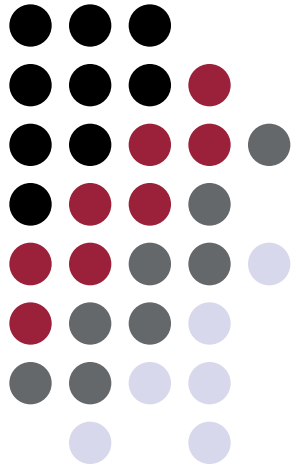
Mineração de Dados: Classificação

CEA462 – Sistemas de Apoio à Decisão

Prof. MSc. George H. G. Fonseca
Universidade Federal de Ouro Preto



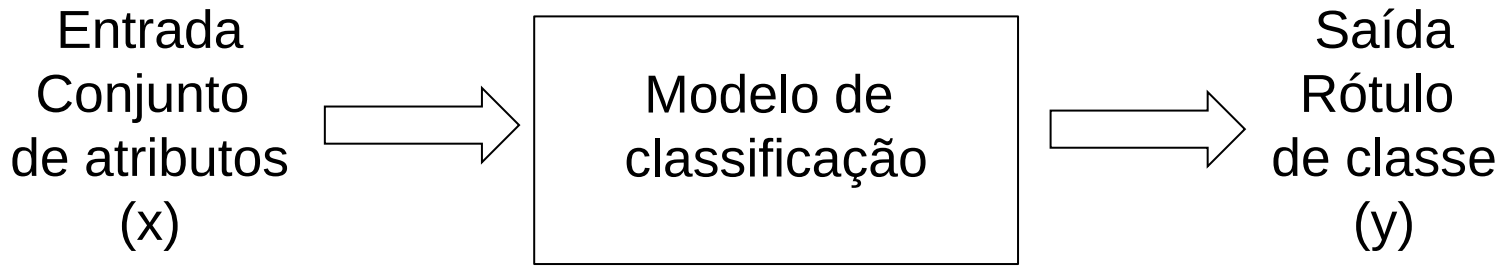
Introdução

- Classificação é a tarefa de organizar objetos em diferentes categorias
  - Detecção de mensagens spam (é spam ou não)
  - Classificação de células como benignas ou cancerígenas

Definições

- Um algoritmo de classificação recebe dados de entrada (exemplos)

- Cada exemplo (x, y) tem um conjunto de atributos (x) e um rótulo associado (y)



Definições

- Exemplo

Espécie	Temp. Corporal	Cobertura de pele	Ser aquático	Ser aéreo	Possui penas	Hiberna	Rótulo
Homem	Quente	Pelo	Não	Não	Sim	Não	Mamífero
Sapo	Fria	Nenhuma	Sim	Não	Sim	Sim	Anfíbio
Gato	Quente	Pelo	Não	Não	Sim	Não	Mamífero
Pinguim	Quente	Penas	Semi	Não	Sim	Não	Ave
Enguia	Fria	Escamas	Sim	Não	Não	Não	Peixe
Jacaré	Fria	Escamas	Semi	Não	Sim	Não	Réptil
Pomba	Quente	Penas	Não	Sim	Sim	Não	Ave

Definições

- Exemplo

Espécie	**Temp. Corporal**	**Cobertura de pele**	**Ser aquático**	**Ser aéreo**	**Possui penas**	**Hiberna**	**Rótulo**
Gila new	Fria	Escamas	Não	Não	Sim	Sim	**???**

Definições

- Formato de entrada de dados a ser considerado
  - csv (compatível com excel)
  - Cada exemplo compõe uma linha
  - Atributos são separados por “;”
  - A classe é o último elemento de cada linha (exemplo)

