Normalização de Dados

Guilherme de Alencar Barreto

gbarreto@ufc.br

Signal and Information Processing for Data Analysis
and Learning Systems (website: spiral.ufc.br)
Departamento de Engenharia de Teleinformática
Bloco 732, Centro de Tecnologia, Campus do Pici
Universidade Federal do Ceará – UFC
http://lattes.cnpq.br/8902002461422112

Objetivos

Objetivo: Entender a necessidade de equalizar as ordens de grandeza dos atributos usados em um problema de classificação/clusterização.

- **Método 1:** Manter constante a norma dos vetores.
- **Método 2:** Mudança da escala original para os intervalos [0, 1] ou [-1,+1].
- **Método 3:** Padronização z-score (i.e. média=0, variância=1).
- **Método 4:** Padronização Robusta (i.e. mediana=0, iqr=1).

Método 1: Norma Constante

- Uma das técnicas mais simples de normalização consiste em manter constantes e iguais a 1 as normas dos vetores de atributos $\mathbf{x}$ e dos centróides $\mathbf{m}_i$.
- Este procedimento deve ser aplicado a todos os vetores de atributos e todas os centróides.
- Para isso, basta dividir cada vetor por sua respectiva norma euclidiana:

$$\boxed{\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}} \quad \text{e} \quad \boxed{\tilde{\mathbf{m}}_i = \frac{\mathbf{m}_i}{\|\mathbf{m}_i\|}} \tag{1}$$

Método 1: Norma Constante

- Por exemplo, considere o seguinte vetor, que não possui norma unitária:

$$\mathbf{x} = \begin{bmatrix} \sqrt{3} \\ 3 \\ -2 \end{bmatrix} \tag{2}$$

- A norma deste vetor é calculada como

$$\|\mathbf{x}\| = \sqrt{(\sqrt{3})^2 + (3)^2 + (-2)^2} = \sqrt{16} = 4. \tag{3}$$

- Assim, a versão normalizada do vetor $\mathbf{x}$ é dada por

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{1}{4} \begin{bmatrix} \sqrt{3} \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} \sqrt{3}/4 \\ 3/4 \\ -1/2 \end{bmatrix} \tag{4}$$

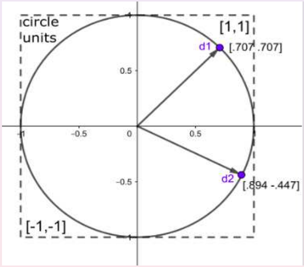
Propriedades do Método 1: Norma Constante

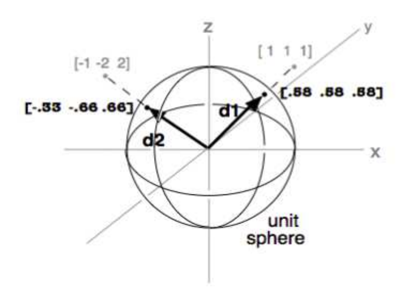
- A normalização descrita no slide anterior não altera a direção do vetor, apenas muda seu comprimento.
- Em outras palavras, o vetor resultante é um múltiplo do vetor original conforme pode ser visto na operação a seguir.

$$\tilde{\mathbf{x}} = \frac{1}{\|\mathbf{x}\|}\mathbf{x} = \alpha\mathbf{x}, \tag{5}$$

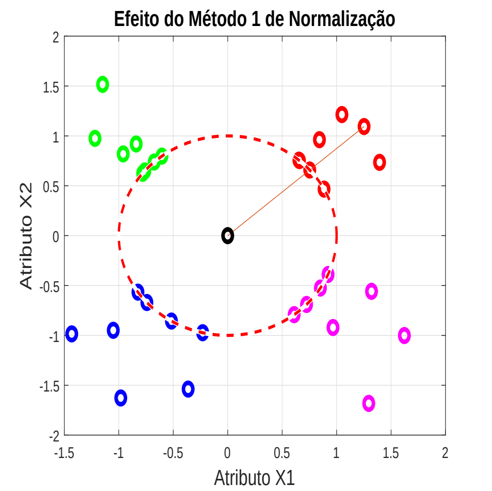
  em que $\alpha = 1/\|\mathbf{x}\|$ é uma constante positiva.
- Note que a normalização assim realizada depende apenas dos valores das componentes do vetor sendo normalizado.
- Assim, chamaremos este tipo de procedimento de *normalização local*.

