ACH5504 – Mineração de Dados

Notas de aulas baseadas no livro

"*Introduction to Data Mining*"

Tan, Steinbach, Karpatne, Kumar

Resumo

- Similaridade de Atributos Nominais
- Similariades de Atributos Ordinais
- Similaridade de Atributos Numéricos
- Distância

Medidas de Similaridade e Dissimilaridade

- Similaridade:
  - Medida numérica de quão parecidos são duas instâncias de dados.
  - Valor maior quanto maior a similaridade.
  - Tipicamente no intervalo [0,1].
- Dissimilaridade:
  - Medida numérica de quão diferentes são duas instâncias de dados.
  - Valor menor indica maior similaridade.
  - Zero normalmente é o valor mínimo.
  - Limite superior pode variar.
- Proximidade se refere a similaridade ou dissimilaridade.

Dados

- Matriz de dados
  - n instâncias com p atributos

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Matriz de dissimilaridade
  - n instâncias por n instâncias
  - matriz triangular

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Similaridade – atributos nominais

- Atributos nominais
  - Método 1: verificação simples
    - m: número de atributos com valores iguais
    - p: número total de atributos

$$d(i, j) = \frac{p - m}{p}$$

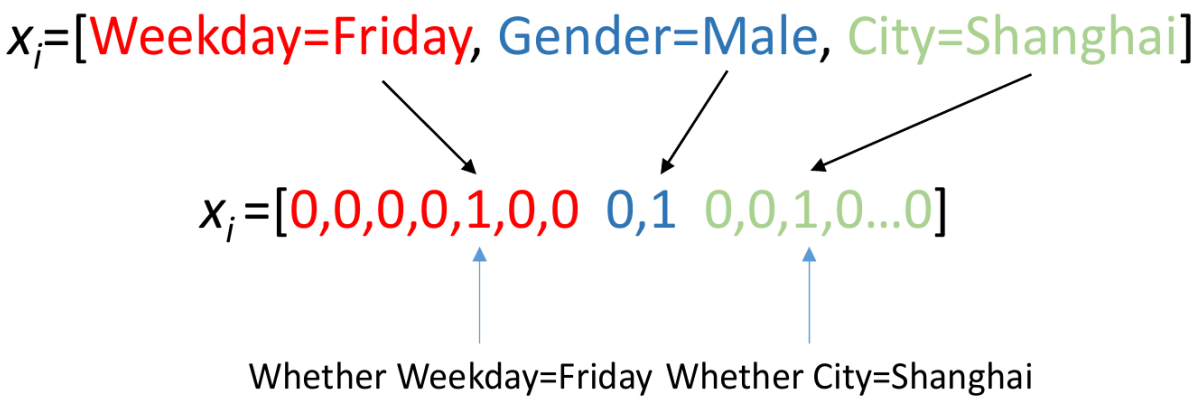
x_1=[Weekday=Friday, Gender=Male, City=Shanghai]

x_2=[Weekday=Friday, Gender=Female, City=Shanghai]

$$d(1,2) = \frac{3-2}{3} = \frac{1}{3}$$

Similaridade – atributos nominais

- Atributos nominais
  - Método 2: usar um grande número de atributos binários
    - Criar um novo atributo binário para cada um dos M estados nominais



Similaridade – atributos nominais

- Tabela de contingência para dados binários:

Instância *i* \ Instância *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Medida de distância para variáveis binárias simétricas:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Medida de distância para variáveis binárias assimétricas:

$$d(i,j) = \frac{r+s}{q+r+s}$$

  - Coeficiente de Jaccard (similaridade): similaridade entre conjuntos

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

Exemplo – similaridade de atributos nominais

- Atributos binários assimétricos
- Sim (S) e Positivo (P) são 1, Negativo (N) é 0

Nome	Febre	Tosse	Teste 1	Teste 2	Teste 3	Teste 4
Jack	S	N	P	N	N	N
Mary	S	N	P	N	P	N
Jim	S	P	N	N	N	N

- d(jack,mary) = (0+1)/(2+0+1) = 0.33
- d(jack,jim) = (1+1)/(1+1+1) = 0.67
- d(mary,jim) = (2+1)/(1+2+1) = 0.75

(Dis)Similaridade de Atributos Simples

A tabela mostra a similaridade e dissimilaridade entre dois objetos, *x* e *y*, com respeito ao único atributo simples.

Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Distância Euclidiana

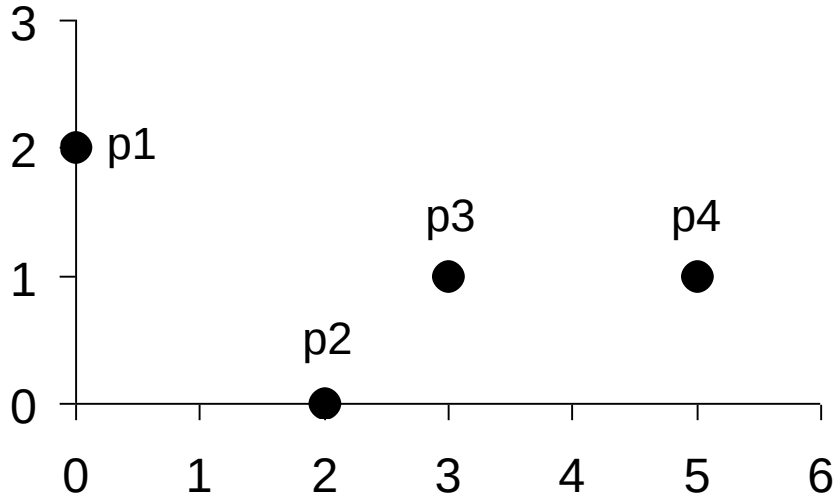
- Distância Euclidiana (dissimilaridade)

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

  - onde n é o número de dimensões (atributos) e x_k e y_k são, respetivamente, o k–ésimos atributos (componentes) ou objetos de dados $\mathbf{x}$ e $\mathbf{y}$.

- A padronização é necessária, se as escalas são diferentes.

Distância Euclidiana - Exemplo



Matriz de dados

ponto	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Matriz de distância euclidiana

Distância de Minkowski

- Distância de Minkowski é a generalização de distância euclidiana

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

  - Onde r é um parâmetro, n é o número de dimensões (atributos) e x_k e y_k são, respetivamente, os k–esimos atributos (componentes) ou objetos de dados $\boldsymbol{x}$ e $\boldsymbol{y}$.

Distância de Minkowski: Casos especiais

- r = 1. Distância de quarteirão (Manhattan, citiblock, L_1 norm).
  - Um exemplo comum disso é a distância de Hamming, que é apenas o número de bits que são diferentes entre dois vetores binários
- r = 2. Distância euclidiana
- $r = \infty$. Distância "supremum" (L_{max} norm, L_∞ norm).
  - Esta é a diferença máxima entre qualquer componente dos vectores
- Não confunda r com n, ou seja, todas essas distâncias são definidas para todos os números de dimensões.

Distância de Minkowski - exemplos

Dados

ponto	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

- Manhattan (L1)
- Euclidiana (L2)
- Supremum (Lmax)

Matrizes de distância

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Propriedades comuns de distâncias

- Distâncias, como a distância euclidiana, têm algumas propriedades bem conhecidas.

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ para todos x e y e $d(\mathbf{x}, \mathbf{y}) = 0$ somente se $\mathbf{x} = \mathbf{y}$. (definida positivamente)
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ para todos $\mathbf{x}$ e $\mathbf{y}$. (simetria)
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ para todos pontos $\mathbf{x}$, $\mathbf{y}$, e $\mathbf{z}$. (desigualdade triangular)

onde $d(\mathbf{x}, \mathbf{y})$ é a distância (dissimilaridade) entre pontos (objetos de dados), $\mathbf{x}$ e $\mathbf{y}$.

- Uma distância que satisfaça essas propriedades é uma métrica.

Propriedades comuns de distâncias

- Similaridades, também possuem algumas propriedades bem conhecidas.
  - $s(\mathbf{x}, \mathbf{y}) = 1$ (ou similaridade máxima) somente se $\mathbf{x} = \mathbf{y}$.
  - $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ para todos $\mathbf{x}$ e $\mathbf{y}$. (simetria)
- onde $s(\mathbf{x}, \mathbf{y})$ é a similaridade entre pontos (objetos de dados), $\mathbf{x}$ e $\mathbf{y}$.

Similaridade de Cosseno

- Se $\mathbf{d}_1$ e $\mathbf{d}_2$ são dois vetores de documento, então

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\|,$$

onde $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indica produto interno do produto ou do vetor do ponto dos vetores, $\mathbf{d}_1$ e $\mathbf{d}_2$, e $\| \mathbf{d} \|$ é o comprimento do vetor $\mathbf{d}$.