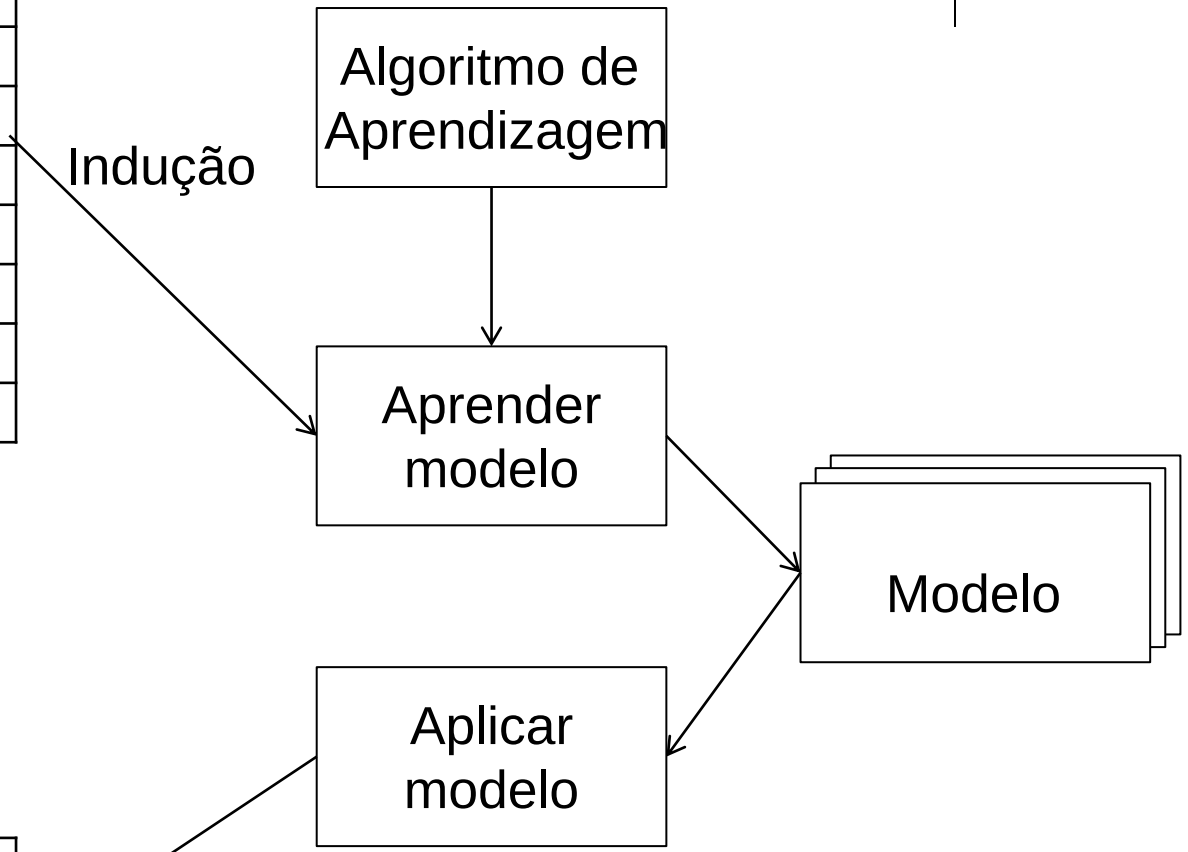
Espécie	**Temp. Corporal**	**Cobertura de pele**	**Ser aquático**	**Ser aéreo**	**Possui penas**	**Hiberna**	**Rótulo**
Homem	Quente	Pelo	Não	Não	Sim	Não	Mamífero

Homem;Quente;Pelo;Não;Não;Sim;Não;Mamífero

Abordagem Geral para Classificação

Conjunto de Treinamento

ID	Atb1	Atb2	Atb3	Classe
1	Sim	Grande	125	1
2	Não	Médio	100	1
3	Não	Pequeno	70	1
4	Sim	Médio	120	1
5	Não	Grande	95	2
6	Não	Médio	60	1
7	Sim	Grande	220	1
8	Não	Pequeno	85	2



Conjunto de Teste

ID	Atb1	Atb2	Atb3	Classe
9	Não	Pequeno	55	?
10	Sim	Médio	80	?
11	Sim	Grande	110	?

