

# Curso 2 – CD, AM e DM

## MBA EM IA e BIGDATA

ARVORES DE DECISÃO

PROFA. ROSELI AP. FRANCELIN ROMERO



# Curso 2 – CD, AM e DM

## MBA EM IA e BIGDATA

### MEDIDAS DE DISTÂNCIA

### KNN

Roseli Ap. Francelin Romero  
SCC – ICMC - USP



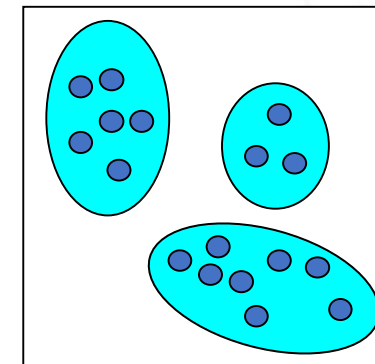
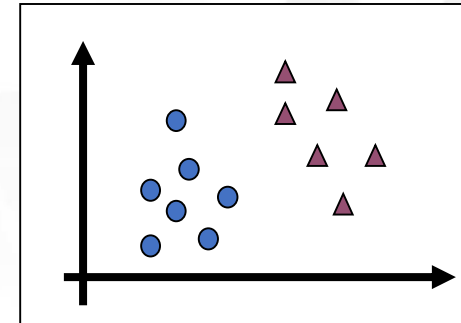
# Principais tópicos

- Aprendizado baseado em instâncias
- 1-vizinho mais próximo
- Medidas de distância
- Similaridade e dissimilaridade
- Proximidade
- K-vizinhos mais próximos
- Conclusão



# AM e Geometria

- Medidas de distância
  - Podem ser usadas para
    - Classificar novos dados
      - Ex.: K-NN
    - Agrupar dados
      - Ex.: K-médias
- Existem várias medidas

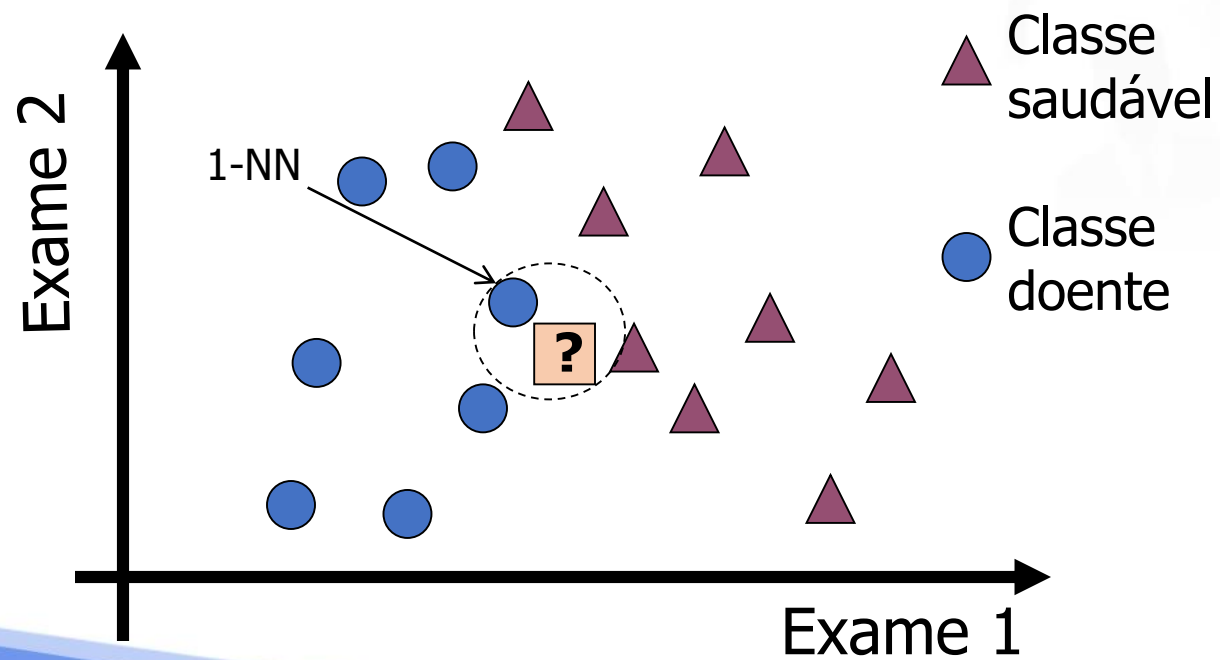


# 1-vizinho mais próximo

- Versão simples do algoritmo k-NN
  - Geralmente usado para classificação
- Algoritmo *lazy* (preguiçoso)
  - Olha os dados de treinamento apenas quando vai classificar um novo objeto
  - Não constrói um modelo explicitamente
  - Diferente de algoritmos
    - Induzem modelo
      - Ex.: ADs, RNs e SVMs



