

**Aprendizado de Máquina**

# **Aula 3: Algoritmos baseados em distância**

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

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

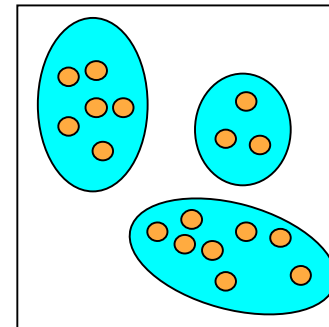
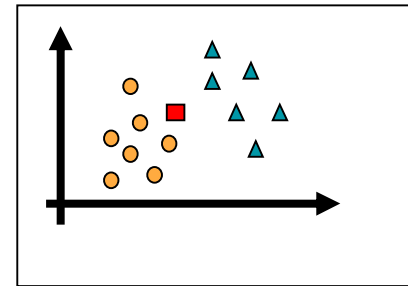


# Tópicos

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

# AM e Geometria

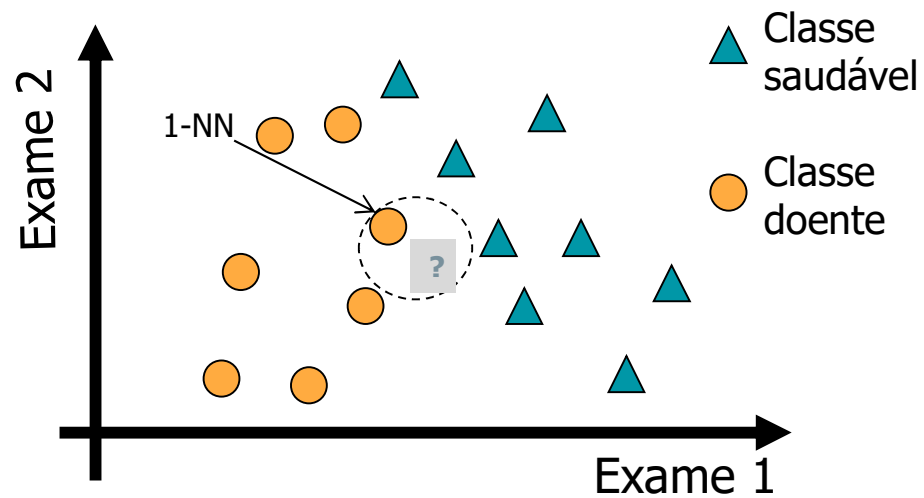
- Medidas de distância
  - Podem ser usadas para
    - Classificar novos dados
      - Ex.: K-NN
    - Agrupar dados
      - Ex.: K-médias
  - Existem várias medidas



# 1-vizinho mais próximo

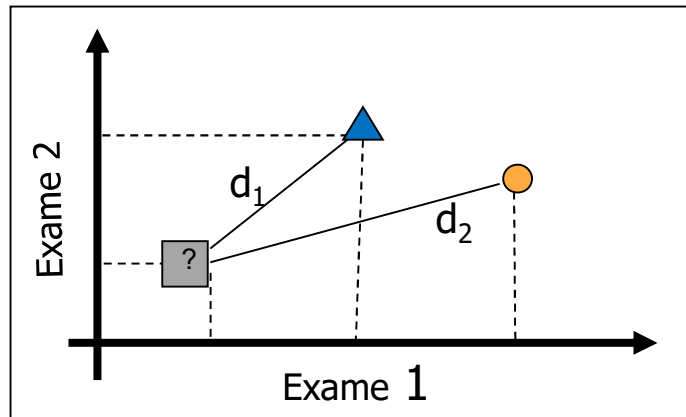
- Versão simples do algoritmo k-NN
  - K-Nearest neighbour
  - Geralmente usado para classificação
- Algoritmo *lazy* (preguiçoso)
  - Olha os dados de treinamento apenas quando vai classificar um novo objeto
  - Não constrói um modelo explicitamente
  - Diferente de algoritmos *eager* (ansioso)
    - Induzem um modelo
      - Ex.: Algoritmos de indução de árvores de decisão, redes neurais, máquinas de vetores de suporte, ...

