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 em 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

Investigue normalização no código (está no CANVAS)

Lendo_e_tratando_arquivo_IRIS.ipynb

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