Aprendizado de Máquina

Aula 3: Algoritmos baseados em distância

André C. P. L. F de Carvalho
ICMC/USP
andre@icmc.usp.br

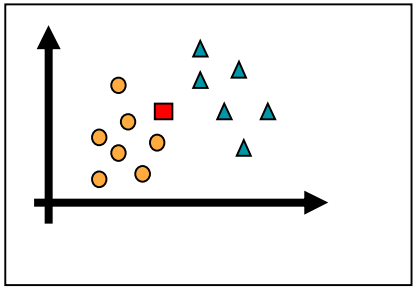
Tópicos

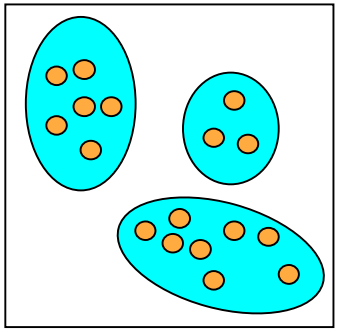
- Aprendizado baseado em instâncias
- 1-vizinho mais próximo
- Medidas de distância
- Similaridade e dissimilaridade
- K-vizinhos mais próximos
- Raciocínio baseado em casos

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

AM e Geometria

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





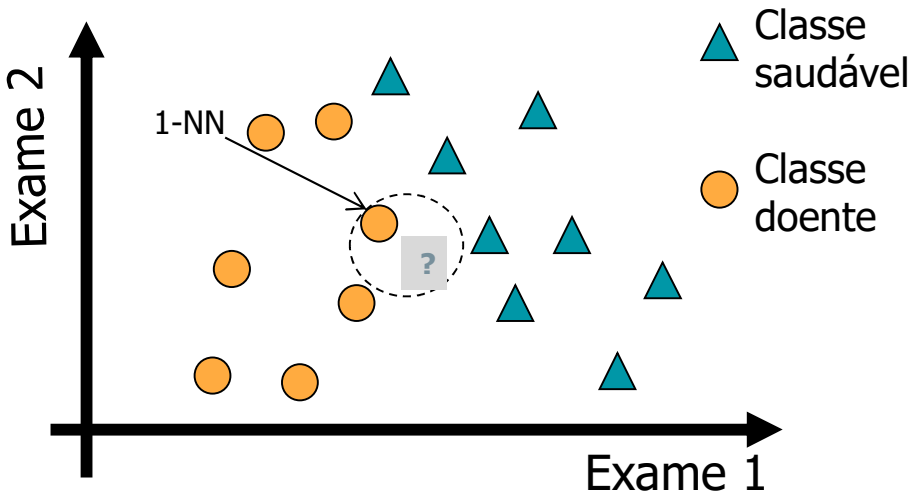
Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

1-vizinho mais próximo

- Versão simples do algoritmo k-NN
  - K-Nearest neighbour
  - 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 *eager* (ansioso)
    - Induzem um modelo
      - Ex.: Algoritmos de indução de árvores de decisão, redes neurais, máquinas de vetores de suporte, ...

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

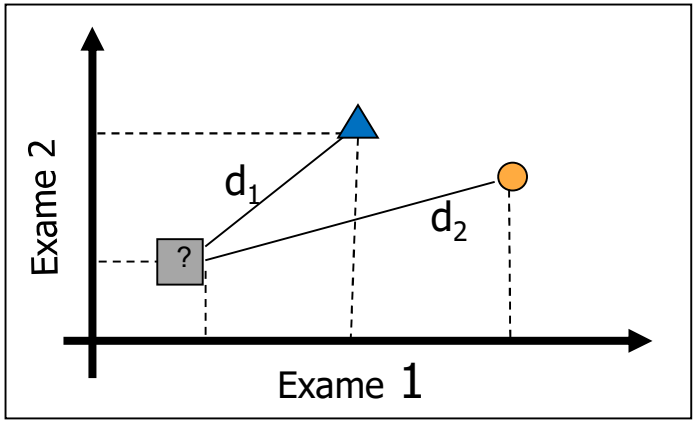
1-vizinho mais próximo



Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Métodos baseados em distância

- Consideram proximidade entre dados
  - Medidas de similaridade
  - Medidas de dissimilaridade (distância)



- Medidas mais usadas:
  - Euclidiana
  - Norma máxima
  - Bloco-cidade (Manhattan)
  - ...