- Método 1: Interpretação Geométrica





- Método 1: Interpretação Geométrica



Propriedades do Método 1: Norma Constante

- A similaridade entre 2 vetores $\mathbf{x}, \mathbf{v} \in \mathbb{R}^p$ de norma unitária pode ser calculada pelo cosseno do ângulo entre eles.
- A partir da fórmula do produto escalar, $\mathbf{x} \cdot \mathbf{v} = \|\mathbf{x}\| \times \|\mathbf{v}\| \times \cos(\theta)$, chega-se à seguinte expressão:

$$s(\mathbf{x}, \mathbf{v}) = \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{v}}{\|\mathbf{x}\| \times \|\mathbf{v}\|} \tag{6}$$

$$= \frac{\mathbf{x} \cdot \mathbf{v}}{1 \times 1} = \mathbf{x}^T \mathbf{v} = \sum_{j=1}^{p} x_j v_j. \tag{7}$$

- Resumo: A similaridade entre dois vetores de norma unitária é computada pelo produto escalar entre eles.

Método 1: Norma constante

- A normalização pelo Método 1 é particularmente útil para o classificador de máxima correlação (MC).
- O classificador MC nada mais é do que uma implementação dos classificadores de distância mínima (NN ou DMC) em que a medida de dissimilaridade é substituída por uma medida de similaridade, no caso, o produto escalar.
- O algoritmo do classificador MC é apresentado no próximo slide.

Classificador de Máxima Correlação

Passo 1 - Encontrar o vetor centróide de cada classe:

$$\boxed{\mathbf{m}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} \mathbf{x}} \tag{8}$$

em que N_i é o número de exemplos da i-ésima classe (cujo rótulo é ω_i), $i = 1, \ldots, C$.

Passo 2 - Atribuir um novo vetor de atributos $\mathbf{x}_{new}$ à mesma classe que $\mathbf{m}_{i^*}$, se

$$\boxed{\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} > \tilde{\mathbf{m}}_i^T \tilde{\mathbf{x}}_{new}, \quad \forall i \neq i^*} \tag{9}$$

em que $\tilde{\mathbf{m}}_i = \mathbf{m}_i / \|\mathbf{m}_i\|$ e $\tilde{\mathbf{x}}_{new} = \mathbf{x}_{new} / \|\mathbf{x}_{new}\|$ são as versões de norma unitária de $\mathbf{m}_i$ e $\mathbf{x}_{new}$, respectivamente.

Sobre Equivalência entre os classificadores DMC e MC

- O classificador MC nada mais é do que uma implementação dos classificadores de distância mínima (NN ou DMC) em que a medida de dissimilaridade é substituída por uma medida de similaridade, no caso, o produto escalar.
- Esta equivalência é fácil de mostrar a partir de um resultado muito conhecida da álgebra linear: $\|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{v}$.
- Assim, considere a desigualdade da Eq. (**??**), em que a função genérica $dist(\cdot,\cdot)$ é instanciada pela distância euclidiana quadrática.

Sobre Equivalência entre os classificadores DMC e MC

Assim, tem-se que

$$\begin{aligned} dist(\mathbf{x}_{new}, \mathbf{m}_i) &= \|\mathbf{x}_{new} - \mathbf{m}_i\|^2, && (10)\\ &= (\mathbf{x}_{new} - \mathbf{m}_i)^T(\mathbf{x}_{new} - \mathbf{m}_i), && (11)\\ &= \mathbf{x}_{new}^T\mathbf{x}_{new} - \mathbf{x}_{new}^T\mathbf{m}_i - \mathbf{m}_i^T\mathbf{x}_{new} + \mathbf{m}_i^T\mathbf{m}_i, && (12)\\ &= \mathbf{x}_{new}^T\mathbf{x}_{new} - 2\mathbf{m}_i^T\mathbf{x}_{new} + \mathbf{m}_i^T\mathbf{m}_i, && (13)\\ &= \|\mathbf{x}_{new}\|^2 - 2\mathbf{m}_i^T\mathbf{x}_{new} + \|\mathbf{m}_i\|^2, && (14) \end{aligned}$$

tal que, para $\|\mathbf{x}_{new}\| = \|\mathbf{m}_i\| = 1$, resulta em

$$dist(\tilde{\mathbf{x}}_{new}, \tilde{\mathbf{m}}_i) = 2 - 2\tilde{\mathbf{m}}_i^T\tilde{\mathbf{x}}_{new}. \qquad (15)$$

