

# Aprendizado de Máquina

## Aula 8: Algoritmos baseados em proximidade (parte 1)

André C. P. L. F de Carvalho  
ICMC/USP

[andre@icmc.usp.br](mailto:andre@icmc.usp.br)



# Tópicos

- Aprendizado baseado em proximidade
  - Aprendizado baseado em instâncias
- Proximidade
  - Similaridade e dissimilaridade (distância)
- 1-vizinho mais próximo
- Distância de Minkowski e suas variações
- K-vizinhos mais próximos
  - Propriedades de medidas de similaridade e de distância
- Variações
- Conclusão

# Tópicos

- Aprendizado baseado em proximidade
  - Aprendizado baseado em instâncias
- Proximidade
  - Similaridade e dissimilaridade (distância)
- 1-vizinho mais próximo
- Distância de Minkowski e suas variações
- K-vizinhos mais próximos
  - Propriedades de medidas de similaridade e de distância
- Variações
- Conclusão

# Aprendizado baseado em proximidade

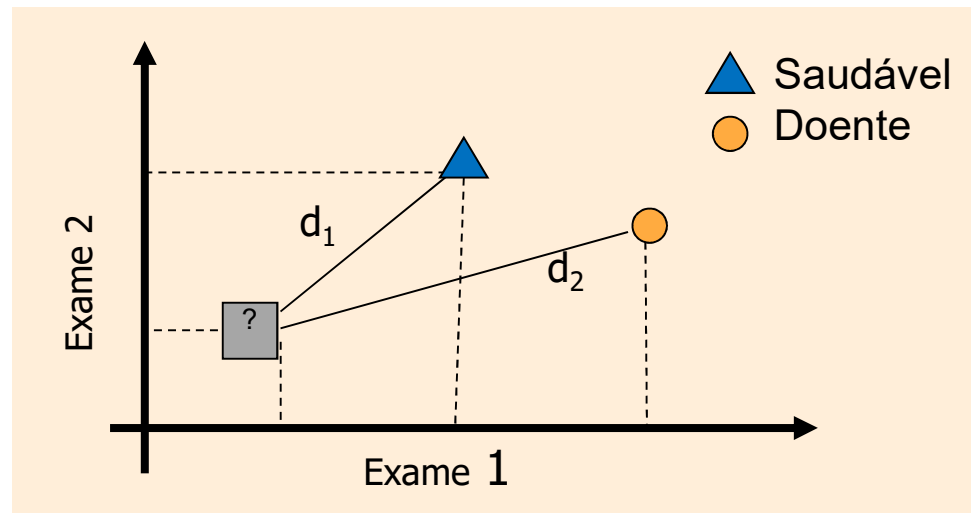
- Uso mais comum é em tarefas de aprendizado preditivo (classificação)
  - Utiliza medidas de proximidade para prever o rótulo de novos objetos
    - Supõe que objetos próximos (parecidos) têm rótulos semelhantes
      - Quando um novo objeto precisa ser classificado, ele recebe o rótulo do objeto (instância) mais próximo
  - Aprendizado baseado em instâncias
    - *Instance-based learning*

# Aprendizado baseado em instâncias

- Não tem uma fase explícita de treinamento
  - Apenas armazena os exemplos de treinamento
- Fase de teste
  - Para classificar um novo objeto  $x$ , compara ele com os demais objetos do “conjunto de treinamento” e selecionar os objetos mais parecidos com  $x$
  - Escolhe a classe que aparece mais vezes
- Como definir os mais parecidos?
  - Quantos e quais?

