / divisão
nominal	✓	✗	✗	✗
ordinal	✓	✓	✗	✗
intervalar	✓	✓	✓	✗
razão	✓	✓	✓	✓

- **Razão**: idade, peso, altura (em metros), temperatura (em Kelvins), entre outras
- Podemos comparar a **igualdade**, **ordenar**, **somar** e **multiplicar**
- Aqui o zero é absoluto

Escala dos dados (6)

- **Qualitativos**: nominal e ordinal
  - Também podem ser chamados de simbólicos
- **Quantitativos**: intervalar e razão
  - Também podem ser chamados de numéricos

Escala dos dados (6)

- Alguns algoritmos só reconhecem atributos qualitativos
- Outros só reconhecem quantitativos
- Conversões
  - Quantitativos → qualitativos (ordinal)
  - Qualitativo → quantitativo (intervalar)
  - Todos → (0, 1) ou (-1, +1)

Escala dos dados (7)

- Nominais de 2 valores → intervalar
  - Eleitor: **não**, **sim** → **0** (não), **1** (sim)
- Nominais de 3 ou mais valores → intervalar (**binarização**)
  - Esporte: **basquete**, **futebol** ou **polo**
  - **joga_basquete** (0 ou 1),
  - **joga_futebol** (0 ou 1),
  - **joga_polo** (0 ou 1)
  - Atributos novos são mutualmente exclusivos

Escala dos dados (8)

- Convertendo: ordinais → intervalar
  - **Segunda**, **terça**, **quarta**, … → **1**, **2**, **3**, …
  - **A**, **B**, **C**, … → **5**, **4**, **3**, …
- Intervalar → ordinal
  - **0-10** graus celsius: **muito_frio**
  - **10-20** graus: **frio**
  - **20-25** graus: **agradável**

Escala dos dados (9)

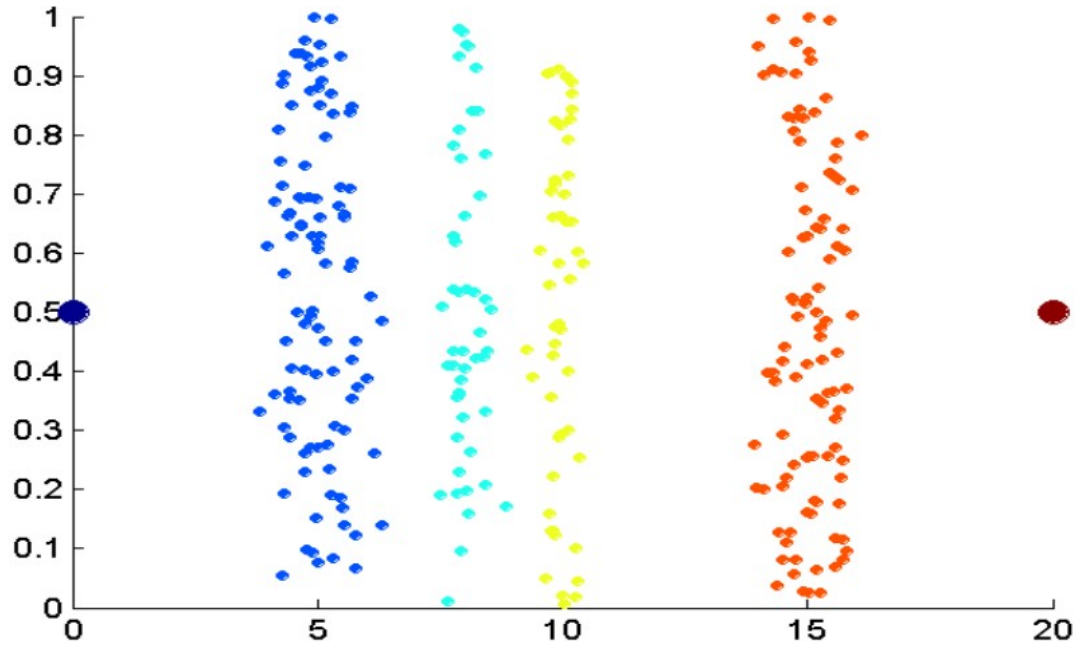
- Razão → ordinal
  - **Até 20** palavras: **pequeno**
  - **Até 200** palavras: **médio**
  - **Mais de 200** palavras: **grande**
- Todas → [0, 1]
  - Converter para números primeiro, se necessário
  - Caso 1 – **regra de três**: maior vira 1, menor vira 0
  - Caso 2 – **distribuição normal** (gaussiana): subtrair a média e dividir pelo desvio padrão + ajustes