Abordagem Geral para Classificação

- Como avaliar um algoritmo de classificação?
  - Matriz de confusão
    - Acerto
    - Erro

		Classe prevista	
		1	0
Classe real	1	f_{11}	f_{10}
	0	f_{01}	f_{00}

Abordagem Geral para Classificação

- Como avaliar um algoritmo de classificação?

		Classe prevista	
		1	0
Classe real	1	f_{11}	f_{10}
	0	f_{01}	f_{00}

- $Precisão = \frac{Classificações\ corretas}{Total\ de\ classificações} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}}$

- $Taxa\ de\ erro = 1 - Precisão$

Classificador de Vizinho Mais Próximo

- "*Se caminhar como um pato, se grasnar como um pato e se parecer com um pato, então provavelmente é um pato*"

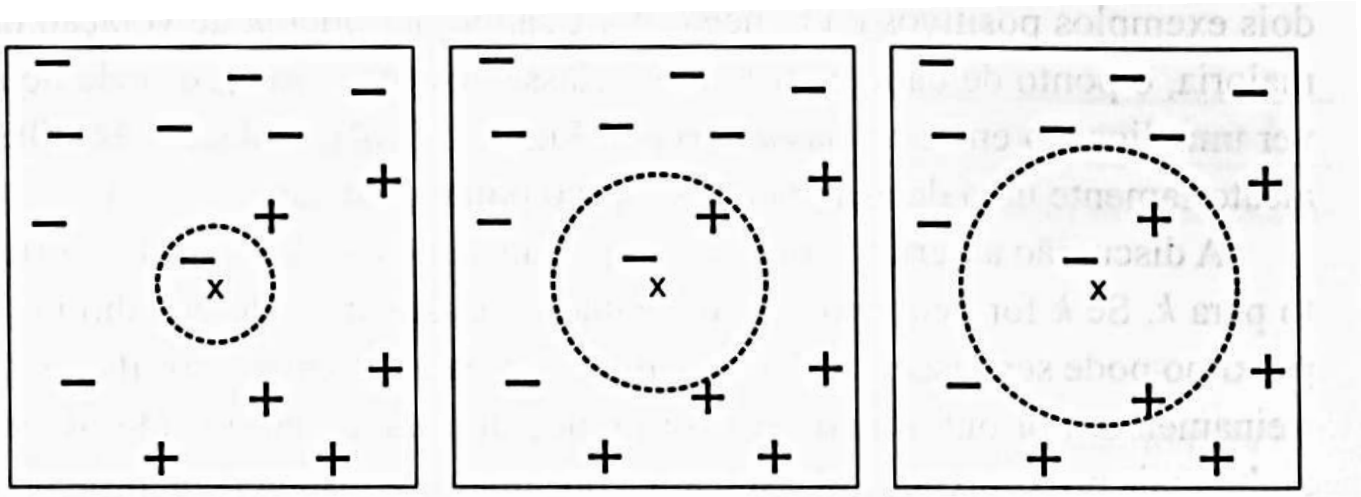
- Representa cada exemplo como um ponto em um espaço *d*-dimensional, onde *d* é o número de dimensões

- Dado um exemplo de teste, calcula-se a proximidade dele aos demais pontos de dados
  - O exemplo de teste será classificado com classe igual à classe predominante nos *k* pontos mais próximos

Classificador de Vizinho Mais Próximo

- Daí o nome K-NN (K-*Nearest Neighbors*)
  - Onde k é um parâmetro

- Exemplo num plano bidimensional



(a) 1-vizinho mais próximo (b) 2-vizinho mais próximo (c) 3-vizinho mais próximo

Classificador de Vizinho Mais Próximo

- Como calcular a proximidade entre dois pontos (exemplos)?
  - Distância euclidiana
    - $d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - yk)^2}$