# Aprendizado baseado em instâncias

- Consideram proximidade entre dados
  - Medidas de similaridade
  - Medidas de dissimilaridade





# Tipos de atributos

- Simbólicos ou qualitativos
  - Nominal ou categórico
    - Ex.: cor, código de identificação, profissão
  - Ordinal
    - Ex.: gosto (ruim, médio, bom), dias da semana
- Numéricos, contínuos ou quantitativos
  - Intervalar
    - Ex.: data, temperatura em Celsius
  - Racional
    - Ex.: peso, tamanho, idade, temperatura em Kelvin

# Dissimilaridade x Similaridade

- Sejam  $a$  e  $b$  valores do atributo para dois objetos de um único atributo

## Tipo de atributo

### Nominal

#### Dissimilaridade

$$d(a, b) = \begin{cases} 1, & \text{se } a \neq b \\ 0, & \text{se } a = b \end{cases}$$

#### Similaridade

$$s(a, b) = \begin{cases} 0, & \text{se } a \neq b \\ 1, & \text{se } a = b \end{cases}$$

### Ordinal

$$d(a, b) = \frac{|pos_a - pos_b|}{n - 1}$$

$n = \text{\#valores}$   
 $n > 1$

$$s(a, b) = 1 - \frac{|pos_a - pos_b|}{n - 1}$$

### Intervalar ou racional

$$d(a, b) = |a - b|$$

$$s(a, b) = -d, \quad s(a, b) = \frac{1}{d} \text{ ou } s(a, b) = 1 - \frac{d - \min_d}{\max_d - \min_d}$$



# Dissimilaridade x Similaridade

- Sejam  $a$  e  $b$  valores do atributo para dois objetos de um único atributo

## Tipo de atributo

### Dissimilaridade

### Similaridade

## Nominal

$$d(\text{azul}, \text{amarelo}) = 1$$

$$s(\text{azul}, \text{amarelo}) = 0$$

## Ordinal

$$d(\text{terça}, \text{quinta}) = 2/6$$

$$s(\text{terça}, \text{quinta}) = 1 - 2/6$$

## Intervalar ou racional

$$d(4, 9) = 5$$

Supor valores variando de 3 a 10

$$s(4, 9) = -5, \quad s(4, 9) = \frac{1}{5} \text{ ou } s(4, 9) = 1 - \frac{5-3}{10-3}$$

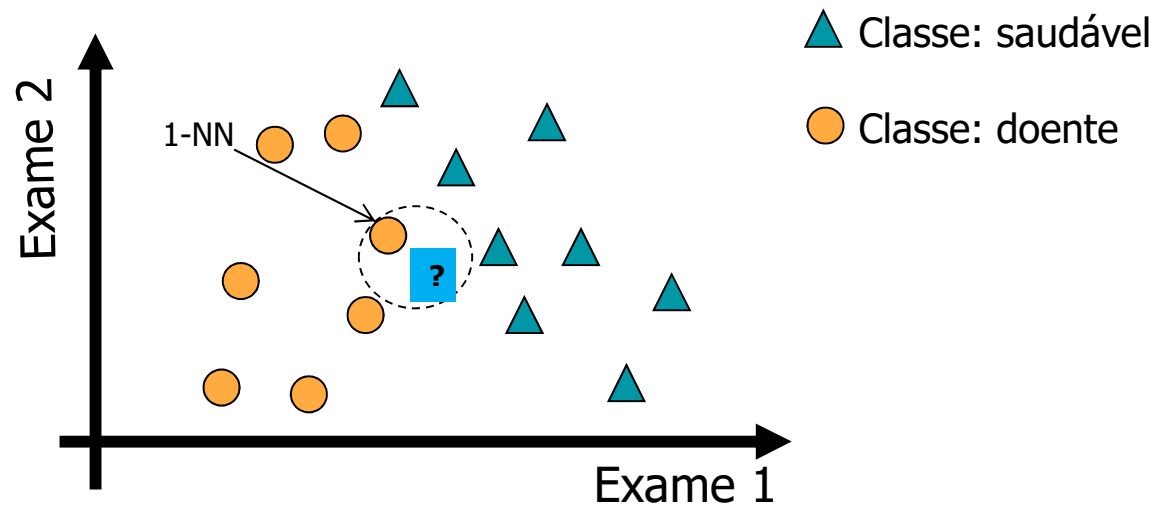
# Funções de transformação

- Convertem medida de similaridade em medida de dissimilaridade
  - E vice-versa
- Fazem com que o valor retornado pela medida:
  - Fique dentro de um dado intervalo
  - Apresente uma dada distribuição

