

TÉCNICAS DE AGRUPAMENTO

Cristiane Neri Nobre

Agrupamento versus Classificação

- **Classificação**

- Aprendizado **Supervisionado**
 - Amostras de treinamento são classificadas
 - Número de Classes é conhecido
- Aprendizado por **Exemplo**

- **Agrupamento**

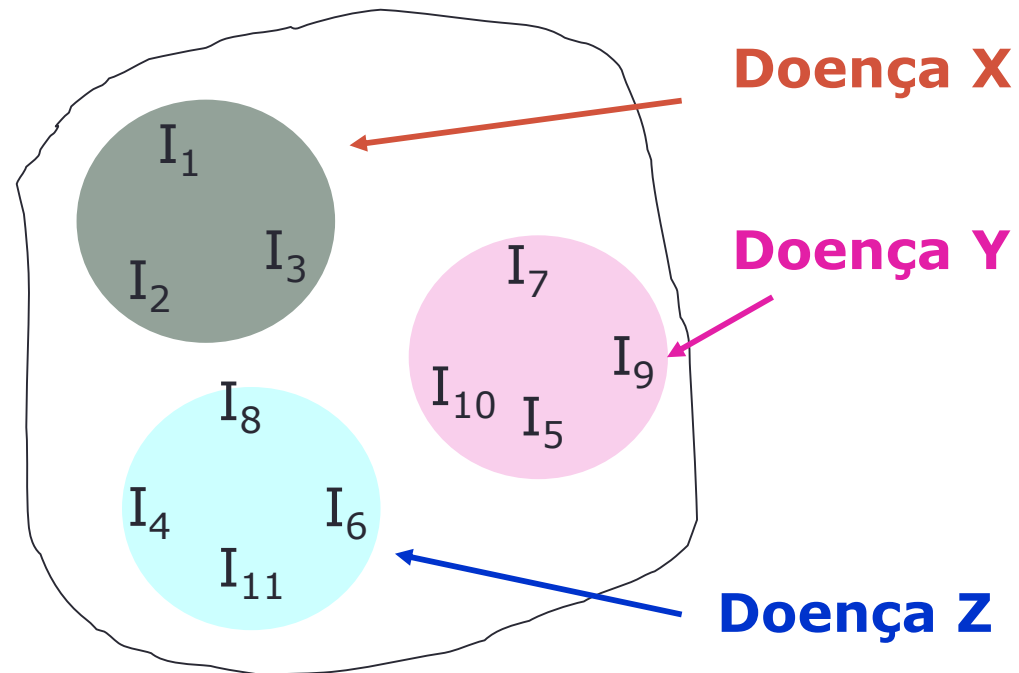
- Aprendizado **Não Supervisionado**
- Aprendizado por Observação

Agrupamento - Análise de Clusters

Número de Clusters = 3

	a_1	a_2	\dots	a_n		
I_1	a	F	1	0	1	1
I_2	b	M	0	0	1	1
\vdots	c	F	1	1	1	0
\vdots	d	F	1	0	0	0
I_n	e	M	1	1	0	1

nome gênero sintomas



Conceito = Doença

Agrupamento -Análise de Clusters

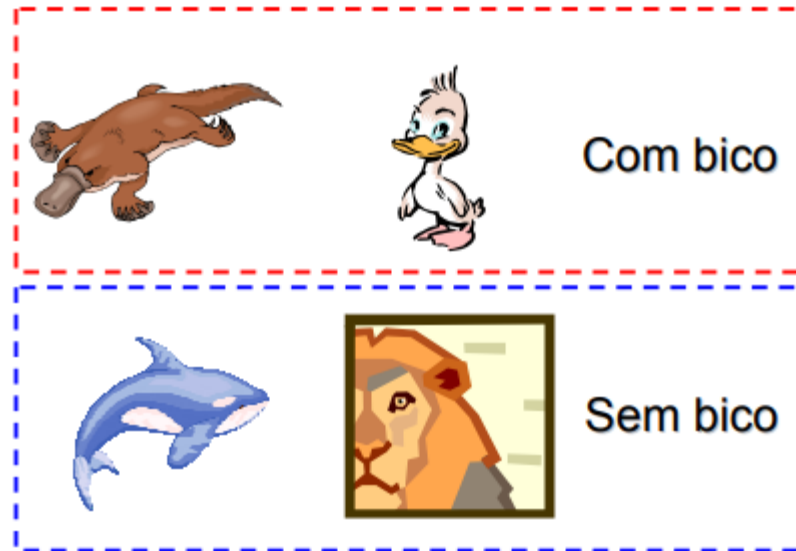
- Por exemplo, como agrupar estes animais*?



* Exemplo extraído de <http://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-Clustering.pdf>

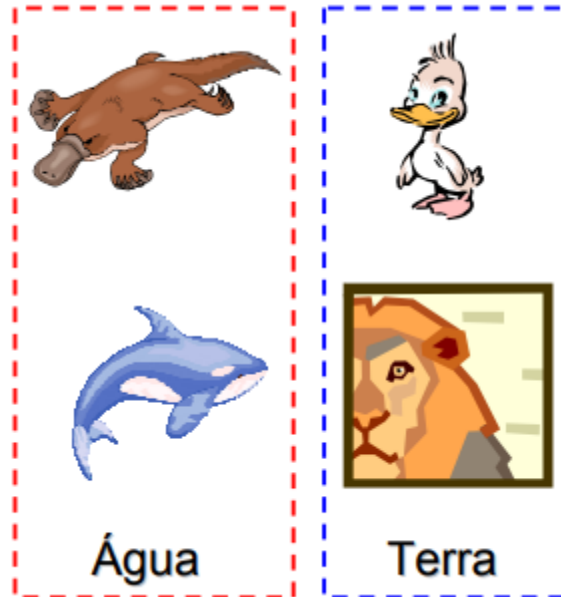
Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais?