Discretização

- Em alguns casos, é difícil definir valores qualitativos ao converter atributos numéricos
- Soluções:
  - Análise gráfica manual
  - Algoritmos de discretização
    - Geralmente projetista precisa definir o número de categorias manualmente (pode ser um processo de tentativa e erro)

Discretização (2)

- **Análise gráfica manual**
  - Exemplo para dois atributos com 4 grupos e 2 outliers



Discretização (3)

- **Algoritmos de discretização**
  - Vários exemplos: k-means, IRD entre outros
  - Dois tipos:
    - Supervisionados
    - Não supervisionados

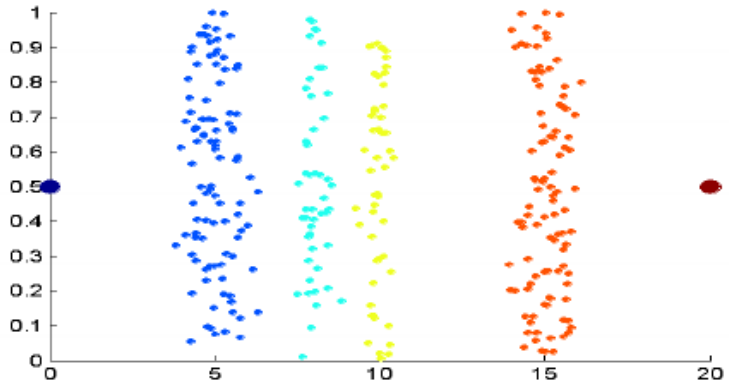
Discretização (4)

- **Algoritmos de discretização supervisionados**
  - Têm acesso a exemplos rotulados
  - Por exemplo: 5º = muito frio; 15º = frio; 25º = agradável; …
  - Tentar achar pontos de corte que maximizem a pureza dos dados
  - Limitação: pontos que são bem separados em duas ou mais dimensões podem não ser em uma

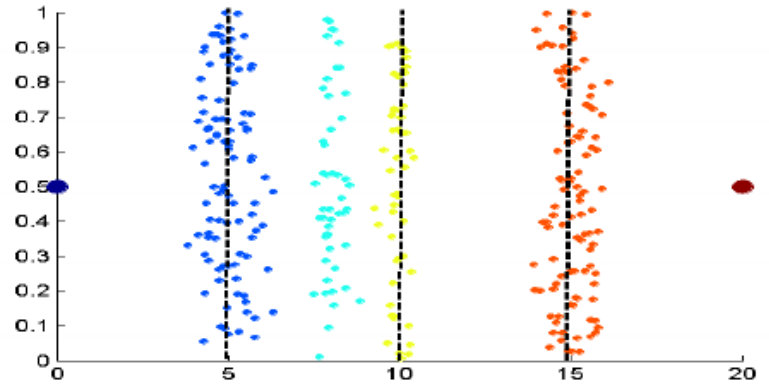
Discretização (5)

- **Algoritmos de discretização não supervisionados**
  - Algoritmos mais simples
  - Larguras iguais: dividir intervalo dos valores pelo número de grupos desejados
  - Frequência: mesmo número de instâncias em cada grupo
  - Aplicar algoritmo de agrupamento