# Algoritmo k-vizinhos mais próximos (k-NN)

- Geralmente usado em tarefas de classificação
- Algoritmo de aprendizado *lazy* (preguiçoso)
  - Olha os dados de treinamento apenas quando vai classificar um novo objeto
    - Processamento é atrasado até o momento de classificação de um novo exemplo
    - Não constrói um modelo explicitamente
  - Diferente de um algoritmo *eager* (ansioso)
    - Olha os dados de treinamento para induzir um modelo, depois usado para classificar novos objetos

# Algoritmo 1-vizinho mais próximo

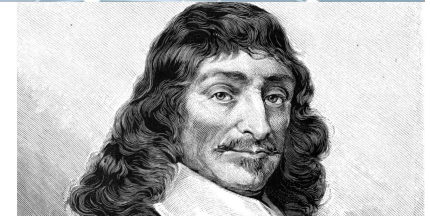


# Algoritmo 1-vizinho mais próximo

- Novo exemplo é atribuído a classe do exemplo mais próximo
  - Medida de distância
    - Valores dos  $d$  atributos definem coordenadas no espaço  $d$ -dimensional
    - Geralmente utiliza a distância euclidiana
  - Superfície de decisão
    - Muito complexas
      - Define poliedros convexos com centro nos exemplos de treinamento
      - Conjunto de poliedros forma um diagrama de Voronoi

# Diagrama de Voronoi

- Estudado por René Descartes
  - Filósofo/físico/matemático francês
    - Mas nome homenageia o matemático ucraniano Georgy Voronoy (que definiu e estudou o caso d-dimensional)
- Criado pela distribuição aleatória de pontos em um plano euclidiano
  - Que é dividido em polígonos convexos (tesselações), um em torno de cada ponto



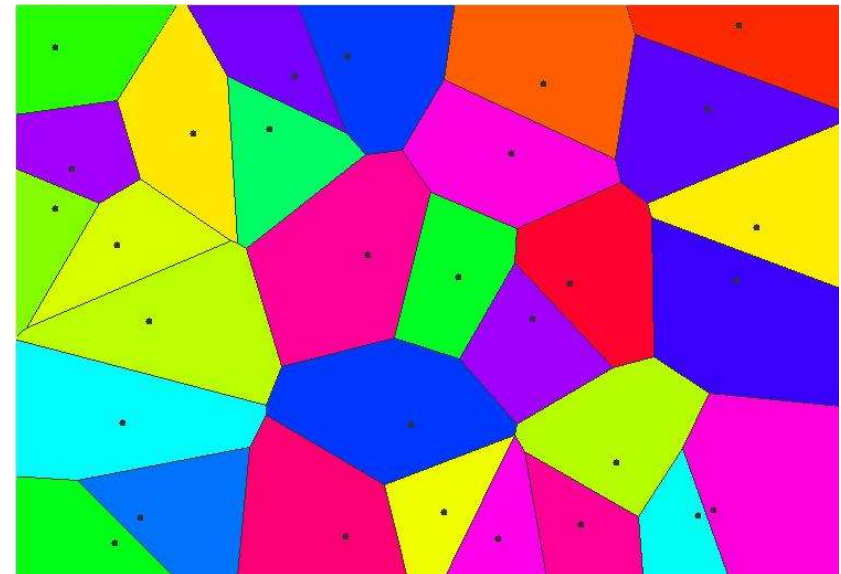
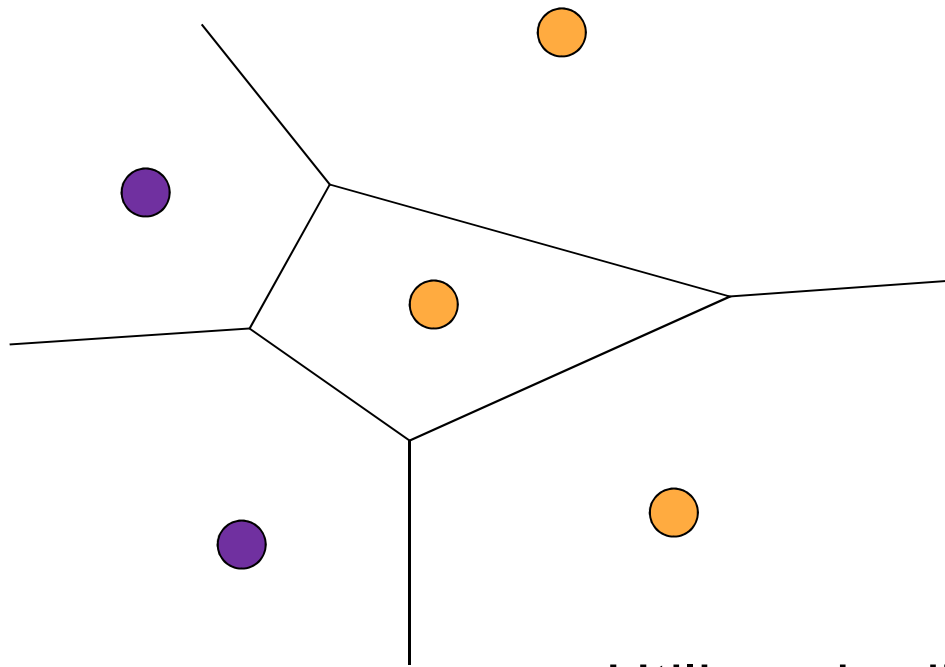