Classificador de Vizinho Mais Próximo

- Algoritmo

classificadorKNN(número de vizinhos *k*, conjunto de treinamento D, conjuto de exemplos de teste T)

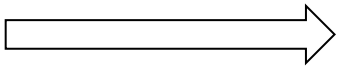
1. **para cada** exemplo de teste z ∈ T **faça**
2. Calcule d(z, x), a distância entre z e cada exemplo x ∈ D.
3. Selecione $D_z \subseteq$ D o conjunto dos *k* exemplos de treinamento para *z*
4. $classe_z$ = classe de maior incidência dentre $k \in D_z$
5. **fim para**

fimClassificador

Classificador de Vizinho Mais Próximo

- Exemplo
  - Normalização

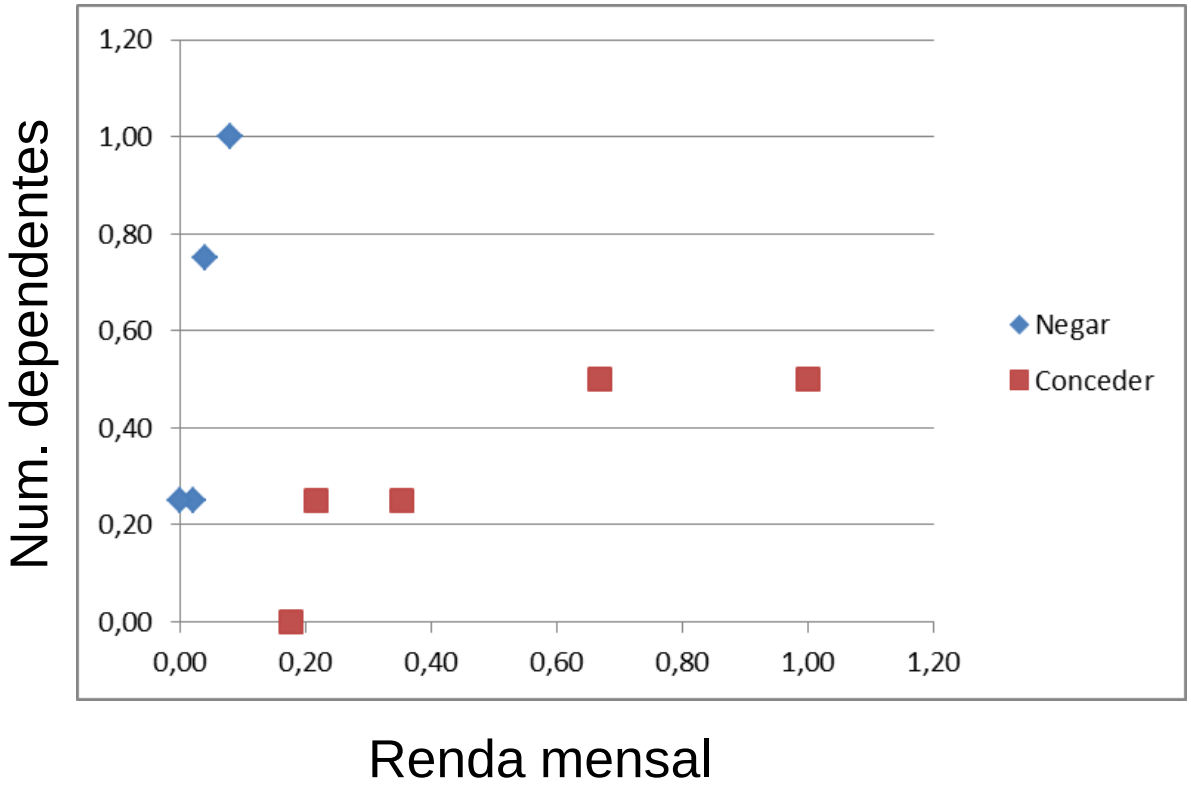
ID	Renda mensal	Depen-dentes	Classe
1	1000	1	1
2	6000	2	2
3	1300	4	1
4	2000	1	2
5	1100	3	1
6	1800	0	2
7	4300	2	2
8	0900	1	1
9	2700	1	2