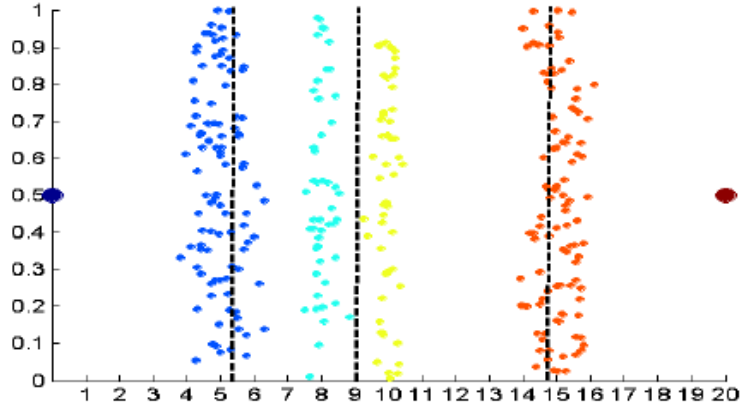
Discretização (6)



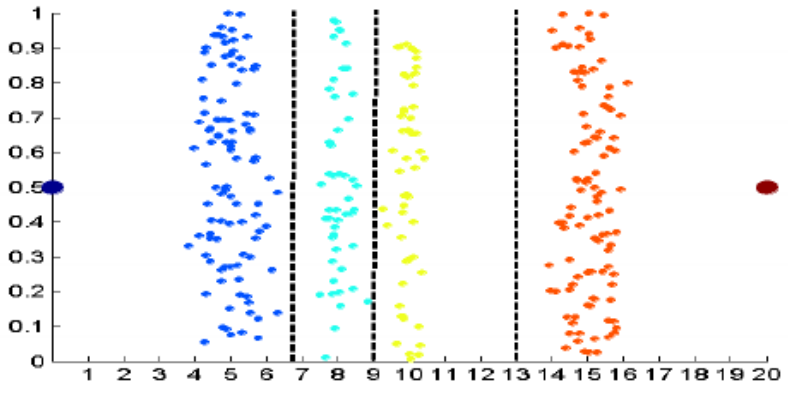
Dados



Mesma largura



Mesma freqüência



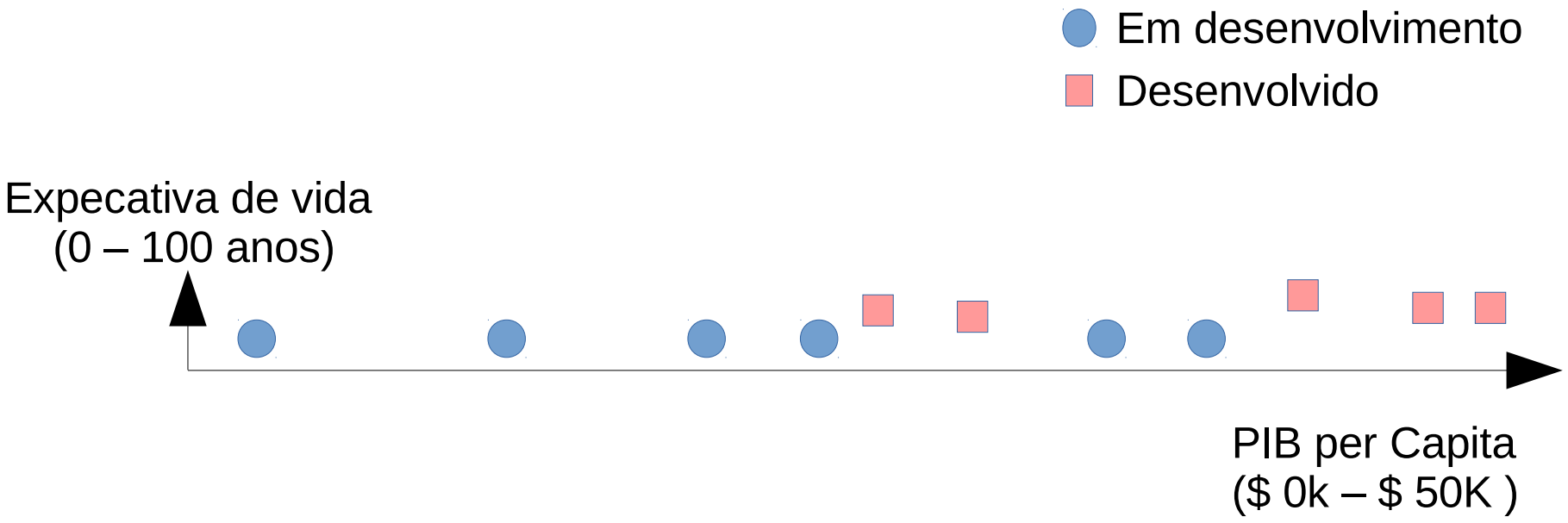
K-médias

Atributos quantitativos: normalização

- Atributos normalmente não tem intervalos de valores parecidos
- Um atributo cujos valores variam muito se sobressai sobre os demais
  - Distâncias euclidianas o favorecem

Atributos quantitativos: normalização (2)

- Atributos normalmente não tem intervalos de valores parecidos
- Ex.: expectativa de vida e PIB per Capita



Atributos quantitativos: normalização (3)

- PIB per Capta é uma medida imperfeita quando a desigualdade social é alta
  - Pequenas variações nele alteram bastante as distâncias entre as instâncias
  - Enquanto grandes variações na expectativa de vida praticamente não afetam as distâncias
- Eixo x_1 varia muito mais que eixo x_2:

  x_1 **domina** ou **mascara** x_2

Atributos quantitativos: normalização (4)

- Solução: transformar os eixos para que variações sejam similares
  - Opção 1: normalizar entre zero e um
    - Maior valor vale 1; menor vale 0; demais são obtidos por regra de três
  - Opção 2: normalizar por média e desvio padrão
    - Forçar dados a terem média 0 e desvio 1

Atributos quantitativos: normalização (5)

- Outras opções:
  - ln(x), log(x)
  - 1/x, $\sqrt{x}$
  - |x|, seno(x)
- Log é particularmente interessante para variações muito altas
  - Ex.: aplicado a faturamento bruto, agrupa empresas em pequenas, médias e grandes

Atributos quantitativos: normalização (6)