# Diagrama de Voronoi

- Estudado por René Descartes
  - Filósofo/físico/matemático francês
    - Mas nome homenageia o matemático ucraniano Georgy Voronoy (que definiu e estudou o caso d-dimensional)
- Criado pela distribuição aleatória de pontos em um plano euclidiano
  - Que é dividido em polígonos convexos (tesselações), um em torno de cada ponto
    - Tesselação: pavimentação, mosaico
    - Define região do plano mais próxima àquele ponto do que a qualquer outro ponto



# Diagrama de Voronoi



Utilizando distância euclidiana

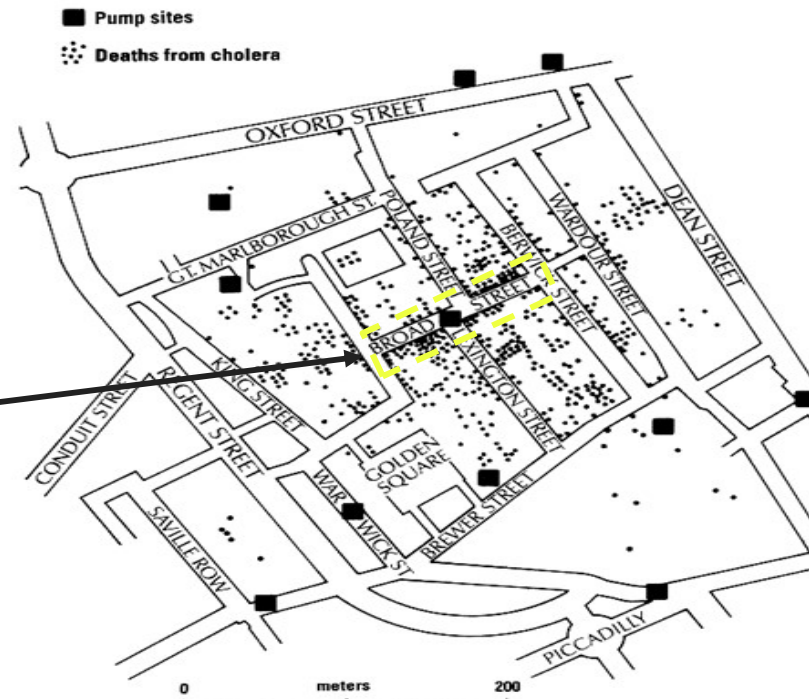
# Diagrama de Voronoi

- Possui várias aplicações (não só na matemática)
  - Modelagem de território animal
  - Navegação de robôs
  - Modelagem de crescimento de cristais
  - Usado em 1954 pelo médico John Snow, durante a epidemia da cólera em Londres
    - Criou diagrama para identificar locais em que havia bomba de água
    - Contou o número de mortes em cada polígono para achar a bomba que provocava a infecção

# Cólera

Mapa da Cólera em Londres (Snow) 1854

Distribuição da doença permitiu identificar que a fonte da cólera era uma bomba de água pública na Broad Street



# Medidas de distância

- Já vimos como calcular similaridade e dissimilaridade entre valores de 1 atributo preditivo
- Supor agora que cada objeto pode ter  $d$  atributos preditivos
  - Para medir dissimilaridade, são utilizadas medidas de distância
    - Existem várias
    - Algumas delas são derivadas da distância de Minkowski



# Distância de Minkowski

- Medida de distância generalizada

$$\text{distânciaMinkovsk}_i(p, q) = (\sum_{k=1}^d |p_k - q_k|^r)^{\frac{1}{r}}$$

- Escolha do valor de r resulta em diferentes medidas de distância:
  - 1 ( $L_1$ ): Distância bloco cidade (Manhattan, geometria do taxi)
    - Hamming (para valores binários ou cadeias de caracteres)
      - Ex.: 100011 e 011011
  - 2 ( $L_2$ ): Distância euclidiana
  - $\infty$  ( $L_\infty$  ou  $L_{\max}$ ): Distância de Chebyshev (máxima, do tabuleiro de xadrez)



# Medidas de distância

- Distância bloco cidade (Manhattan)
  - Medida de menor complexidade (e exatidão)

$$distância_{Bloco}(p, q) = \sum_{k=1}^d |p_k - q_k|$$

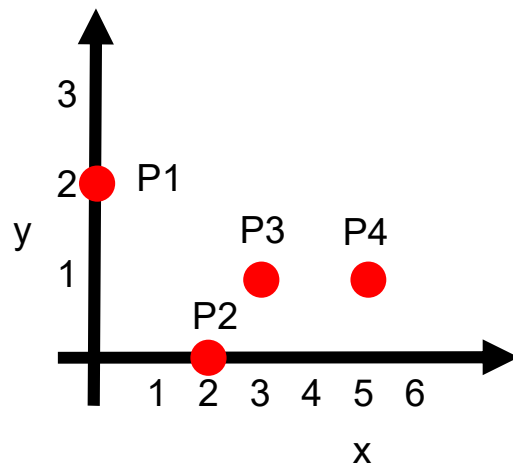
- Distância euclidiana
  - Sistemas de coordenadas cartesianas

$$distância_{Euclidiana}(p, q) = \sqrt{\sum_{k=1}^d (p_k - q_k)^2}$$

- Distância máxima (Chebyshev)