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Distância de Minkowski

- Medida de distância generalizada

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

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Medidas de distância

- Distância Euclidiana
  - Sistema de coordenadas cartesianas

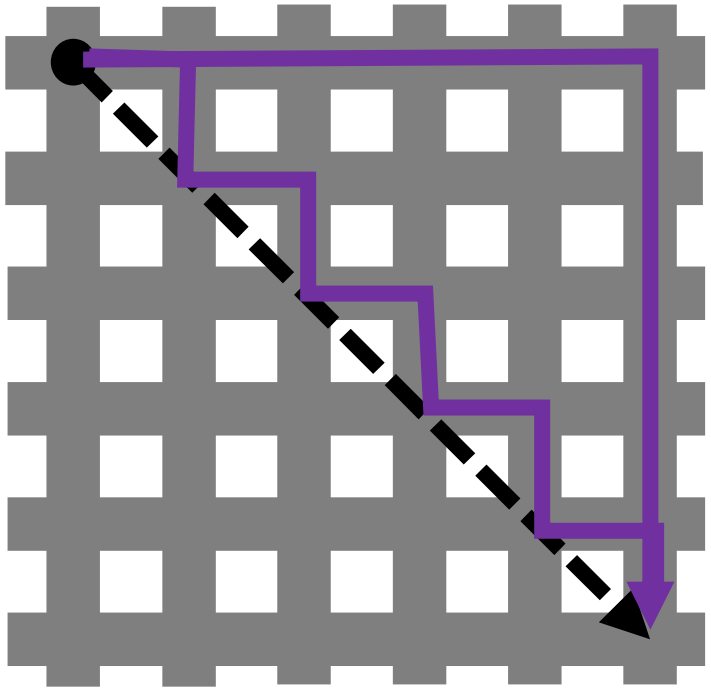
$$dist = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$

- Distância de norma máxima
  - Menor complexidade (e exatidão)

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Medidas de distância



Distância Euclidiana

Distância Bloco Cidade (Manhattan)

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Distância entre objetos

- Mede o quanto dois objetos são diferentes
  - Dissimilaridade (d)
    - Quanto mais diferentes, maior o valor
    - Nominal
    - Ordinal
    - Numérico

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

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

n = #valores
n > 1

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Similaridade entre vetores binários

- Coeficiente de Casamento Simples

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

- Coeficiente Jaccard

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Similaridade cosseno

- Muito usada quando dados são textos
  - *Bag of words*
    - Grande número de atributos
    - Esparsos
- Sejam p e q vetores representando documentos
  - $\cos(p, q) = (p \cdot q) / ||p|| \, ||q||$
    - $\cdot$: produto interno entre vetores
    - $||x||$: tamanho (norma) do vetor x

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

K-vizinhos mais próximos

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

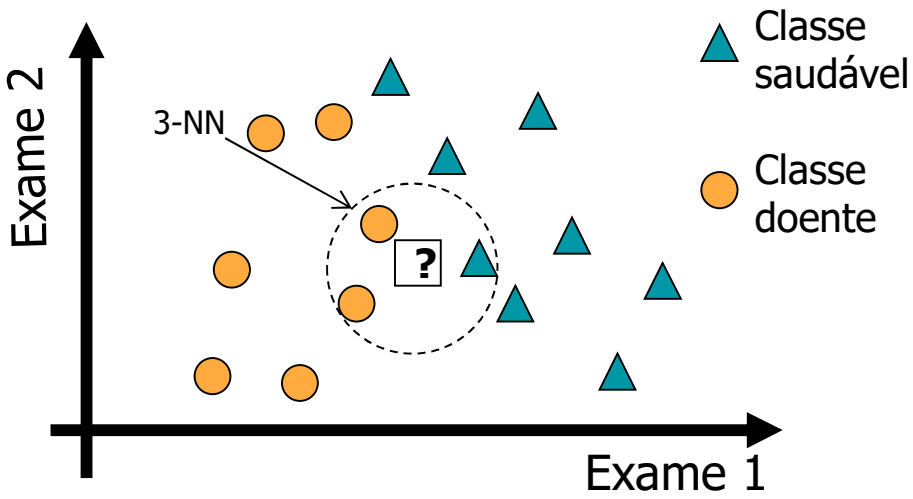
Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Quantos vizinhos?