Sobre Equivalência entre os classificadores DMC e MC

Substituindo a Eq. (15) na Eq. (**??**), tem-se que

$$
\begin{aligned}
dist(\tilde{\mathbf{x}}_{new}, \tilde{\mathbf{m}}_{i^*}) &< dist(\tilde{\mathbf{x}}_{new}, \tilde{\mathbf{m}}_{i}) \\
2 - 2\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} &< 2 - 2\tilde{\mathbf{m}}_{i}^T \tilde{\mathbf{x}}_{new} \\
-2\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} &< -2\tilde{\mathbf{m}}_{i}^T \tilde{\mathbf{x}}_{new} \\
\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} &> \tilde{\mathbf{m}}_{i}^T \tilde{\mathbf{x}}_{new}
\end{aligned}
$$

sendo a última desigualdade a regra de decisão do classificador de máxima correlação.

Método 2: Mudança de escala

- Para classificadores baseados em distância euclidiana, uma normalização que promove uma mudança na escala das variáveis, é mais comum.
- Este procedimento é realizado variável a variável e requer a determinação do valor mínimo (x_{min}) e do valor máximo (x_{max}) da variável sendo normalizada.
- Por isso, chamaremos este tipo de procedimento de *normalização global*.
- Este tipo de normalização torna a variável adimensional.

Método 2: Mudança de escala para o intervalo [0,1]

$$\boxed{\tilde{x}_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}, \quad j = 1, \ldots, p} \tag{16}$$

com $\max(x_j)$ e $\min(x_j)$ sendo os valores máximo e mínimo do atributo x_j no conjunto de dados, respectivamente.

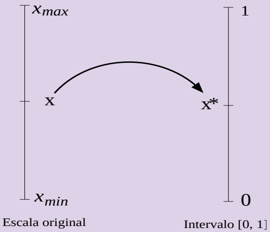


Figura: Mudança da escala do atributo x_j para o intervalo [0,1].

Método 2: Exemplo 1

- Considere o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $\min(x_1)$=11,03 e $\max(x)_1$=14,83.
- Assim, a função de normalização é dada por

$$\tilde{x}_1 = \frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)} = \frac{x_1 - 11,03}{14,83 - 11,03} = \frac{x_1 - 11,03}{3,80} \quad (17)$$

- Assim, a observação x_1=13,50 na escala original, terá o seguinte valor no intervalo [0, 1]:

$$\tilde{x}_1 = \frac{13,50 - 11,03}{3,80} = 0,65. \quad (18)$$

Método 2: Mudança de escala para o intervalo [-1,+1]

$$\boxed{\tilde{x}_j = 2\left(\frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}\right) - 1} \qquad (19)$$

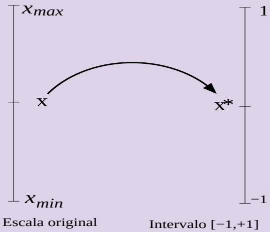


Figura: Mudança da escala original de x_j para o intervalo [-1,+1].

Método 2: Exemplo 2

- Considere o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos min(x_1)=11,03 e max(x_1)=14,83.
- Assim, a função de normalização é dada por

$$\tilde{x}_1 = 2\left(\frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)}\right) - 1 = 2\left(\frac{x_1 - 11,03}{3,80}\right) - 1 \tag{20}$$

- Assim, a observação x_1=13,50 na escala original, terá o seguinte valor no intervalo [-1,+1]:

$$\tilde{x}_1 = 2\left(\frac{13,50 - 11,03}{3,80}\right) - 1 = 0,30. \tag{21}$$

Método 3: Padronização da variável (média=0, variância=1)

- Assim como as normalizações descritas no Método 2, devemos aplicar a padronização às variáveis do problema, uma a uma.
- Este tipo de normalização requer o cálculo da média ($\hat{\mu}_j$) e do desvio-padrão ($\hat{\sigma}_j$) da variável x_j.
- Por isso, a padronização também pode ser chamada de normalização estatística, normalização pelo desvio-padrão, ou ainda normalização *z-score*.
- Este procedimento também é um tipo de *normalização global*.
- Este tipo de normalização torna a variável adimensional.