$$distância_{Máxima}(p, q) = MAX(|p_k - q_k|)$$

# Distância euclidiana



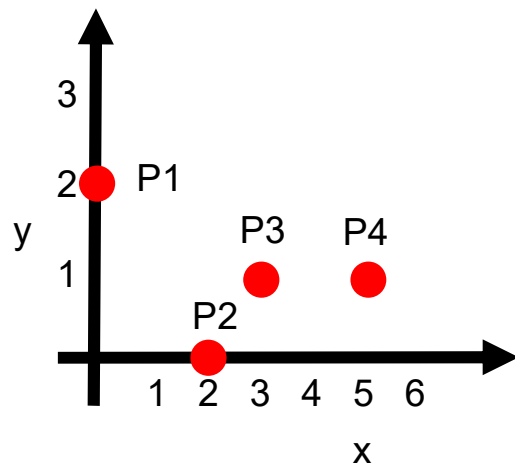
Coordenadas  
dos objetos

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

Matriz de distâncias  
entre os objetos

	p1	p2	p3	p4
p1				
p2				
p3				
p4				

# Distância euclidiana



Coordenadas  
dos objetos

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

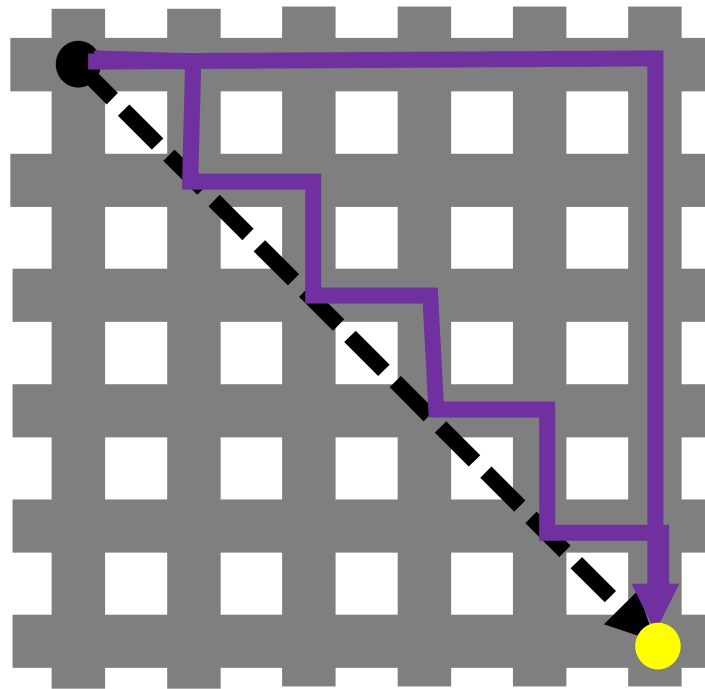
Matriz de distâncias  
entre os objetos

	p1	p2	p3	p4
p1	0,00	2,828	3,162	5,099
p2	2,828	0,00	1,414	3,162
p3	3,162	1,414	0,00	2,000
p4	5,099	3,162	2,000	0,00

# Distância euclidiana

- Medida de distância mais utilizada
- Atributos com escalas de valores diferentes
  - Pode ser necessário padronização ou re-escala
- O cálculo da raiz quadrada tem um custo elevado
  - Que pode ser evitado por outras medidas
    - Distância bloco cidade (Manhattan)
    - Distância máxima

# Medidas de distância



Distância euclidiana



Distância bloco cidade (Manhattan)




# Distância máxima

- Também conhecida como distância de Chebyshev, **distância quadrática** ou **do tabuleiro de xadrez**
  - Distância de menor complexidade
    - E de menor precisão
  - Supor  $p = [1, 2, -4]$  e  $q = [2, 0, 3]$ 
    - Retorna maior distância entre os atributos
    - Distâncias entre atributos:
      - $|1 - 2| = 1$
      - $|2 - 0| = 2$
      - $|-4 - 3| = 7$



# Distância máxima

- Quantas casas o rei percorre entre sua posição inicial e sua posição alvo
  - Em uma ou mais jogadas
- Ex.: mover de f6 para b4

