







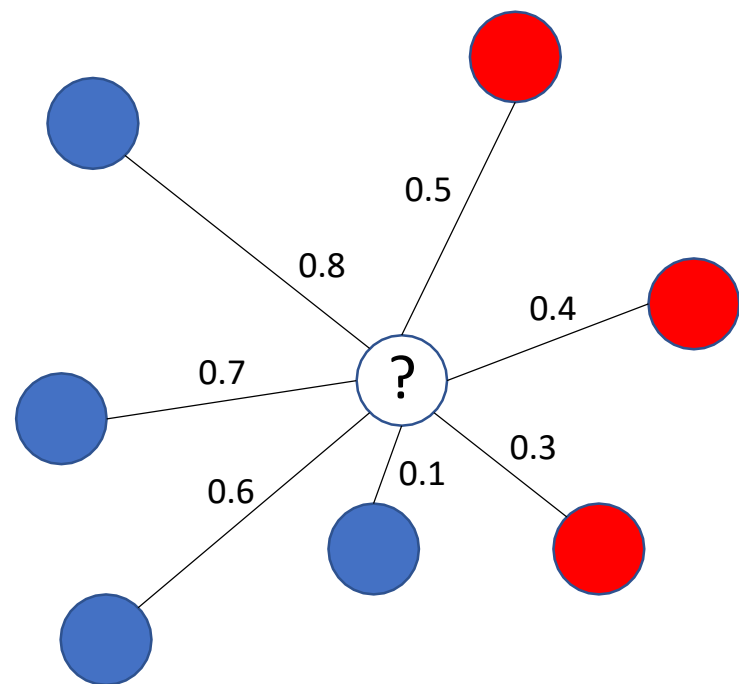
# Aprendizagem baseada em instâncias

Jones Granatyr



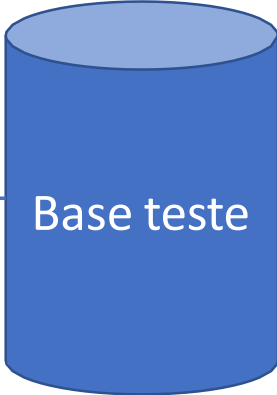
# K-Nearest Neighbour (kNN) – K vizinhos mais próximos

- K = 1 
- K = 2  
- K = 3 
- K = 4 
- K = 7 

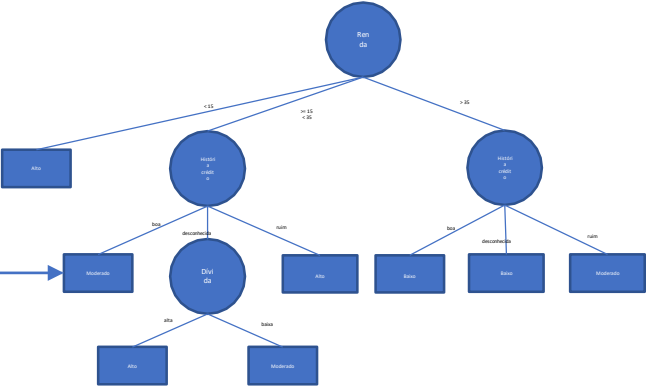


# kNN

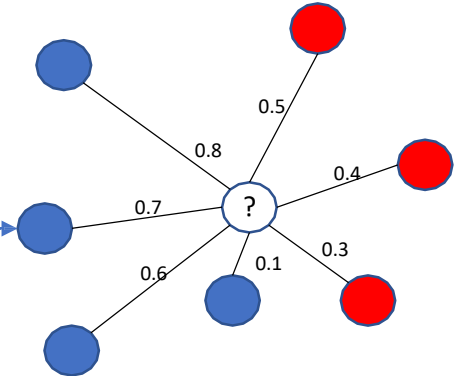
- A maioria dos métodos de aprendizagem constroem um modelo após o treinamento (os dados são descartados após a criação do modelo)
- Métodos baseados em instâncias simplesmente armazenam os exemplos de treinamento
- A generalização/previsão é feita somente quando uma nova instância precisa ser classificada (lazy)
- Paradigmas de aprendizagem de máquina



Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa	Desconhecida	Ruim	Alta	Baixa	Suficiente	Adequada	< 1500	1500 - 3500	> 3500
Alto	1/5	2/5	3/4	4/7	2/7	6/11	0	3/3	2/4	1/7
Moderado	1/5	1/5	1/4	1/7	2/7	2/11	1/3	0	2/4	1/7
Baixo	3/5	2/5	0	2/7	3/7	3/11	2/3	0	0	5/7



Regra	Resultado
Se renda = >35.000 E história_crédito = BOA	Risco = BAIXO
Se renda = >35.000 e história_crédito = DESCONHECIDA	Risco = BAIXO
Default (padrão)	Risco = ALTO



Naive bayes

Árvore de decisão

Regras

kNN

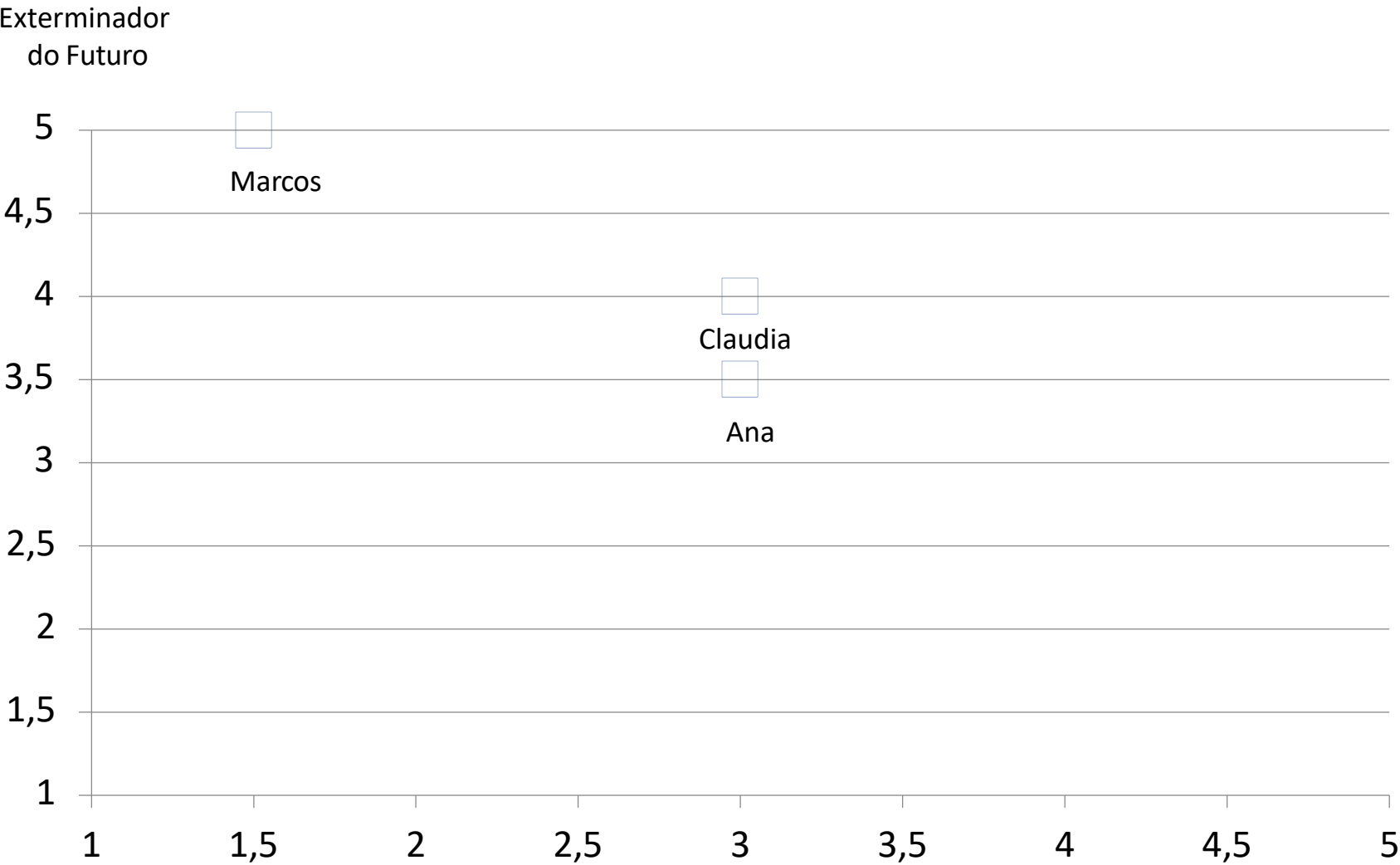
Registros  
% acerto

$$DE(x, y) = \sqrt{\sum_i^p (x_i - y_i)^2}$$

- $x = 5, 7, 9$
- $y = 5, 5, 5$
- Subtração de cada posição do vetor
  - $5 - 5 = 0$
  - $7 - 5 = 2$
  - $9 - 5 = 4$
- Elevação ao quadrado
  - $0^2 = 0$
  - $2^2 = 4$
  - $4^2 = 16$
- Somatório
  - $0 + 4 + 16 = 20$
- Raiz quadrada
  - $\text{Raiz}(20) = 4,47$
- **Distância Euclidiana = 4,47**

# Sistemas de recomendação (filtragem colaborativa)

$$DE(x,y) = \sqrt{\sum_i^p (x_i - y_i)^2}$$



Ana x Marcos

Ana x Cláudia

X = 3,0 3,5

X = 3,0 3,5

Y = 1,5 5,0

Y = 3,0 4,0

$(3,0 - 1,5)^2 = 2,25$

$(3,0 - 3,5)^2 = 0,25$

$(3,5 - 5,0)^2 = 2,25$

$(3,0 - 4,0)^2 = 1,00$

$2,25 + 2,25 = 4,5$

$0,25 + 1,00 = 1,25$

Raiz(4,5) = **2,12**

Raiz(1,25) = **1,11**