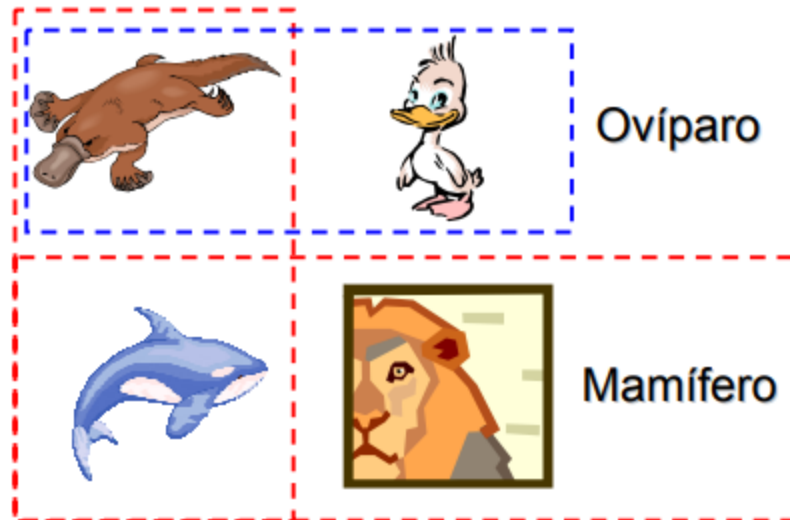
Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais?



Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais?



Agrupamento versus Classificação

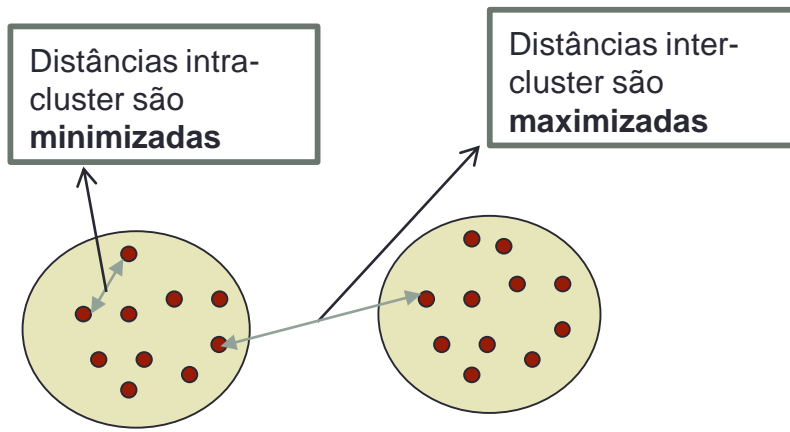
Objetivo do agrupamento:

Dado um conjunto de **instâncias** descritos por múltiplos **atributos**:

- 1) Atribuir grupos (clusters) aos objetos particionando-os objetivamente em grupos homogêneos de maneira a:
 - a) Maximizar a similaridade de objetos dentro de um mesmo cluster
 - b) Minimizar a similaridade de objetos entre clusters distintos
- 2) atribuir uma descrição (**rótulo**) para cada cluster formado

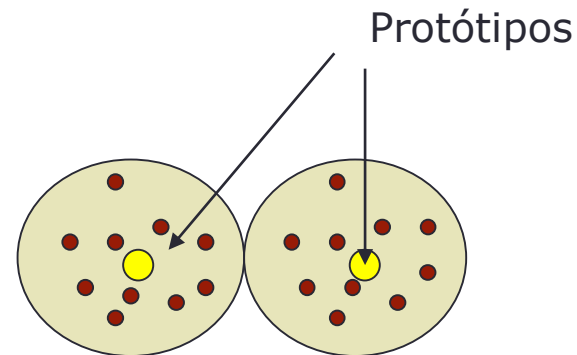
O que é um cluster ?

Como definir a noção de Cluster?



Bem separados

Um *cluster* é um conjunto de objetos no qual cada objeto está mais próximo (ou é mais similar) a objetos dentro do cluster do que qualquer objeto fora do cluster.



Baseados em Protótipos

Um *cluster* é um conjunto de objetos no qual cada objeto está mais próximo ao *protótipo que define o cluster* do que dos protótipos de quaisquer outros clusters.

Em geral: Protótipo = centróide

Como avaliar se as instâncias estão no mesmo grupo?

Pela distância!

Considere os pontos: $P_1 = (X_1, X_2, \dots, X_p)$, $P_2 = (Y_1, Y_2, \dots, Y_p)$, quais as distâncias entre estes dois pontos?

Manhattan

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_p - y_p|$$

Minkowski

$$d(x, y) = \sqrt[m]{(x_1 - y_1)^m + (x_2 - y_2)^m + \dots + (x_p - y_p)^m}$$

Distância Euclidiana

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2}$$

Investigue outras distâncias!

Exercício

- Sejam $X1 = (1,2)$ e $X2 = (4,6)$.
Calcule as distâncias euclidianas e Manhattan entre $X1$ e $X2$.

$$X1 = (1,2)$$

$$X2 = (4,6).$$

- $D_Manhattan(X1, X2) = |4-1| + |6-2| = 3 + 4 = 7$
- $D_Euclidiana(X1,X2) = ?$
- Ilustre no plano xy os segmentos representando tais distâncias.

Neste vídeo, aprendemos:

1. Sobre o funcionamento dos métodos de agrupamento

No próximo vídeo, veremos como **é o funcionamento do algoritmo K-means**, um algoritmo de agrupamento.