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	






# Distância máxima

- Quantas casas o rei percorre entre sua posição inicial e sua posição alvo

- Em uma ou mais jogadas
- Ex.: mover de f6 para b4

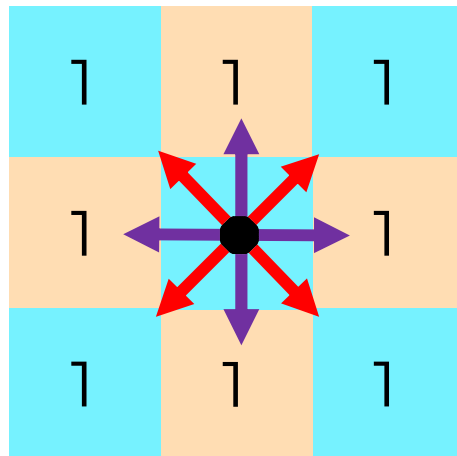
- $\text{Max} (|x_{\text{alvo}} - x_{\text{inicial}}|, |y_{\text{alvo}} - y_{\text{inicial}}|)$

- $\text{Max} (4, 2) = 4$

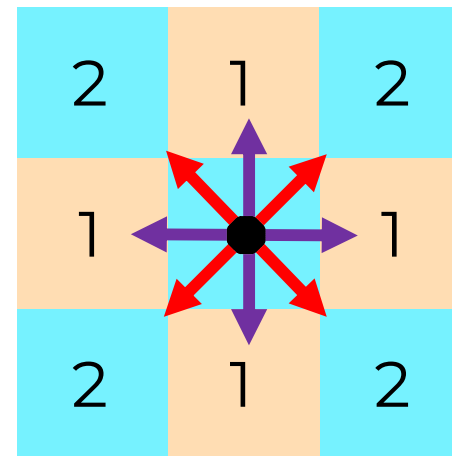
	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2		1	1	2	5
4	5				2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