ID	Renda mensal	Depen-dentes	Classe
1	0,02	0,25	1
2	1,00	0,50	2
3	0,08	1,00	1
4	0,22	0,25	2
5	0,04	0,75	1
6	0,18	0,00	2
7	0,67	0,50	2
8	0,00	0,25	1
9	0,35	0,25	2

Classificador de Vizinho Mais Próximo

- Exemplo
  - Dados de treinamento no plano



Classificador de Vizinho Mais Próximo

- Exemplo
  - Teste

ID	Renda mensal	Depen-dentes	Classe
1	0,02	0,25	1
2	1,00	0,50	2
3	0,08	1,00	1
4	0,22	0,25	2
5	0,04	0,75	1
6	0,18	0,00	2
7	0,67	0,50	2
8	0,00	0,25	1
9	0,35	0,25	2

ID	Renda mensal	Depen-dentes	Classe
10	0,80	0,50	?
11	0,02	0,75	?
12	0,35	0,50	?

Classificador de Vizinho Mais Próximo

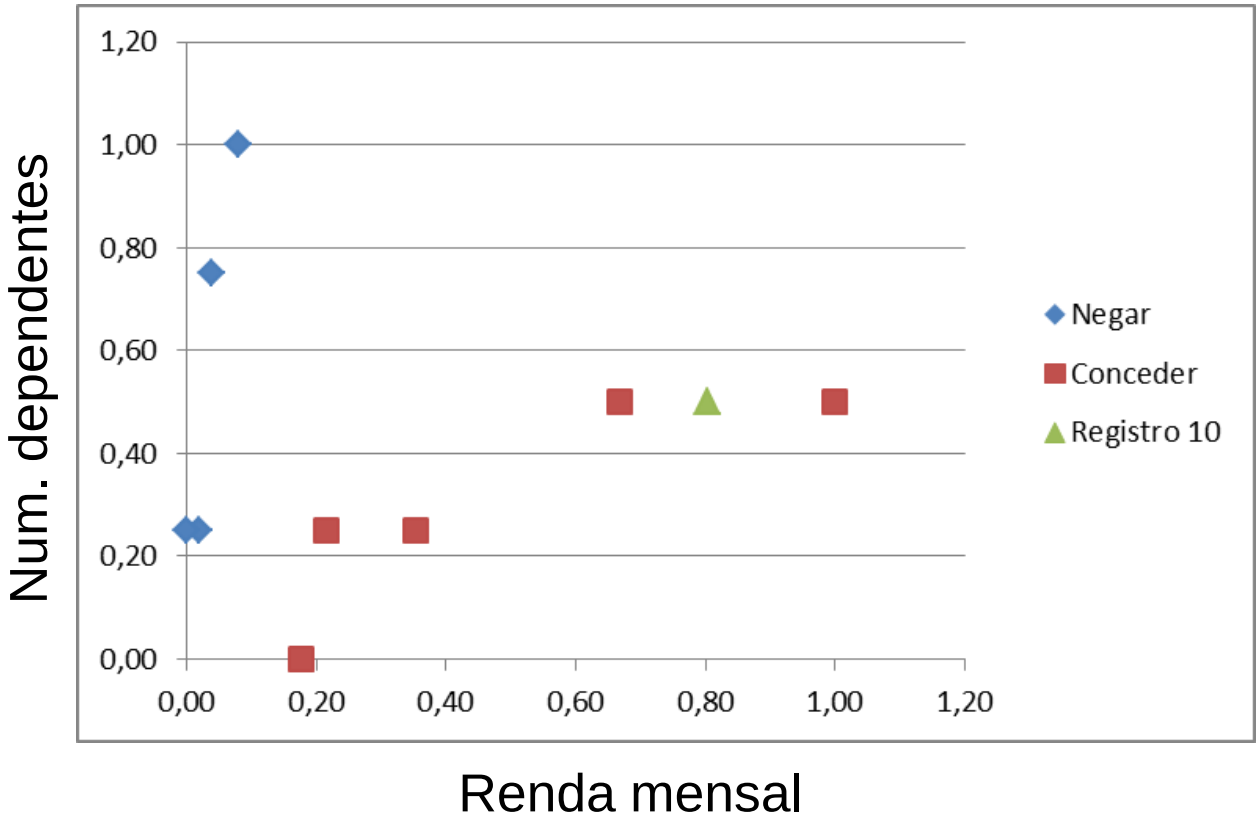
- Exemplo
  - Teste (considerando *k* = 3)