# 1-vizinho mais próximo



# Métodos baseados em distância

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



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

■ ...



# 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

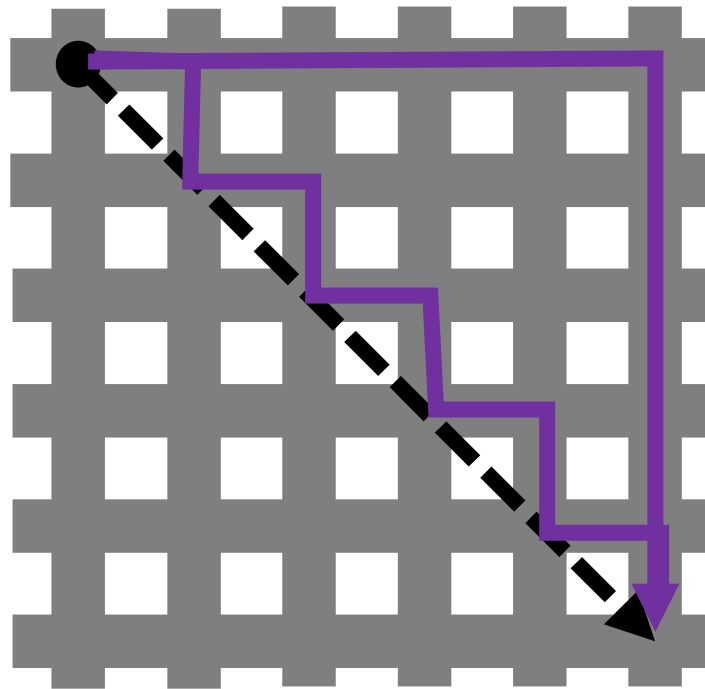
$$dist = \sqrt{\sum_{k=1}^m (p_k - q_k)^2}$$

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

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



# Medidas de distância



Distância Euclidiana



Distância Bloco Cidade (Manhattan)



# Proximidade entre valores

- Similaridade (s)

- Mede o quanto dois objetos são parecidos
  - Quanto mais parecidos, maior o valor

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

- Dissimilaridade (d)

- Mede o quanto dois objetos são diferentes
  - Distância

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

- Quanto mais diferentes, maior o valor

- Medida de proximidade pode ser usada

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

# 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

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

# 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  $x$

# K-vizinhos mais próximos

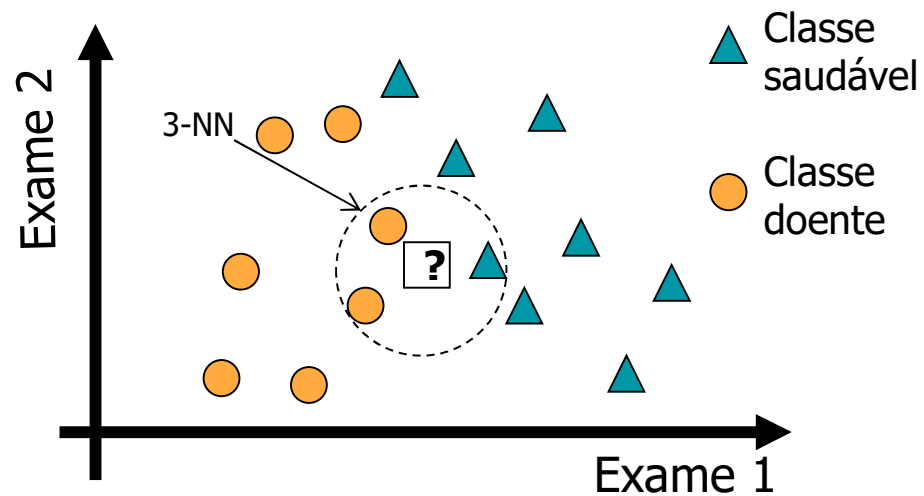
- Generalização do 1-vizinho mais próximo
- Algoritmo de AM baseado distância muito simples
  - Baseado em memória
- Número de vizinhos ( $k$ ) pode variar