# Chebyshev versus Manhattan

Chebyshev

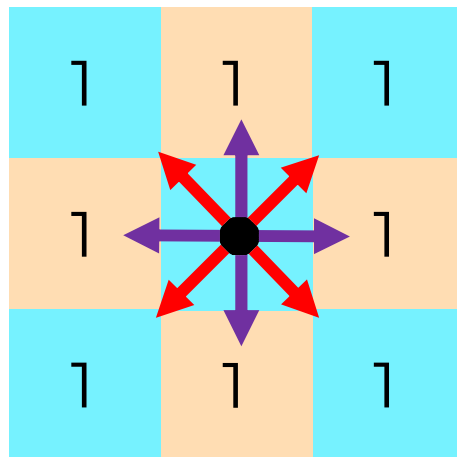


Manhattan



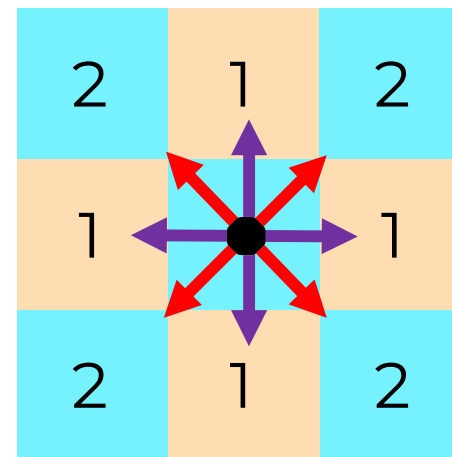
# Chebyshev versus Manhattan

Chebyshev



**Veículo voador**

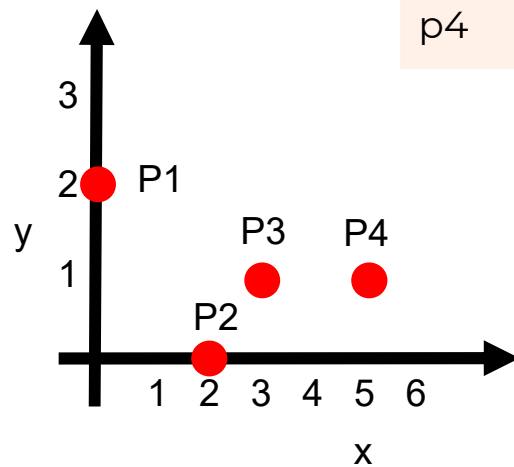
Manhattan



**Veículo terrestre**

# Distância de Minkowski

Coordenadas  
dos objetos



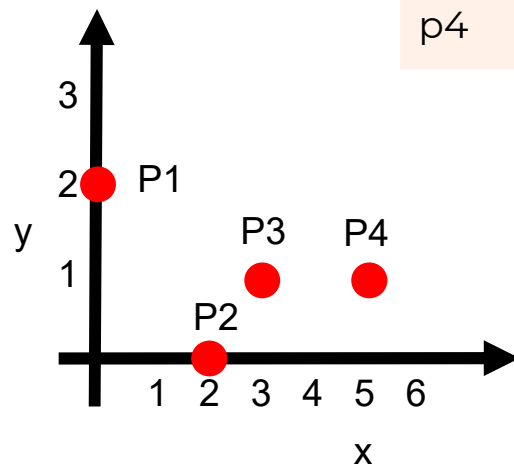
Objeto	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Matriz de distâncias  
entre os objetos

$L_1$	p1	p2	p3	p4
p1				
p2				
p3				
p4				
$L_2$	p1	p2	p3	p4
p1				
p2				
p3				
p4				
$L_\infty$	p1	p2	p3	p4
p1				