Método 3: Padronização z-score (média=0, variância=1)

A normalização estatística é dada por

$$\boxed{\tilde{x}_j = \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}} \tag{22}$$

com a média e o desvio-padrão amostrais de x_j calculados como

$$\boxed{\hat{\mu}_j = \frac{\sum_{n=1}^{N} x_j(n)}{N}} \quad \text{e} \quad \boxed{\hat{\sigma}_j = \sqrt{\left(\frac{\sum_{n=1}^{N}(x_j(n) - \hat{\mu}_j)^2}{N-1}\right)}} \tag{23}$$

tal que $x_j(n)$ é a n-ésima observação de x_j e N é o número total de observações de x.

Método 3: Exemplo numérico

- Usando o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $\hat{\mu}_1$=13,00 e $\hat{\sigma}_1$=0,81.
- Assim, a função de normalização é dada por

$$\tilde{x}_1 = \frac{x_1 - 13,00}{0,81} \tag{24}$$

- Assim, a observação x_1=13,50 na escala original, terá o seguinte valor padronizado:

$$\tilde{x}_1 = \frac{13,50 - 13,00}{0,81} = 0,617. \tag{25}$$

Método 4: Padronização robusta (mediana=0, iqr=1)

- Esta normalização usa estatísticas robustas, como mediana e o intervalo interquartil[a] (IQR):

$$\boxed{\tilde{x}_j = \frac{x_j - \text{mediana}(x_j)}{\text{IQR}(x_j)}} \tag{26}$$

em que a mediana é uma estatística robusta de tendência central, enquanto o IQR é uma medida robusta de dispersão das das medidas.

- Maiores detalhes sobre o IIQ em `https://pt.wikipedia.org/wiki/Amplitude_interquartil`

[a] Por vezes chamado de *amplitude* ou *faixa interquartil*.

Método 4: Padronização robusta (mediana=0, iqr=1)

- No Octave/Matlab, a mediana de um conjunto de medidas de um atributo x_j pode ser estimada de 2 diferentes maneiras por meio dos seguintes comandos:

```
» med1=median(x);
» med2=prctile(x,50);
```

- De modo similar, o iqr para um conjunto de medidas de x_j pode ser estimado pelos seguintes comandos:

```
» iqr1=iqr(x);
» iqr2=prctile(x,75)-prctile(x,25);
```

Normalização e Unidade da Variável

- Deve-se atentar para o fato de que os Métodos 2, 3 e 4 de normalização tornam a variável normalizada adimensional.
- Ou seja, perde-se unidade original da grandeza.
- Tomando como exemplo o Método 3, se x_j é tem unidade de tensão (volt, [V]), tanto o seu valor médio μ_j e seu desvio-padrão σ_j possuem unidade de tensão [V].
- Porém, a variável normalizada não terá mais unidade alguma; ou seja, passará a ser adimensional.
- O mesmo só ocorre com o Método 1 se as componentes do vetor $\mathbf{x}$ tiverem a mesma unidade.

Normalização como Transformação Linear

- As técnicas para normalização de variáveis descritas anteriormente podem ser vistas como uma transformação linear aplicada à variável original.
- Por exemplo, a normalização via Método 2 pode ser escrita da seguinte forma:

$$
\begin{aligned}
\tilde{x}_j &= \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)} \\
&= \left(\frac{1}{\max(x_j) - \min(x_j)}\right) x_j - \left(\frac{\min(x_j)}{\max(x_j) - \min(x_j)}\right) \\
&= a x_j + b
\end{aligned}
$$

em que $a = \frac{1}{\max(x_j) - \min(x_j)}$ e $b = -\frac{\min(x_j)}{\max(x_j) - \min(x_j)}$.

Propriedade 1 dos Métodos de Normalização

- Por serem transformações lineares, as normalizações descritas anteriormente não alteram a distribuição da variável normalizada em relação à variável original não-normalizada.
- Em outras palavras, o tipo de distribuição da variável permanece o mesmo. Por exemplo, se for gaussiana, continua gaussiana após a transformação.
- Os parâmetros da distribuição podem mudar, mas a forma dela não.
- Este resultado é suportado por um resultado teórico muito importante, que discutiremos a seguir.