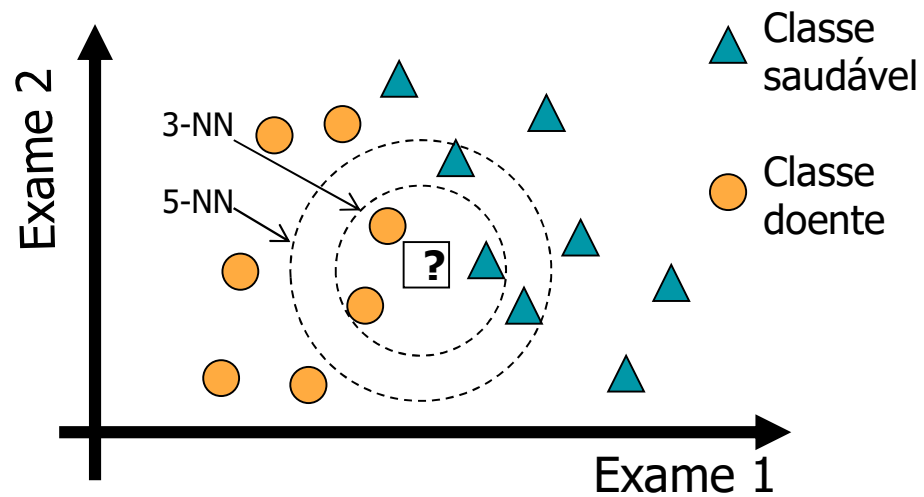
# Quantos vizinhos?

- 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

# Quantos vizinhos?



# Quantos vizinhos?



# 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*

# 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

# K-vizinhos mais próximos

- Algoritmos iterativos para eliminação
  - Seleccionam 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)



# K-vizinhos mais próximos

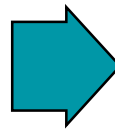
- Normalizar atributos
- Ponderar atributos pela importância
- Ponderar votos pela distância entre exemplos
- Regressão
- Naturalmente incremental

# Raciocínio baseado em casos (RBC)

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



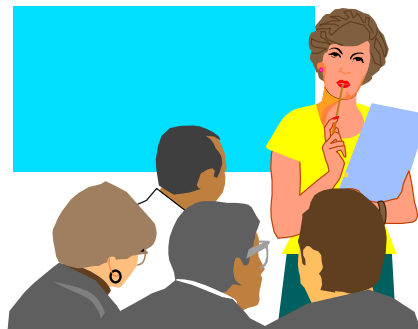
**EXPERIÊNCIA**



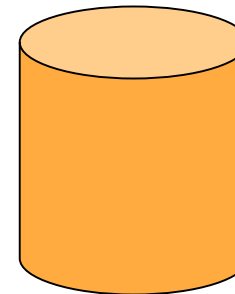
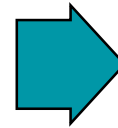
*If ....  
Then ...  
Else...*