- Exemplo:

$$\mathbf{d}_1 = (3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0)$$

$$\mathbf{d}_2 = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2)$$

$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\| \mathbf{d}_1 \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d}_2 \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$$

Similaridade – atributos ordinais

- A ordem (rank) é importante
- Pode ser tratado como numérico:
  - Troque x_{if} pelo seu rank $r_{if} \in \{ 1, \ldots, M_f \}$
  - Mapeamento do alcance de cada variável em [0, 1] trocando i-ésima instância da f-ésima variável por

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Compute a distância usando métodos para atributo numérico

A correlação mede a relação linear entre objetos

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

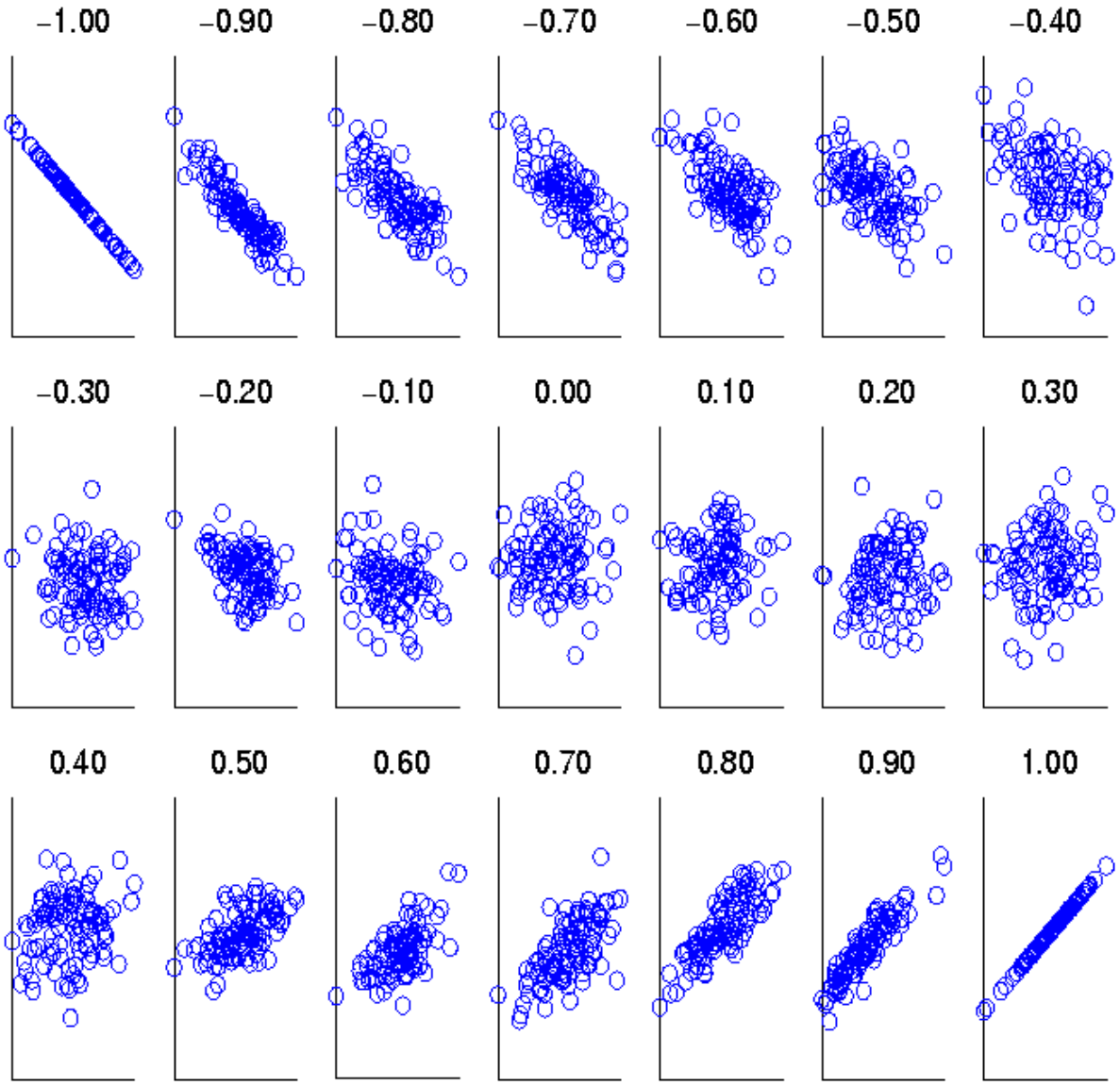
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Avaliando visualmente a correlação



Gráficos de dispersão mostram a similaridade de – 1 a 1.