- Antes de aplicar uma transformação
  - Verificar se a ordem precisa ser mantida:
    1/x: (1, 2, 3, 4) → (1, 0.5, 0.33, 0.25)
  - O que acontece no intervalo entre 0 e 1?
    $\sqrt{0.25} = 0.5$

Conversões avançadas

- Atributos nominais com muitos valores:
  - São custosos
  - Abordagem tradicional leva a vetores muito esparsos
  - Solução: usar conhecimento do domínio e aplicar atributos alternativos

Conversões avançadas (2)

- Exemplo: atributo país com 195 possíveis valores
- Usar 7 atributos alternativos
  - Continente (nominal): 7 valores possíveis
  - PIB: 1 valor
  - População: 1 valor
  - IDH: 1 valor
  - Temperatura média anual: 1 valor

Conversões avançadas (3)

- Tradução: alguns formatos de dados podem ser difíceis de lidar para alguns algoritmos
  - Data: converter para inteiro
  - Hora: converter para inteiro
  - Rua: converter para CEP
  - Etc

Exercícios

- 1. Converter atributos abaixo no intervalo [0, 1]

Febre	Enjôo	Mancha	Dor	Diagnóstico
baixa	sim	pequena	A	doente
média	não	média	C	saudável
alta	sim	grande	B	saudável
alta	não	pequena	A	doente
baixa	não	grande	D	saudável
média	não	sem	C	doente

Exercícios

- 2. Para a base Iris:
  - (a) aplicar discretização não-supervisionada
  - (b) aplicar discretização supervisionada
  - (c) normalizar dados entre [0, 1] via regra de três (**normalize**)
  - (d) normalizar dados entre [-3, 3] via distribuição normal (**standardize**)

Pontos chaves

- Conversão entre escalas
- Discretização
- Conversões avançadas

Agradecimentos/referências

- Notas de aula do Prof. André de Carvalho (USP)