INSTITUTO FEDERAL
Ceará

Programa de Pós-Graduação
em Ciência da Computação

05 Normalização de Dados

APRENDIZAGEM PROFUNDA

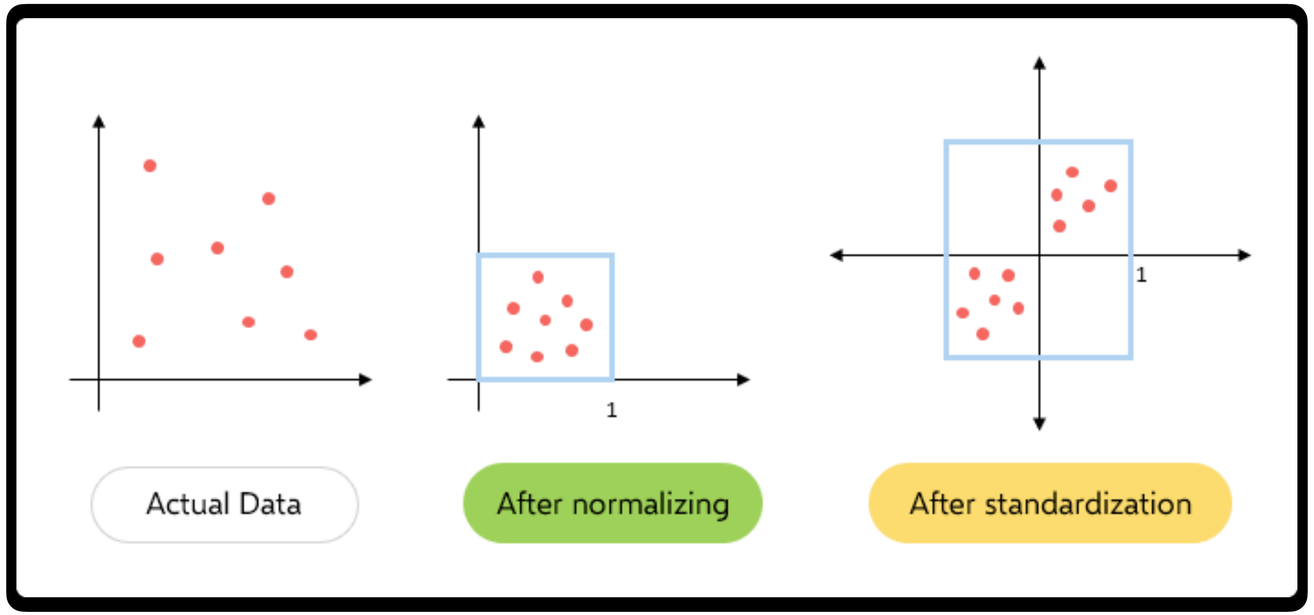
PPGCC – 2023.1

Prof. Saulo Oliveira <saulo.oliveira@ifce.edu.br>

ESCALAS DIFERENTES PODEM SER UM PROBLEMA

NORMALIZAR OS DADOS

O objetivo da normalização é alterar os valores das colunas numéricas no conjunto de dados para uma escala comum, sem distorcer as diferenças nos intervalos de valores.



https://www.someka.net/blog/how-to-normalize-data-in-excel/

Maneiras de realizá-la:

1. Manter constante a norma dos vetores;
2. Mudança da escala original para os intervalos $[0, 1]$ ou$[-1, +1]$;
3. Padronização dos dados (i.e. $\mu = 0$, $\sigma^2 = 1$).

NORMALIZAÇÃO POR NORMA CONSTANTE

NORMALIZAÇÃO POR NORMA CONSTANTE

- Uma das técnicas mais simples de normalização consiste em manter constantes e iguais a 1 as normas dos vetores de atributos de $\mathbf{x}$. Para isso, basta dividir cada vetor por sua respectiva norma euclidiana, i.e., $\tilde{\mathbf{x}} = \dfrac{\mathbf{x}}{\| \mathbf{x} \|}$.

- Por exemplo, considere o seguinte vetor $\mathbf{x}$, que não possui norma unitária $\mathbf{x} = \left[\sqrt{3},\ 3,\ -2\right]^{\mathsf{T}}$.
- A norma de $\mathbf{x}$ é calculada como $\| \mathbf{x} \| = \sqrt{\left(\sqrt{3}\right)^2 + 3^2 + (-2)^2} = \sqrt{16} = 4.$

- Assim, a versão normalizada pela norma é dada por $\tilde{\mathbf{x}} = \frac{1}{4}\left[\sqrt{3},\ 3,\ -2\right]^{\mathsf{T}} = \left[\frac{\sqrt{3}}{4}, \frac{3}{4}, \frac{-2}{4}\right]^{\mathsf{T}}.$

  $\tilde{\mathbf{x}} = \left[0.43,\ 0.75,\ -0.50\right]^{\mathsf{T}}.$
- Quanto vale $\| \tilde{\mathbf{x}} \|$?