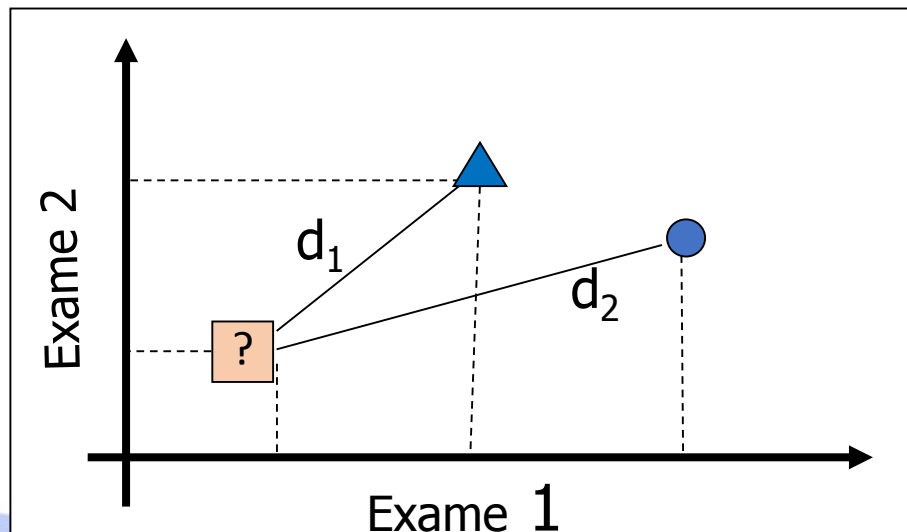
# 1-vizinho mais próximo





# Métodos baseados em distância

- Consideram proximidade entre dados
  - Similaridade
  - Dissimilaridade



- Existem várias
  - Euclidiana
  - Norma máxima
  - Bloco-cidade
  - ...



# Propriedade de Distâncias

- Medidas de distância, em geral, têm as seguintes propriedades
  - Seja  $d(p, q)$  a distância (dissimilaridade) entre dois objetos  $p$  e  $q$ 
    - $d(p, q) \geq 0 \quad \forall p \text{ e } q$  e  $d(p, q) = 0$  se e somente se  $p = q$  (definida positiva)
    - $d(p, q) = d(q, p) \quad \forall p \text{ e } q$  (simetria)
    - $d(p, r) \leq d(p, q) + d(q, r) \quad \forall p, q \text{ e } r$  (desigualdade triangular)
- Medidas que satisfazem essas propriedades são denominadas métricas





# Propriedade de Distâncias

- Medidas de similaridade também têm propriedades bem definidas:
  - Seja  $s(p, q)$  a similaridade entre dois objetos  $p$  e  $q$ 
    - $s(p, q) = 1$  (similaridade máxima) apenas se  $p = q$
    - $s(p, q) = s(q, p) \forall p \text{ e } q$  (simetria)



# Distância de Minkowski

- Medida de distância generalizada

$$dist = \left( \sum_{k=1}^m |p_k - q_k|^r \right)^{\frac{1}{r}}$$

- Valor de r leva a diferentes distâncias:
  - 1 ( $L_1$ ): Distância bloco cidade (Manhattan)
    - Hamming (valores binários)
  - 2 ( $L_2$ ): Distância Euclidiana