Desvantagem de correlação

- **x** = (-3, -2, -1, 0, 1, 2, 3)
- **y** = (9, 4, 1, 0, 1, 4, 9)

- $y_i = x_i^2$

- mean(**x**) = 0, mean(**y**) = 4
- std(**x**) = 2.16, std(**y**) = 3.74

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)
  = **0**

Comparação de medidas de proximidade

- Domínio de aplicações
  - As medidas de similaridade tendem a ser específicas para o tipo de atributo e dados
  - Dados de registro, imagens, gráficos, sequências, estrutura de proteína 3D, etc., tendem a ter diferentes medidas
- No entanto, pode-se falar sobre várias propriedades que você gostaria de uma medida de proximidade para ter
  - Simetria é uma propriedade comum
  - A tolerância ao ruído e aos outliers é outra
  - Capacidade de encontrar mais tipos de padrões?
  - Muitas outras possíveis
- A medida deve ser aplicável aos dados e produzir resultados que concordem com o conhecimento do domínio

Entropia

- Para
  - a variável (evento), X,
  - com n valores possíveis (resultados), $x_1, x_2 \ldots, x_n$
  - cada resultado com probabilidade, $p_1, p_2 \ldots, p_n$
  - a entropia de X , $H(X)$, é dada por

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- A entropia está entre 0 e $\log_2 n$ e é medida em bits
  - Assim, a entropia é uma medida de quantos bits leva para representar uma observação de X em média

Exemplos para Entropia

- Para uma moeda com probabilidade p de caras e probabilidade $q = 1 - p$ q = 1 – p de coroas
  - $H = -p \log_2 p - q \log_2 q$
  - Para p= 0.5, $q = 0.5$ (moeda justa) $H = 1$
  - Para $p = 1$ ou $q = 1$, $H = 0$

- Qual é entropia de um dado de seis lados justo (cada lado tem mesma probabilidade)?

Entropia para dados: exemplo

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Entropia máxima é $\log_2 5 = 2.3219$

Entropia para dados amostrais

- Suponha que tenhamos
  - uma série de observações (m) de um atributo, X, e.g., a cor do cabelo dos alunos da turma,
  - onde temos n possíveis valores diferentes
  - e o número de observação em i–ésima categoria é m_i
  - Então, para essa amostra

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- Para dados contínuos, o cálculo é mais difícil

Maximal Information Coefficient

- Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.

- Applies mutual information to two continuous variables
- Consider the possible binnings of the variables into discrete categories
  - $n_X \times n_Y \leq N^{0.6}$ where
    - n_X is the number of values of X
    - n_Y is the number of values of Y
    - N is the number of samples (observations, data objects)
- Compute the mutual information
  - Normalized by $\log_2(\min(n_X, n_Y))$
- Take the highest value

Abordagem geral para combinar similaridades

- Às vezes, os atributos são de muitos tipos diferentes, mas uma similaridade geral é necessária.

1: Para o k–ésimo atributo, compute a similaridade, $s_k(\mathbf{x}, \mathbf{y})$, no intervalo [0, 1].

2: Define uma variável indicadora, δ_k, para o k–ésimo atributo com:

δ_k = 0 se o k–ésimo atributo é um atributo assimétrico e ambos os objetos têm um valor de 0, ou se um dos objetos tiver um valor ausente para o k–ésimo atributo

δ_k = 1 caso contrário

3. Compute

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

Usando pesos para combinar similaridades

- Pode não querer tratar todos os atributos igualmente.
  - Use pesos não negativos ω_k
  - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$
- Também pode definir uma forma ponderada de distância