**REGRAS**

# Raciocínio baseado em casos (RBC)



**EXPERIÊNCIA**

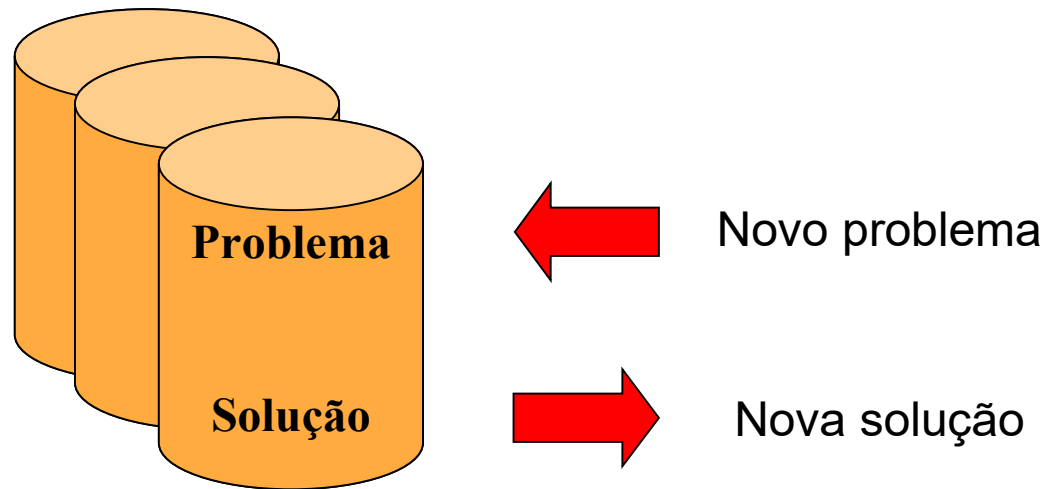


**BASE DE  
EXPERIÊNCIAS**

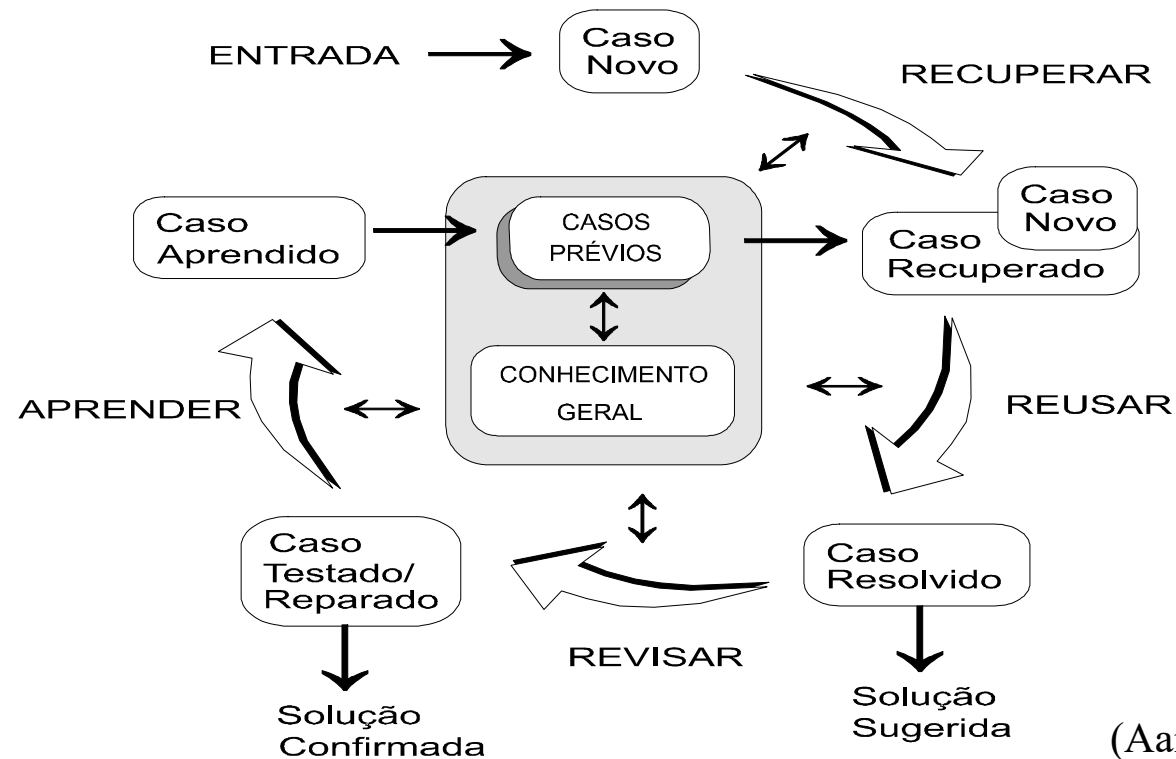
Mais que uma Base de Dados!