$$d(10,1) = \sqrt{(0{,}80 - 0{,}02)^2 + (0{,}50 - 0{,}25)^2} = 0{,}82$$
$$d(10,2) = \sqrt{(0{,}80 - 1{,}00)^2 + (0{,}50 - 0{,}50)^2} = \mathbf{0{,}19} \quad 2$$
$$d(10,3) = \sqrt{(0{,}80 - 0{,}08)^2 + (0{,}50 - 1{,}00)^2} = 0{,}88$$
$$d(10,4) = \sqrt{(0{,}80 - 0{,}22)^2 + (0{,}50 - 0{,}25)^2} = 0{,}64$$
$$d(10,5) = \sqrt{(0{,}80 - 0{,}04)^2 + (0{,}50 - 0{,}75)^2} = 0{,}80$$
$$d(10,6) = \sqrt{(0{,}80 - 0{,}18)^2 + (0{,}50 - 0{,}00)^2} = 0{,}80$$
$$d(10,7) = \sqrt{(0{,}80 - 0{,}67)^2 + (0{,}50 - 0{,}50)^2} = \mathbf{0{,}14} \quad 2$$
$$d(10,8) = \sqrt{(0{,}80 - 0{,}00)^2 + (0{,}50 - 0{,}25)^2} = 0{,}84$$
$$d(10,9) = \sqrt{(0{,}80 - 0{,}35)^2 + (0{,}50 - 0{,}25)^2} = \mathbf{0{,}52} \quad 2$$