# Medidas de distância

- Distância Euclidiana
  - Sistema de coordenadas cartesianas

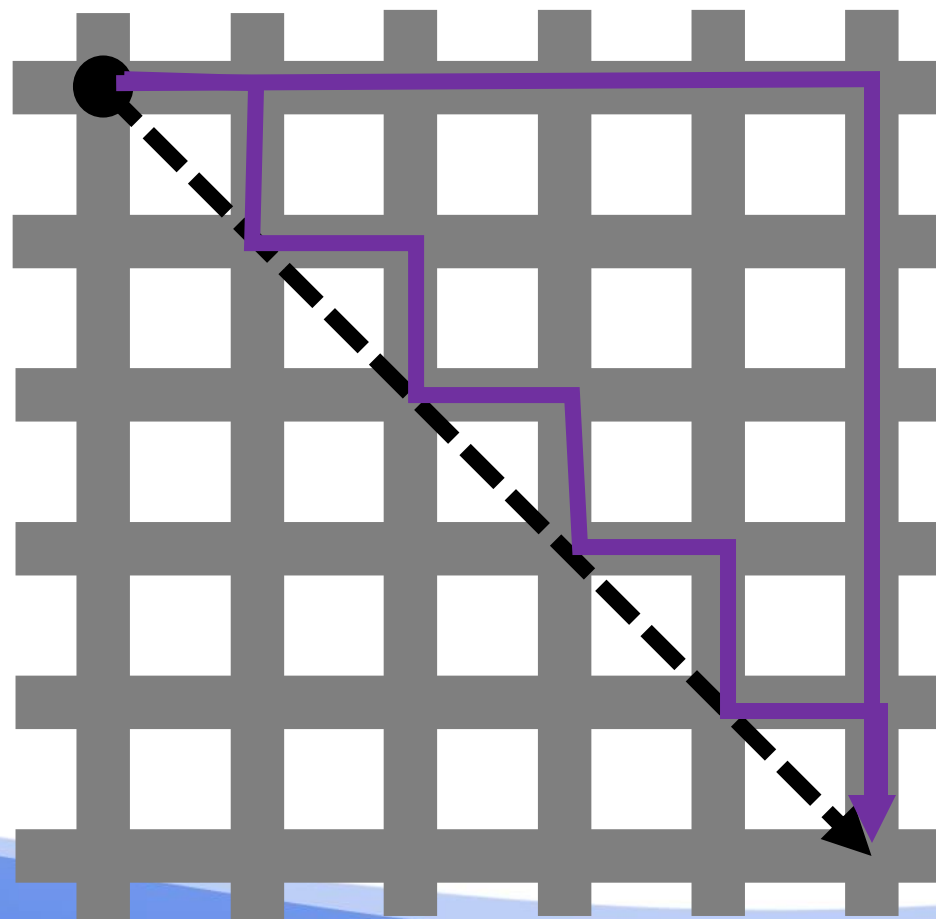
$$dist = \sqrt{\sum_{k=1}^m (p_k - q_k)^2}$$

- Distância de norma máxima
  - Menor complexidade e menos exatidão

$$dist = MAX(|p_k - q_k|)$$



# Medidas de distância



Distância Euclidiana



Distância Manhattan



# Exercício

- Qual das três medidas resulta na maior e na menor distância entre os exemplos abaixo?
  - Manhattan
  - Euclidiana
  - Norma máxima

Ex1 = (3, 1, 10, 2)

Ex2 = (2, 5, 3, 2)



# Exercício

■ Utilizando distância de Manhattan, definir:

■ Qual par dos números binários abaixo tem a distância mais semelhante à diferença entre seus valores na base decimal?

110000, 111001, 000111, 001011, 100111, 101001





# Similaridade x Dissimilaridade

- Similaridade
  - Mede o quanto dois objetos são parecidos
    - Quanto mais parecidos, maior o valor
- Dissimilaridade
  - Mede o quanto dois objetos são diferentes
    - Distância
    - Quanto mais diferentes, maior o valor
- Medida de proximidade pode ser usada



# Proximidade entre valores

■ Sejam a e b dois valores de um atributo

■ Nominal

■  $\text{sim} = 1 - d$

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

■ Ordinal

■  $\text{sim} = 1 - d$

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

■ Intervalar ou racional

■  $\text{sim} = 1 - d$  ou  $\text{sim} = 1/(1+d)$

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



# Exercício

- Para cada medida de distância
  - Quais são os dois exemplos da tabela abaixo mais próximos e os dois mais distantes?
  - Usar distâncias Euclidiana, bloco cidade e norma máxima

Estado	Escolaridade	Altura	Salário	Classe
SP	Médio	180	3000	A
RJ	Superior	174	7000	B
RJ	Fundamental	100	2000	A



# K-vizinhos mais próximos

- Generalização do 1-vizinho mais próximo
- Algoritmo de AM baseado distância muito simples
  - Memória
- Número de vizinhos (k) pode variar