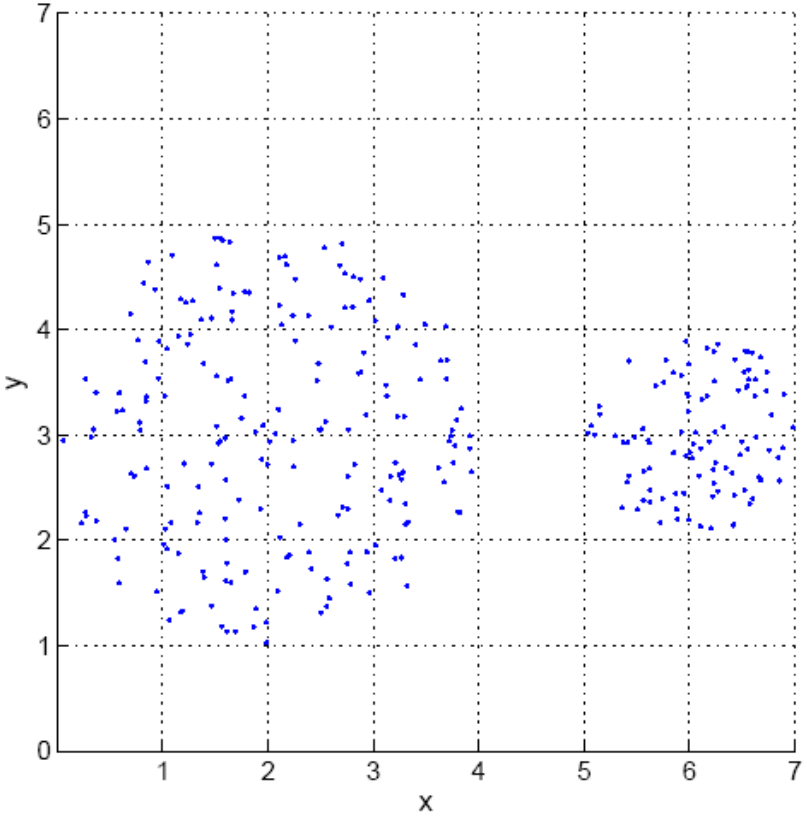
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

Densidade

- Mede o grau em que os objetos de dados estão próximos uns dos outros em uma área especificada
- A noção de densidade está intimamente relacionada com a proximidade
- O conceito da densidade é usado tipicamente para a aglomeração e a deteção de anomalias
- Exemplos:
  - Densidade euclidiana
    - Densidade euclidiana = número de pontos por unidade de volume
  - Densidade de probabilidade
    - Estima o que a distribuição dos dados se parece
  - Densidade baseada em gráficos
    - Conectividade

Densidade euclidiana: abordagem baseada em grade

- Abordagem mais simples é dividir a região em um número de células retangulares de volume igual e definir a densidade



Densidade baseada em grade.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Contagens para cada célula.

Densidade Euclidiana: Centro-baseada

- A densidade euclidiana é o número de pontos dentro de um raio especificado do ponto

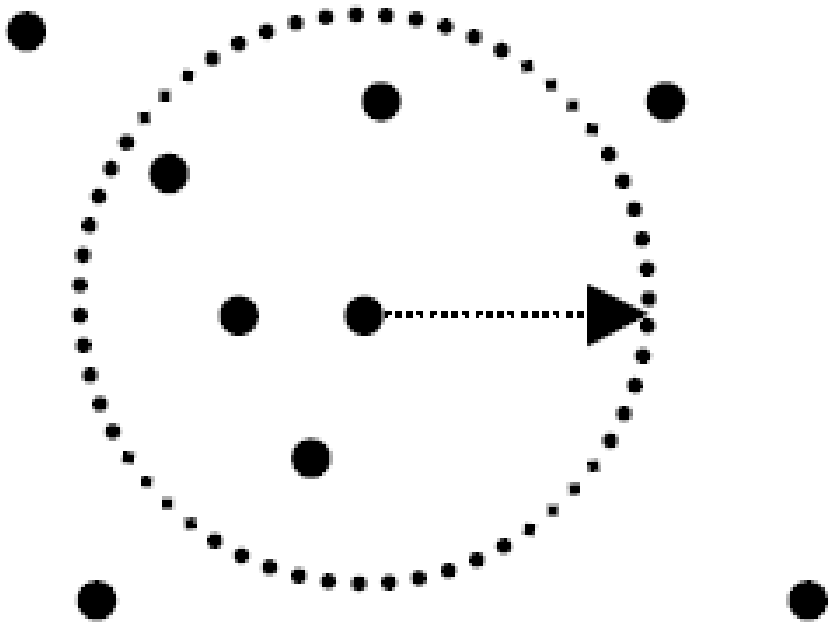


Ilustração da densidade centro-baseada.