Classe de 10 é 2

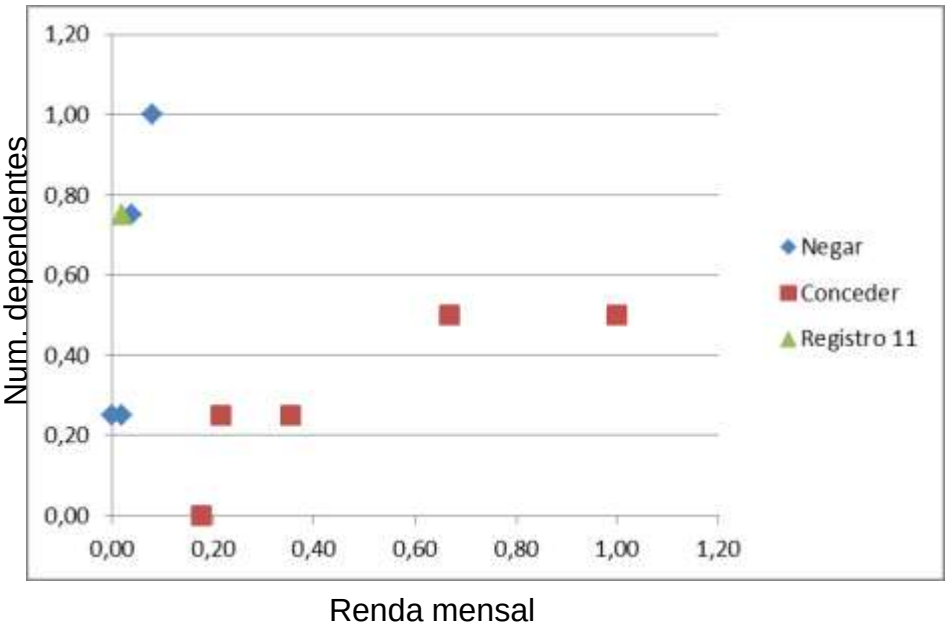
Classificador de Vizinho Mais Próximo

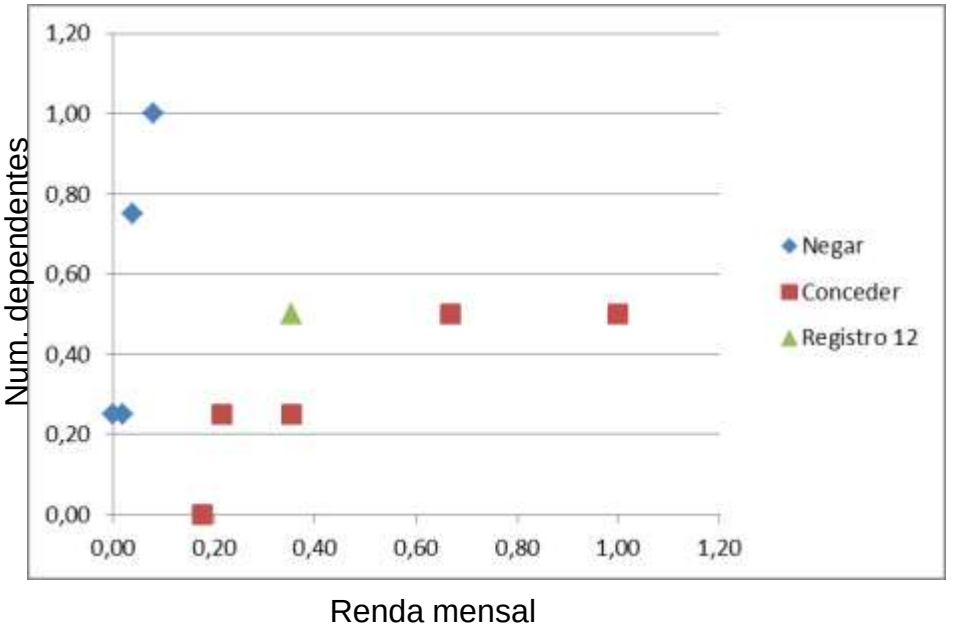
- Exemplo
  - Teste (considerando *k* = 3)



Classificador de Vizinho Mais Próximo

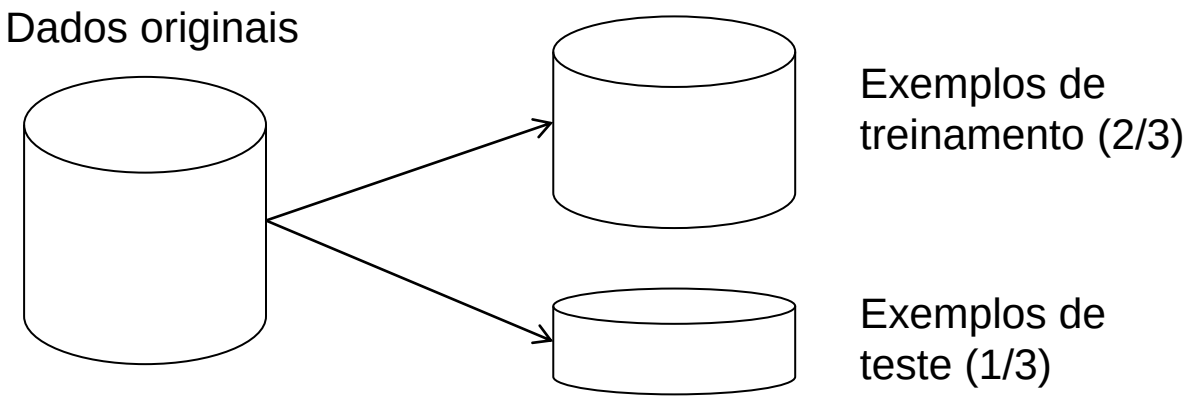
- Exemplo
  - Teste (considerando *k* = 3)