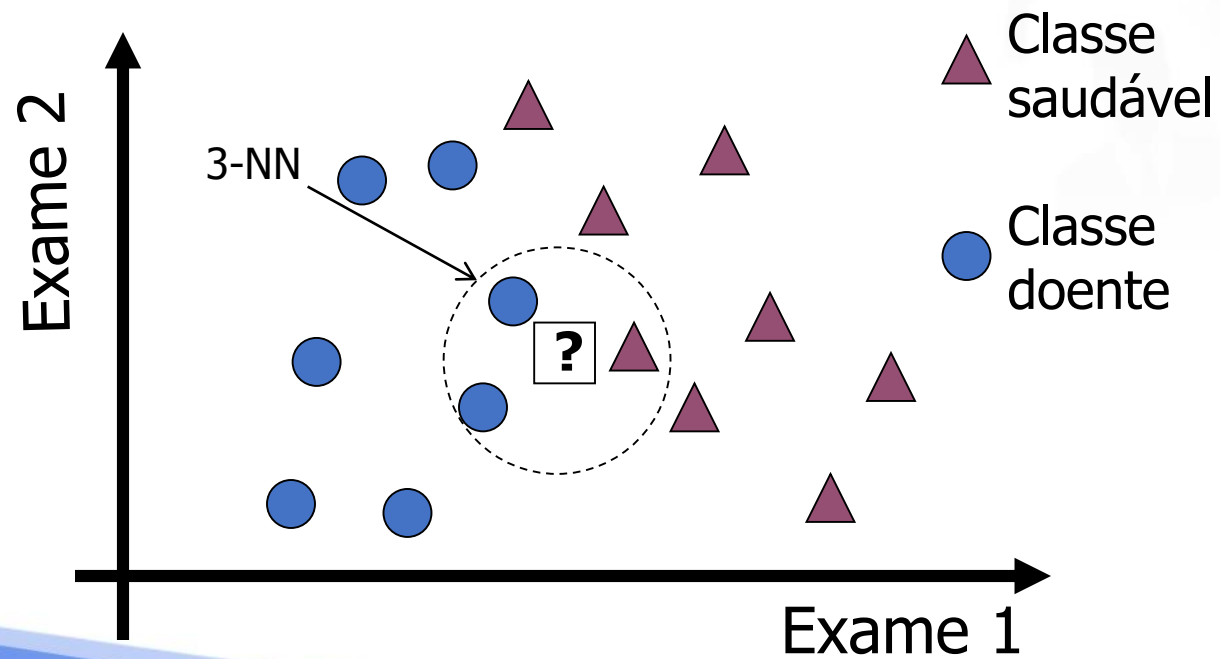
# Quantos vizinhos?

- K muito grande
  - Vizinhos podem ser muito diferentes
  - Predição tendenciosa para classe majoritária
  - Custo computacional mais elevado
- K muito pequeno
  - Considera apenas os objetos muito parecidos
    - Não usa quantidade suficiente de informação
  - Previsão pode ser instável
    - Ruído