- K muito grande
  - Vizinhos podem ser muito diferentes
    - Aumenta incerteza na classificação
  - Predição tendenciosa para classe majoritária
- K muito pequeno
  - Considera apenas objetos muito próximos
    - Não usa quantidade suficiente de informação
  - Previsão pode ser instável
    - Ruído

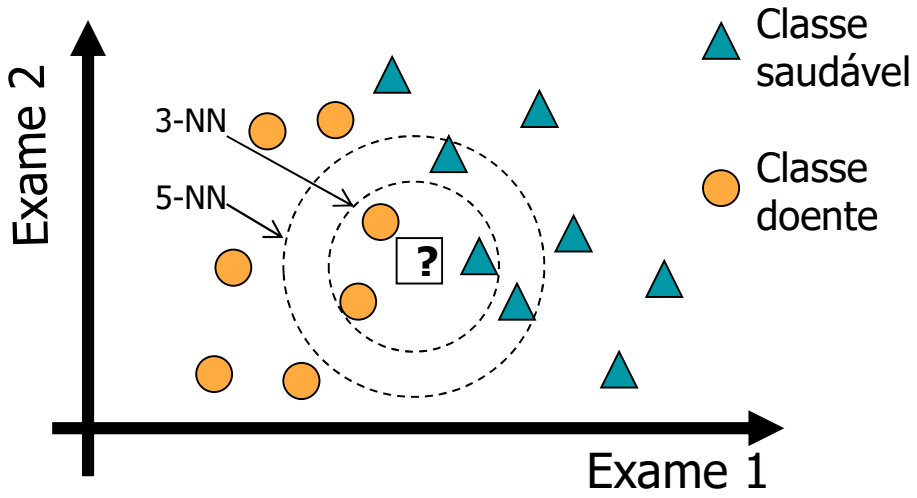
Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Quantos vizinhos?



Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Quantos vizinhos?



Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

K-Vizinhos mais próximos

Seja k o número de vizinhos mais próximos
Para cada novo exemplo x
Retornar a classe dos k exemplos (vizinhos) mais próximos
Classificar x na classe majoritária dentre as retornadas

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

K-vizinhos mais próximos

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

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

K-vizinhos mais próximos

- Algoritmos iterativos para eliminação
  - Selecionam protótipos
  - Eliminação sequencial
    - Conjunto inicial começa com todos os exemplos
    - Descarta exemplos corretamente classificados pelos protótipos (- protótipos)
  - Inserção sequencial
    - Conjunto inicial inclui apenas os protótipos
    - Inclui exemplos incorretamente classificados pelos protótipos (+ protótipos)

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

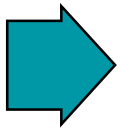
K-vizinhos mais próximos

- Normalizar atributos
- Ponderar atributos pela importância
- Ponderar votos pela distância entre exemplos
- Regressão
- Naturalmente incremental

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Raciocínio baseado em casos (RBC)

- Sistemas Baseados em Regras
  - Populares no passado
  - Dificuldade de especialistas em transformar experiência em regras



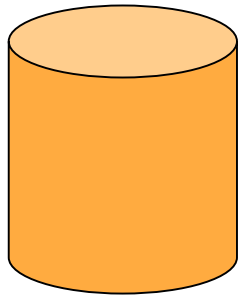
If ....
Then ...
Else...