Filme	Violência	Romance	Ação	Comédia	Classe
Invocação do Mal	0,6	0,0	0,3	0,0	Terror
Floresta Maldita	0,9	0,0	0,5	0,1	Terror
Meu Passado me Condena	0,1	0,2	0,1	0,9	Comédia
Tirando o atraso	0,0	0,2	0,2	0,8	Comédia

$$DE(x,y)=\sqrt{\sum_i^p(x_i-y_i)^2}$$

Violência = 0.8	<b>Pesadelo x Invocação</b>	<b>Pesadelo x Floresta</b>
Romance = 0.1	0,8 0,1 0,5 0,0	0,8 0,1 0,5 0,0
Ação = 0.5	0,6 0,0 0,3 0,0	0,9 0,0 0,5 0,1
Comédia = 0.0		
A Hora do Pesadelo	$0,2^2 + 0,1^2 + 0,2^2 + 0$ $0,04 + 0,01 + 0,04 = 0,09$ Raiz(0,09) = <b>0,30</b>	$0,1^2 + 0,1^2 + 0 + 0,1^2$ $0,01 + 0,01 + 0,01 = 0,03$ Raiz(0,03) = <b>0,17</b>
	<b>Pesadelo x Passado</b>	<b>Pesadelo x Atraso</b>
	0,8 0,1 0,5 0,0	0,8 0,1 0,5 0,0
	0,1 0,2 0,1 0,9	0,0 0,2 0,2 0,8
	$0,7^2 + 0,1^2 + 0,4^2 + 0,9^2$ $0,49 + 0,01 + 0,16 + 0,8 = 1,46$ Raiz(1,46) = <b>1,20</b>	$0,8^2 + 0,1^2 + 0,4^2 + 0,8^2$ $0,64 + 0,01 + 0,16 + 0,64 = 1,45$ Raiz(1,45) = <b>1,20</b>

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto



História do crédito	Dívida	Garantias	Renda anual	Risco
3	1	1	1	Alto
2	1	1	2	Alto
2	2	1	2	Moderado
2	2	1	3	Alto
2	2	1	3	Baixo
2	2	2	3	Baixo
3	2	1	1	Alto
3	2	2	3	Moderado
1	2	1	3	Baixo
1	1	2	3	Baixo
1	1	1	1	Alto
1	1	1	2	Moderado
1	1	1	3	Baixo
3	1	1	2	Alto

História = Boa (1)