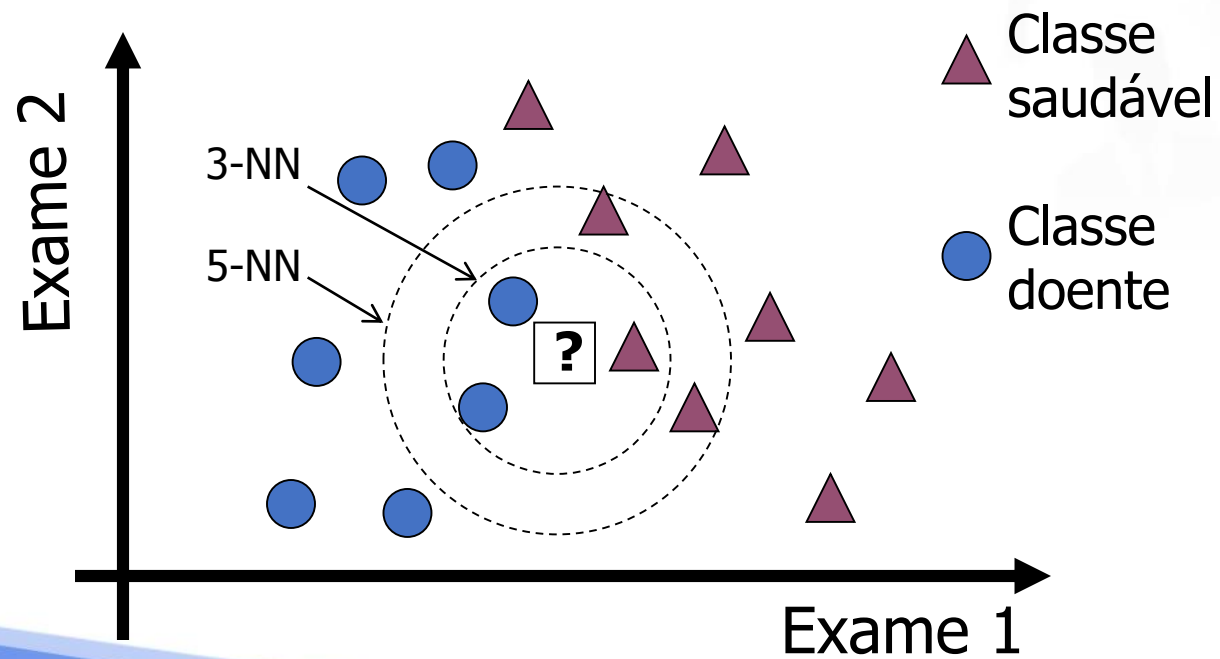


# Quantos vizinhos?





# Quantos vizinhos?



# K-Vizinhos mais próximos

Seja  $k$  o número de vizinhos mais próximos  
Para cada novo exemplo  $x$   
Definir a classe dos  $k$  exemplos  
(vizinhos) mais próximos  
Classificar  $x$  na **classe majoritária**  
entre seus  $k$  vizinhos



# K-vizinhos mais próximos

- Abordagem local
- Processo de classificação pode ser lento
  - Seleção de atributos
  - Eliminação de exemplos
    - Guardar conjunto de protótipos para cada classes
    - Algoritmos iterativos
      - Eliminação sequencial
      - Inserção sequencial



# K-vizinhos mais próximos

- Algoritmos iterativos para eliminação
  - Eliminação sequencial
    - Começa com todos os exemplos
    - Descarta exemplos corretamente classificados pelos protótipos
  - Inserção sequencial
    - Conjunto inicial tem apenas os protótipos
    - Acrescenta exemplos incorretamente classificados pelos protótipos (expande protótipos)



# Similaridade entre vetores binários

- Algumas vezes, objetos  $p$  e  $q$  têm apenas valores binários
  - Ex.: 0110 e 1100
- Similaridades podem ser computadas usando:
  - $M_{01}$  = número de atributos em que  $p = 0$  e  $q = 1$
  - $M_{10}$  = número de atributos em que  $p = 1$  e  $q = 0$
  - $M_{00}$  = número de atributos em que  $p = 0$  e  $q = 0$
  - $M_{11}$  = número de atributos em que  $p = 1$  e  $q = 1$





# Similaridade entre vetores binários

- Coeficiente de Casamento Simples

$$CCS = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

- Coeficiente Jaccard

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$$

- Agrupamento de dados





# Exercício

- Que medida de similaridade binária gera o maior valor de similaridade entre vetores  $p$  e  $q$ ?

$p = 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0$   
 $q = 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1$



# Similaridade cosseno

- Muito usado quando dados são textos
  - *Bag of words*
    - Grande número de atributos
    - Vetores esparsos
- Sejam  $p$  e  $q$  vetores representando documentos
  - $\cos(p, q) = \frac{||p|| \cdot ||q|| \cos\theta}{||p|| \cdot ||q||} = (p \bullet q) / (||p|| \cdot ||q||)$ 
    - $\bullet$ : vector produto interno entre vetores
    - $||p||$ : é o tamanho (norma) do vetor  $p$



# Distância cosseno

- Distância angular entre dois vetores
  - Invariante a escala dos atributos
  - 1 – similaridade cosseno

$$dist_{\text{cosseno}} = 1 - \frac{\sum_{k=1}^m p_k \cdot q_k}{\sum_{k=1}^m p_k^2 \cdot \sum_{k=1}^m q_k^2}$$



# Distância de Pearson

- Muito usada em bioinformática e séries temporais
  - 1 – correlação entre dois vetores

$$dist_{Pearson} = 1 - \frac{\sum_{k=1}^m (p_k - \bar{p}) \cdot (q_k - \bar{q})}{\sqrt{\sum_{k=1}^m (p_k - \bar{p})^2 \cdot \sum_{k=1}^m (q_k - \bar{q})^2}}$$