Avaliando o Desempenho de um Classificador

- Método *Holdout*
  - Dados originais são particionados em um conjunto de treinamento e outro de teste
  - 2/3 para treinamento e 1/3 para teste
  - Se o conjunto de dados for pequeno, poucos exemplos servirão para avaliar o desempenho
  - Pode ser tendencioso (e.g. nos 1/3 de teste haver apenas exemplos de uma classe)

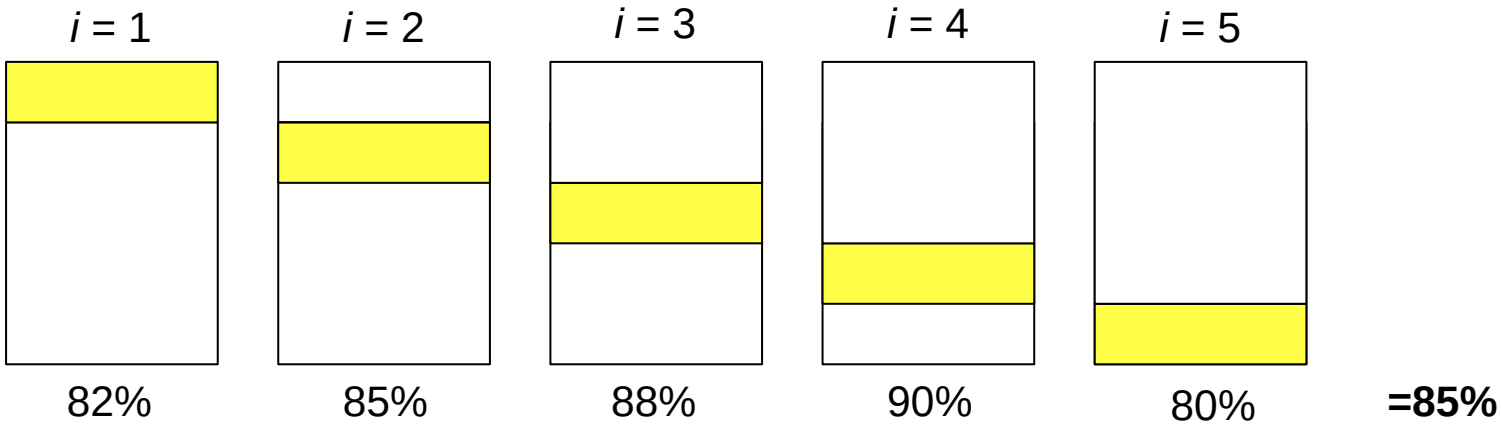


Avaliando o Desempenho de um Classificador

- Validação cruzada
  - Particiona o conjunto de dados em *k* partes (tipicamente *k* = 10)
  - A cada iteração *i*, o conjunto k_i é selecionado para teste e os demais para treinamento
    - A precisão é dada pela média da precisão nas *k* iterações

Avaliando o Desempenho de um Classificador

- Validação cruzada
  - Exemplo, considerando $k = 5$, onde exemplos de treinamento estão realçados em amarelo



  - Indicado quando há poucos exemplos disponíveis
    - Considera todos os dados no teste
    - Computacionalmente custoso: repetir a classificação k vezes

Bibliografia

- *Introdução ao Data Mining*. Steinbach, Michael; Kumar, Vipin; Tan, Pang-ning, Rio de Janeiro: Ed. Ciência Moderna, 2009. Capítulo 4 e 5.