EXPERIÊNCIA **REGRAS**

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Raciocínio baseado em casos (RBC)

EXPERIÊNCIA



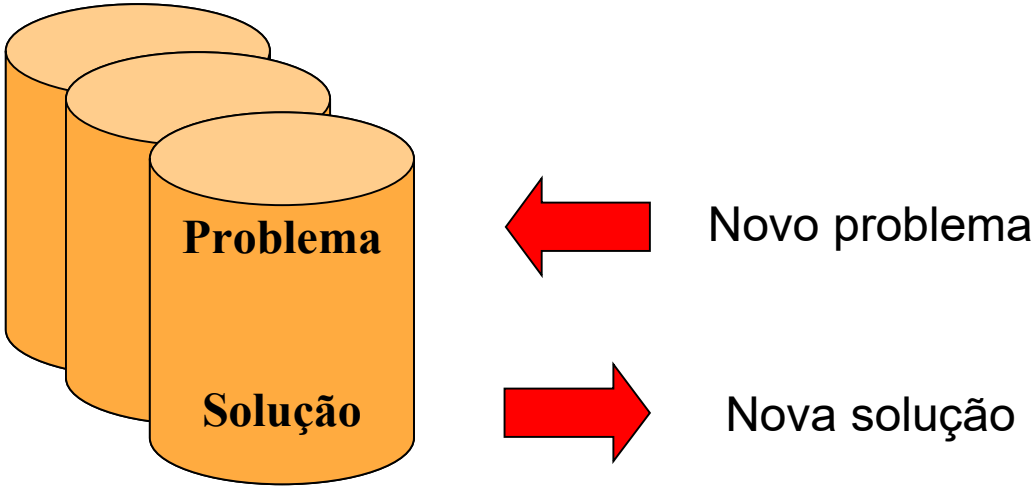
BASE DE EXPERIÊNCIAS

Mais que uma Base de Dados!