p2				
p3				
p4				

# Distância de Minkowski

Coordenadas dos objetos



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

Matriz de distâncias entre os objetos

$L_1$	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

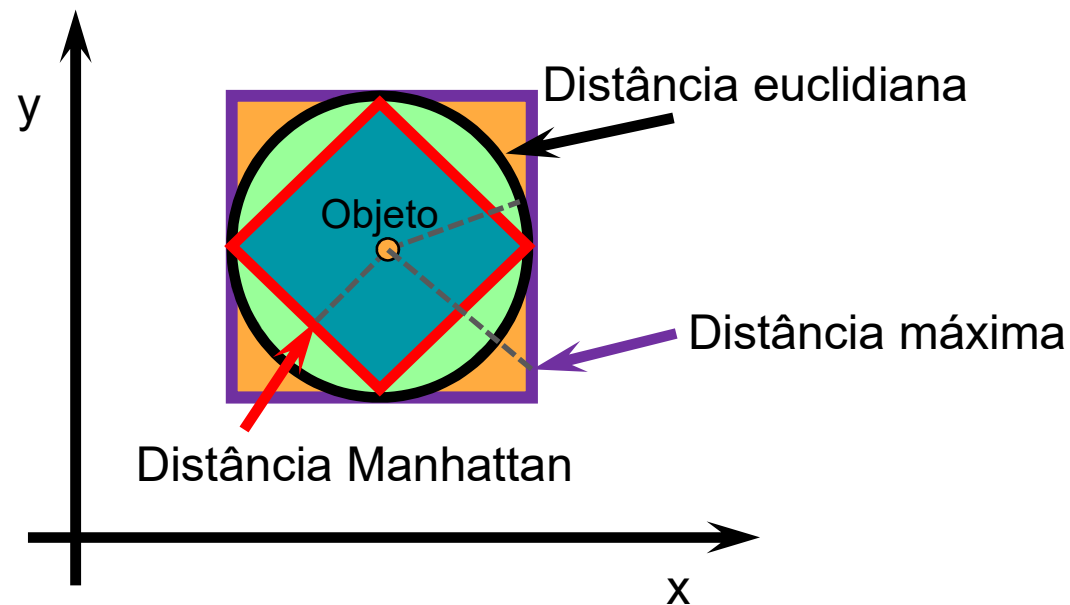
$L_2$	p1	p2	p3	p4
p1	0,00	2,828	3,162	5,099
p2	2,828	0,00	1,414	3,162
p3	3,162	1,414	0,00	2,000
p4	5,099	3,162	2,000	0,00

$L_\infty$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0



# Medidas de distância

- Onde se situam os pontos equidistantes de um objeto representado por um vetor



Contínua na  
próxima aula