# Como funciona

- Resolve novos problemas adaptando soluções de problemas anteriores semelhantes



# Ciclo de um sistema de RBC



(Aamodt, 1993)

# O que é um caso?

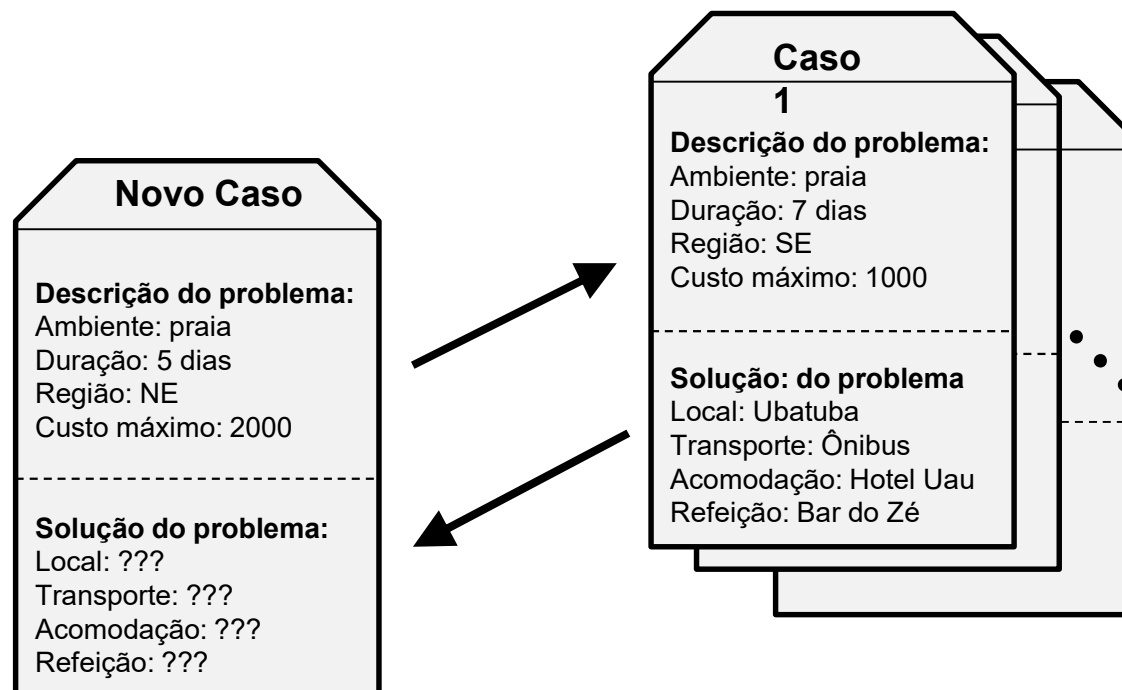
- Existem dois tipos de casos
  - Casos de entrada:
    - Descrições de características (situações) de problemas específicos
  - Casos armazenados:
    - Possuem descrições de características (situações) de problemas anteriores junto com soluções e resultados



# 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

# Raciocínio baseado em casos



# Conclusão

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

# *Final da Apresentação*

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



# 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 a quantidade de seus valores na base decimal?
  - 110000, 111001, 000111, 001011, 100111, 101001



# 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

# Exercício

- Seja o seguinte cadastro de pacientes

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

# Exercício

- Usar K-NN e os exemplos anteriores para definir as classes dos exemplos de teste
  - Usar  $k = 1, 3$  e  $5$
- Exemplos de teste
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

## Exercício 2

- Dada a tabela abaixo, com  $k = 1$  e 3, definir a classe dos exemplos:
  - (RJ, Médio, 178, 2000)
  - (SP, Superior, 200, 800)

Estado	Escolaridade	Altura	Salário	Classe
SP	Médio	180	3000	A
RJ	Superior	174	7000	B
RS	Médio	180	600	B
RJ	Superior	100	2000	A
SP	Fundamental	178	5000	A
RJ	Fundamental	188	1800	A

- Coeficiente de Casamento Simples

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

- Coeficiente Jaccard

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