Propriedade 1 dos Métodos de Normalização

- Seja $x \in \mathbb{R}$ uma variável aleatória contínua, de média μ_x e variância σ_x^2, cuja densidade de probabilidade é $f_X(x)$.
- Seja $y \in \mathbb{R}$ a variável aleatória resultante de uma operação matemática sobre x: $y = g(x)$.
- Pode-se mostrar que a FDP de y é dada por

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{dy}{dx}\right|} \tag{27}$$

- Assim, para $y = ax + b$, com a e b constantes reais, tem-se $|dy/dx| = |a|$ e $f_Y(y) = |a| f_X(x)$.
- Além disso, temos que $\mu_y = E[y] = E[ax + b] = aE[x] + b = a\mu_x + b$. E também $\sigma_y^2 = a^2\sigma_x^2$ (demonstrar!)

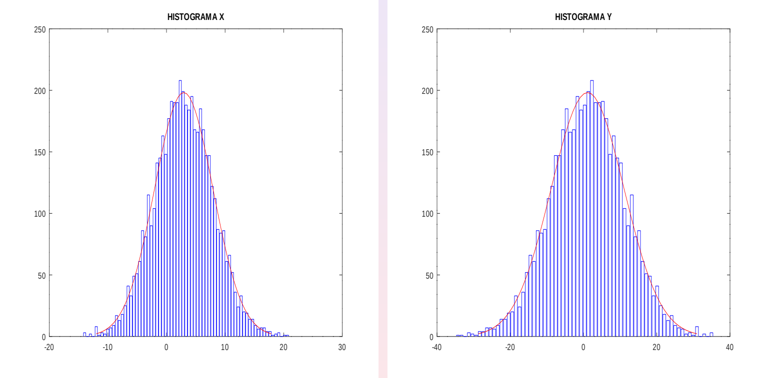
Propriedade 2 dos Métodos de Normalização

- Como estas técnicas de normalização só usam 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.
- Em outras palavras, transformações lineares preservam a correlação entre as duas 2 variáveis envolvidas: correlação antes da normalização = correlação depois da normalização.

Verificação da Propriedade 1: Código Octave/Matlab

```
» mtx=3; stx=5; % estatisticas teoricas de x
» x=normrnd(mtx,stx,5000,1); % gera 5000 observacoes N(mtx,stx2)
» STATSx=[mean(x) std(x)] % estatisticas amostrais de x
STATSx = 2.9447   4.9757
» a=-2; b=7; % parametros da transformacao linear
» y=a*x+b; % aplica transf.  linear a x
» figure; histfit(x); % histograma de x
» figure; histfit(y); % histograma de y
» mty=a*mtx+b, sty=abs(a)*stx % estatisticas teoricas de y
mty= 1
sty= 10
» STATSy=[mean(y) std(y)] % estatisticas amostrais de y
STATSy = 1.1105   9.9514
```

- Propriedade 1: Histogramas de X e Y.



Verificação da Propriedade 2: Código Octave/Matlab

```
» Cd=[4 2.8;2.8 9]; % matriz de covar desejada
» x=normrnd(0,1,5000,2); % gera 5000 observacoes 2 VA's
» A=chol(Cd); % gera matriz de mistura
» z=x*A; % gera VA's correlacionadas
» z1=z(:,1); z2=z(:,2);
» r12=corr(z1,z2) % correlacao entre z1 e z2
r12 = 0.45630
» z1n=(z1-mean(z1))/std(z1); % padroniza z1
» z2n=(z2-mean(z2))/std(z2); % padroniza z2
» STATSz1n=[mean(z1n) std(z1n)] % estatisticas de z1
STATSz1n = 1.6742e-17    1.0000e+00
» STATSz2n=[mean(z2n) std(z2n)] % estatisticas de z1
STATSz2n = -3.9257e-17    1.0000e+00
» r12n=corr(z1n,z2n) % correlacao entre z1 e z2
r12n = 0.45630
```

Implementação do Método 3 no Excel e LibreOffice Calc

- Dadas N observações conjuntas de um atributo qualquer, a normalização estatística (ou padronização) podem ser facilmente implementada em planilhas numéricas.
  - No Excel, usar os comandos `PADRONIZAR` ou `NORMALIZAR`.
  - No LibreOffice Calc, usar o comando `PADRONIZAR`.