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

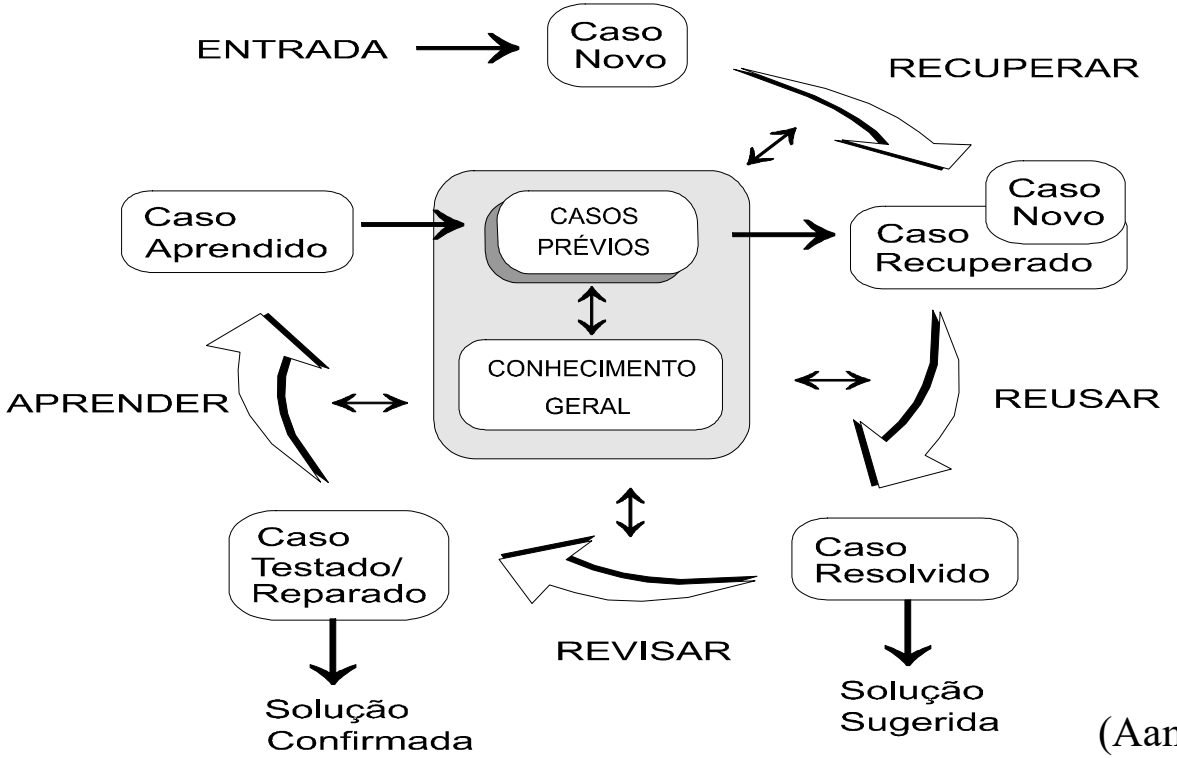
Como funciona

- Resolve novos problemas adaptando soluções de problemas anteriores semelhantes



Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Ciclo de um sistema de RBC



(Aamodt, 1993)

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

O que é um caso?

- Existem dois tipos de casos
  - Casos de entrada:
    - Descrições de características (situações) de problemas específicos
  - Casos armazenados:
    - Possuem descrições de características (situações) de problemas anteriores junto com soluções e resultados

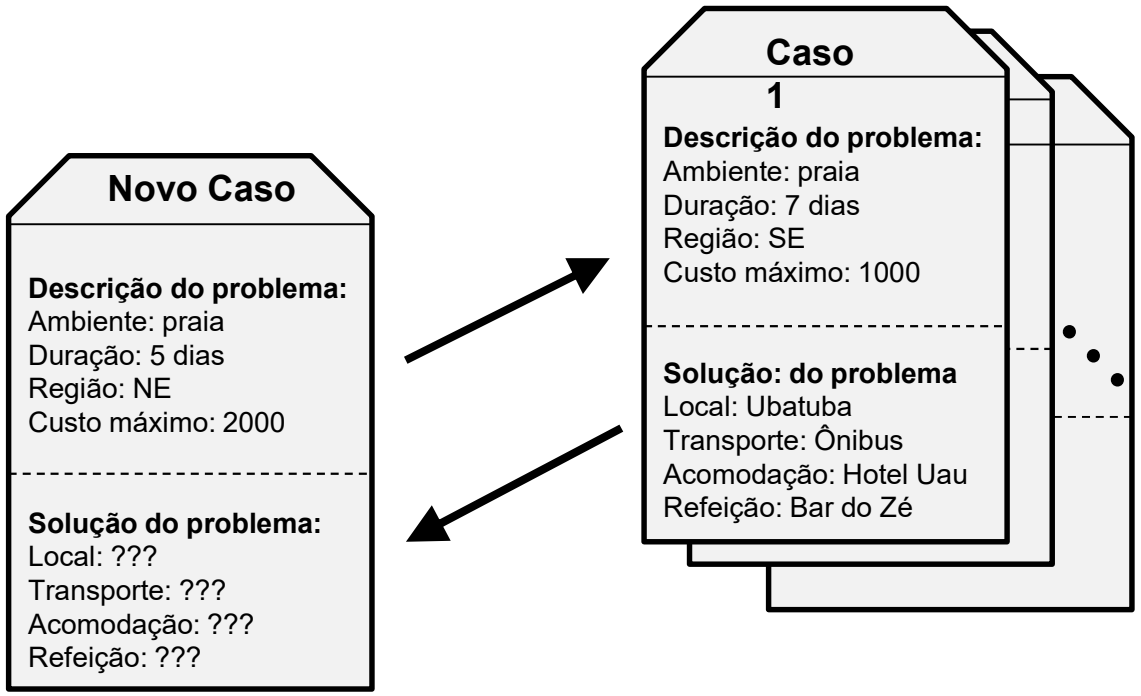
Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

O que é um caso?

- Um caso armazenado geralmente tem duas partes:
  - Uma parte caso
    - Descrição do problema
    - Usada para identificar o caso
      - Indexação e recuperação
  - Uma parte solução
    - Explica como este caso foi resolvido anteriormente de forma bem (mal) sucedida
    - Adaptada quando o caso é recuperado

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Raciocínio baseado em casos



Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Conclusão

- Aprendizado baseado em distância
- Conceitos básicos
- Medidas de distância
- K-vizinhos mais próximos
- Variações
- Exemplos
- Raciocínio baseado em casos

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização

Final da Apresentação

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização