TRANSFORMAÇÃO DE DADOS
TRANSFORMAÇÃO DE ATRIBUTOS NUMÉRICOS

Cristiane Neri Nobre

Transformação de atributos numéricos

- Algumas vezes, o **valor numérico** de um atributo precisa ser transformado em **outro valor numérico**
- Isso ocorre quando os **limites inferior e superior de valores dos atributos são muito diferentes**, o que leva a uma grande variação de valores, ou ainda quando vários **atributos estão em escalas diferentes**
- Essa transformação é geralmente realizada para evitar que um atributo predomine sobre outro
- Quando necessário, a operação de **transformação** é aplicada aos valores de um dado atributo de todas as instâncias

Transformação de atributos numéricos

- Uma transformação que é muito utilizada é a **normalização** de dados.
- A **normalização** de dados é recomendável quando os limites de valores de atributos distintos são muito diferentes, para evitar que um atributo predomine sobre outro
- Pode-se utilizar a **normalização por amplitude**
  - A normalização por **amplitude** pode ser por **reescala** ou **padronização**.

Transformação de atributos numéricos

- A normalização **por reescala** define uma nova escala de valores, limites mínimo e máximo, para todos os atributos.
  - Também chamada de normalização min-max

- As operações são realizadas para cada atributo.

$$v_{Novo} = \min + \frac{v_{Atual} - \text{menor}}{\text{maior} - \text{menor}}(\max - \min)$$

Para que os limites superior e inferior sejam 1 e 0, respectivamente, basta fazer max=1 e min=0.

Transformação de atributos numéricos

- **Exemplo:**

$$v_{Novo} = \min + \frac{v_{Atual} - \text{menor}}{\text{maior} - \text{menor}}(\max - \min)$$

Peso	Novo valor
10	**0**
80	0,24
150	0,48
300	**1**
30	0,068

$$vnovo = \frac{10-10}{300-10} = \frac{0}{290} = 0$$

$$vnovo = \frac{80-10}{300-10} = \frac{70}{290} = 0{,}24$$

$$vnovo = \frac{150-10}{300-10} = \frac{140}{290} = 0{,}48$$

$$vnovo = \frac{300-10}{300-10} = \frac{290}{290} = 1$$

$$vnovo = \frac{30-10}{300-10} = \frac{20}{290} = 0{,}068$$

Transformação de atributos numéricos

- Para a normalização **por padronização (fórmula Zscore)**, a cada valor do atributo a ser normalizado é adicionada ou subtraída uma medida de localização e o valor resultante é sem seguida multiplicado ou dividido por uma medida de escala

- Com isso, diferentes atributos podem ter limites inferiores e superiores diferentes, mas terão os mesmos valores para as medidas de escala e espalhamento.

- Se as medidas de localização e de escala forem a média (μ) e o desvio padrão (σ), respectivamente, os valores de um atributo são convertidos para um novo conjunto de valores com **média 0 e desvio padrão 1**

$$v_{Novo} = \frac{v_{Atual} - \mu}{\sigma}$$

  - Geralmente, é preferível padronizar a reescalar, pois a padronização lida melhor com *outliers*

Transformação de atributos numéricos

- **Exemplo:**

$$v_{Novo} = \frac{v_{Atual} - \mu}{\sigma}$$

Peso	Novo valor
10	**-0,8875**
80	-0,2901
150	0,3072
300	**1,5874**
30	-0,7169

μ=114 e σ =117,1751

$$vnovo = \frac{10 - 114}{117,1751} = \frac{-104}{117,1751} = -0.8875$$

$$vnovo = \frac{80-114}{117,1751} = \frac{-34}{117,1751} = -0.2901$$

$$vnovo = \frac{150-114}{117,1751} = \frac{36}{117,1751} = 0.3072$$

$$vnovo = \frac{300-114}{117,1751} = \frac{186}{117,1751} = 1.5874$$

$$vnovo = \frac{30 - 114}{117,1751} = \frac{-84}{117,1751} = -0.7169$$

Transformação de atributos numéricos

- De uma maneira geral, **se a distribuição não é Gaussiana ou o desvio padrão é muito pequeno**, normalizar os dados é uma escolha a ser tomada.
- Muitos artigos falarão que normalizar é melhor que padronizar!
- E muitos outros artigos falarão o contrário
- Lembre-se "**Não existe almoço grátis (**No free lunch theorem**)**"

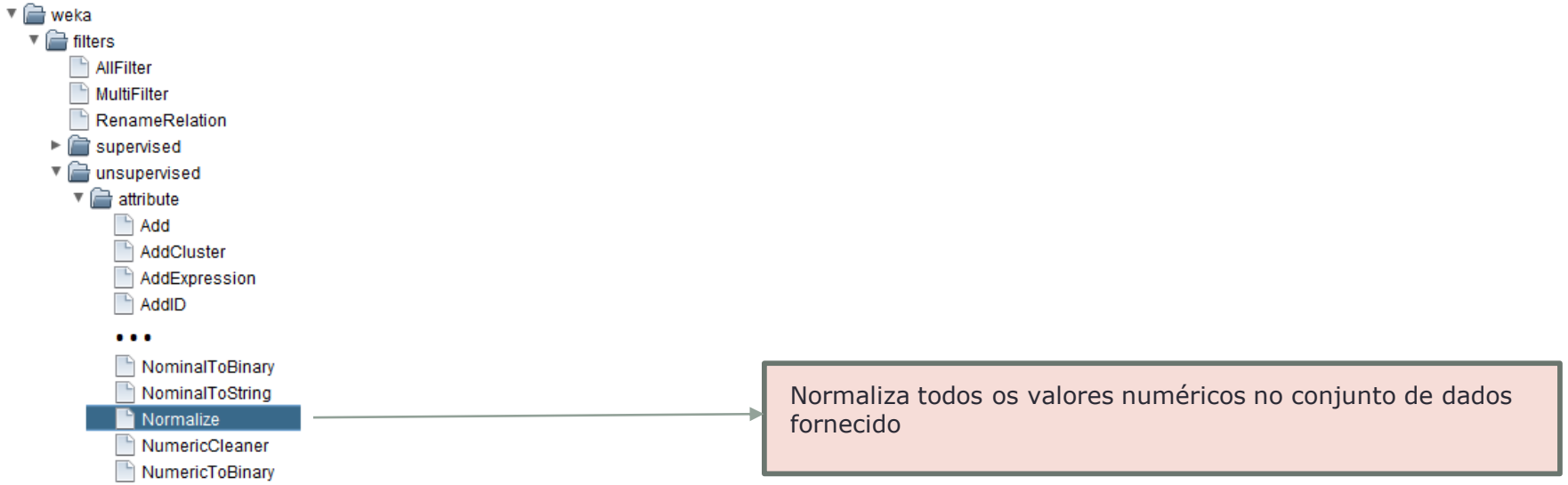
Então sugiro que você teste para a sua base de dados!!

Transformação de atributos numéricos

Como realizar o **normalização** dos atributos no ambiente WEKA?

- Carregue o arquivo: weather.numeric.arff que fica na pasta **Data** onde o WEKA está instalado
  - Veja que os atributos **temperatura** e **umidade** assumem valores inteiros

- Na tela principal do WEKA, vá até a opção **weka/Filters/unsupervised/atribute/Normalize**



Transformação de atributos numéricos

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Base Original

$$vnovo = \frac{80-64}{85-64} = \frac{16}{21} = 0.7619$$

Atributos normalizados

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	1.0	0.64516...	FALSE	no
2	sunny	0.76190476...	0.80645...	TRUE	no
3	overcast	0.90476190...	0.67741...	FALSE	yes
4	rainy	0.28571428...	1.0	FALSE	yes
5	rainy	0.19047619...	0.48387...	FALSE	yes
6	rainy	0.04761904...	0.16129...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny	0.38095238...	0.96774...	FALSE	no
9	sunny	0.23809523...	0.16129...	FALSE	yes
10	rainy	0.52380952...	0.48387...	FALSE	yes
11	sunny	0.52380952...	0.16129...	TRUE	yes
12	overcast	0.38095238...	0.80645...	TRUE	yes
13	overcast	0.80952380...	0.32258...	FALSE	yes
14	rainy	0.33333333...	0.83870...	TRUE	no

Referências:

Capítulo 3 do livro (Seção 3.6.3)

- Katti Faceli et al. Inteligência Artificial, Uma abordagem de Aprendizado de Máquina, LTC, 2015.

Artigo:

- https://arxiv.org/ftp/arxiv/papers/1503/1503.06462.pdf