NORMALIZAÇÃO POR NORMA CONSTANTE

- Não se altera a direção de $\mathbf{x}$, apenas muda-se seu comprimento, i.e., o vetor resultante $\tilde{\mathbf{x}}$ é um múltiplo de $\mathbf{x}$ conforme pode ser visto na operação a seguir $\tilde{\mathbf{x}} = \frac{1}{\| \mathbf{x} \|}\mathbf{x} = \alpha\mathbf{x}$, em que $\alpha = \frac{1}{\| \mathbf{x} \|}$, é uma constante positiva;

- Observer também que a normalização assim realizada depende apenas dos valores das componentes do vetor sendo normalizado;

- Assim, chamaremos este tipo de procedimento de *normalização local*;

- É particularmente útil para classificadores de máxima correlação.

NORMALIZAÇÃO POR MUDANÇA DE ESCALA

NORMALIZAÇÃO POR MUDANÇA DE ESCALA

- Para classificadores com base em distância euclidiana, uma normalização que promove uma mudança na escala dos atributos, é mais comum;
- O procedimento se dá por atributo e requer a determinação do valor mínimo ($x_{\min}$) e do valor máximo ($x_{\max}$) do atributo a ser normalizado;
- Por isso (detectar os valores máximos e mínimos), chamaremos este procedimento de normalização global;
- No próximo slide, utilizaremos a notação em função de um único atributo.

NORMALIZAÇÃO POR MUDANÇA DE ESCALA

Iris (150 rows)

Detail | Compact | Column

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1 – 150	4.3 – 7.9	2 – 4.4	1 – 6.9	0.1 – 2.5	3 unique values
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2		
16	5.7	4.4	1.5		

https://www.kaggle.com/uciml/iris

SepalLengthCm

Length of the sepal (in cm)

4.3 – 7.9

Valid	150	100%
Mismatched	0	0%
Missing	0	0%
Mean	5.84	
Std. Deviation	0.83	
Quantiles	4.3	Min
	5.1	25%
	5.8	50%
	6.4	75%
	7.9	Max

- Para atributos no intervalo $[0,\ 1]$, normalize o atributo de acordo com:

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

- Para atributos no intervalo $[-1, +1]$, normalize conforme a fórmula que segue:

$$\tilde{x} = 2\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) - 1$$

NORMALIZAÇÃO POR PADRONIZAÇÃO

NORMALIZAÇÃO POR PADRONIZAÇÃO

- Assim como as normalizações descritas na normalização por mudança de escala, devemos aplicar a padronização aos atributos do problema, um a um;

- No entanto, este tipo de normalização requer o cálculo da média $\bar{x}$ e do desvio-padrão σ_x do atributo em questão x;

- Por isso, a padronização também pode ser chamada de normalização estatística;

- Este procedimento também é um tipo de normalização global.

NORMALIZAÇÃO POR PADRONIZAÇÃO

A normalização por padronização é dada por

$$\tilde{x} = \frac{x - \bar{x}}{\sigma_x},$$

com a média e desvio-padrão amostrais de x calculados como

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{e} \qquad \sigma_x = \sqrt{\frac{1}{(N-1)}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2},$$

tal que x_i é a i-ésima observação de x e N é o número total de observações de x.

NORMALIZAÇÃO POR MUDANÇA DE ESCALA

Iris (150 rows)

Detail | Compact | Column

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1 – 150	4.3 – 7.9	2 – 4.4	1 – 6.9	0.1 – 2.5	3 unique values
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2		
16	5.7	4.4	1.5		

SepalLengthCm

Length of the sepal (in cm)

Valid	150	100%
Mismatched	0	0%
Missing	0	0%
Mean	5.84	
Std. Deviation	0.83	
Quantiles	4.3	Min
	5.1	25%
	5.8	50%
	6.4	75%
	7.9	Max

$$\tilde{x} = \frac{x - \bar{x}}{\sigma_x}$$

https://www.kaggle.com/uciml/iris

PROPRIEDADES DOS MÉTODOS DE NORMALIZAÇÃO

PROPRIEDADES DOS MÉTODOS DE NORMALIZAÇÃO

- Por serem transformações lineares, não alteram a distribuição da variável normalizada em relação à variável original não-normalizada (Parâmetros podem mudar, mas a forma não).

- Como estas técnicas de normalização só utilizam estatísticas descritivas (min, max, média e desvio-padrão) das variáveis, tomadas individualmente, a correlação entre duas variáveis quaisquer permanece a mesma antes e depois da normalização;

- Variáveis normalizadas pelos métodos descritos anteriormente serão adimensionais (não apresentam unidades de medida).

Referências

- Richard O. Duda, Peter E. Hart, David G. Stork. **Pattern Classification**. John Wiley & Sons, 2012.
- Guilherme A. Barreto. **Introdução à Classificação de Padrões**. Grupo de Aprendizado de Máquinas – GRAMA, 2021.