Dívida = Alta (1)

Garantias = Nenhuma (1)

Renda = > 35 (3)

**Novo x 9º**

1 1 1 3

1 2 1 3

0 + 1<sup>2</sup> + 0 + 0

0 + 1 + 0 + 0 = 1

Raiz(1) = **1**

**Novo x 3º**

1 1 1 3

2 2 1 2

1<sup>2</sup> + 1<sup>2</sup> + 0 + 1<sup>2</sup>

1 + 1 + 0 + 1 = 3

Raiz(3) = **1,7**

$$DE(x,y)=\sqrt{\sum_i^p(x_i-y_i)^2}$$

# kNN – variáveis na mesma escala

Idade	Renda anual
60	30.000
65	75.000
20	29.500

**1º x 2º**

60 30.000

65 75.000

$$5^2 + 45.000^2$$

$$25 + 2.025.000.000 = 2.025.000.025$$

$$\text{Raiz}(2.025.000.000) = \mathbf{45.000}$$

$$DE(x, y) = \sqrt{\sum_i^p (x_i - y_i)^2}$$

**1º x 3º**

60 30.000

20 29.500

$$40^2 + 500^2$$

$$1.600 + 250.000 = 251.600$$

$$\text{Raiz}(251.600) = \mathbf{501,59}$$

# Normalização (Normalization)

$$x = \frac{x - \text{mínimo}(x)}{\text{máximo}(x) - \text{mínimo}(x)}$$

$$x = \frac{60 - 20}{65 - 20} = 0,80$$
$$x = \frac{35 - 20}{65 - 20} = 0,30$$
$$x = \frac{20 - 20}{65 - 20} = 0,00$$

$$x = \frac{30.000 - 29.500}{45.000 - 29.500} = 0,03$$
$$x = \frac{45.000 - 29.500}{45.000 - 29.500} = 1,00$$
$$x = \frac{29.500 - 29.500}{45.000 - 29.500} = 0,00$$

Idade	Renda anual
60	30.000
35	45.000
20	29.500

Idade	Renda anual
0,80	0,03
0,30	1,00
0,00	0,00

1º x 2º

0,80 0,03  
0,30 1,00  
0,50² + 0,97²  
0,25 + 0,940 = 1,19  
Raiz(1,19) = **1,09**

1º x 3º

0,80 0,03  
0,00 0,00  
0,80² + 0,03²  
0,64 + 0,0009 = 0,6409  
Raiz(0,6409) = **0,80**

# Padronização (Standardization)

$$x = \frac{x - média(x)}{desvio\ padrão(x)}$$

$$x = \frac{60 - 38,33}{20,20} = 1,07$$
$$x = \frac{35 - 38,33}{20,20} = -0,16$$
$$x = \frac{20 - 38,33}{20,20} = -0,90$$

$$x = \frac{30.000 - 34.833,33}{8.808,14} = -0,54$$
$$x = \frac{45.000 - 34.833,33}{8.808,14} = 1,15$$
$$x = \frac{29.500 - 34.833,33}{8.808,14} = -0,60$$

Idade	Renda anual	Idade	Renda anual
60	30.000	1,07	-0,54
35	45.000	-0,16	1,15
20	29.500	-0,90	-0,60

Idade  
Média = 38,33  
Desvio padrão = 20,20

**1º x 2º**  
0,80 0,03  
0,30 1,00  
0,50² + 0,97²  
0,25 + 0,940 = 1,19  
Raiz(1,19) = **1,09**

Renda  
Média = 34.833,33  
Desvio padrão = 8.808,14  
**1º x 3º**  
1,07 -0,54  
-0,90 -0,60  
1,97² + 0,006²  
3,88 + 0,000036 = 3,880036  
Raiz(3,880036) = **1,96**

# kNN

- Algoritmo simples e poderoso
- Indicado quando o relacionamento entre as características é complexo
- Valor de k pequeno: dados com ruídos ou outliers podem prejudicar
- Valor de k grande: tendência a classificar a classe com mais elementos (overfitting) – valor default 3 ou 5
- Lento para fazer as previsões
- Outras distâncias
  - Coeficiente de Pearson
  - Índice de Tanimoto
  - City Block

# Conclusão

