



# **Escola Politécnica de Pernambuco**

*Especialização em Ciência de Dados e Analytics*

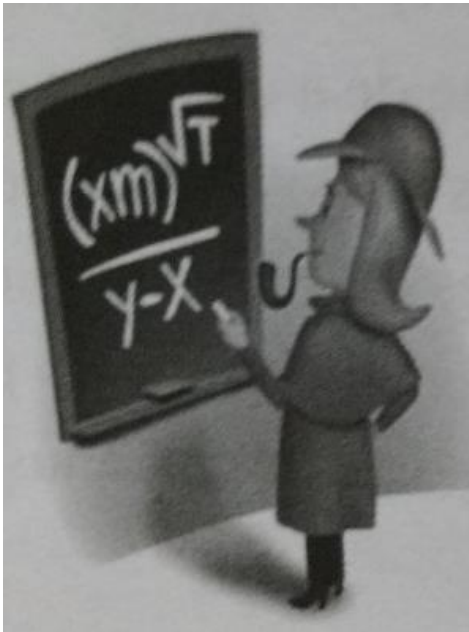
## **Introdução à Ciência de Dados**

### **Aula 4**

Prof. Dr. Alexandre Maciel  
***amam@ecomp.poli.br***

# MODELAGEM

---



- Representação deliberadamente simplificada do problema
- O modelo deve ser minimalista
- Todos os modelos são falhos, mas alguns são úteis

# ANÁLISE PREDITIVA DOS DADOS

---

**“... é a arte de se obter informação a partir de dados coletados e a utilizar para prever padrões de comportamento e tendências.”**

**Covington**

# APRENDIZAGEM DE MÁQUINA

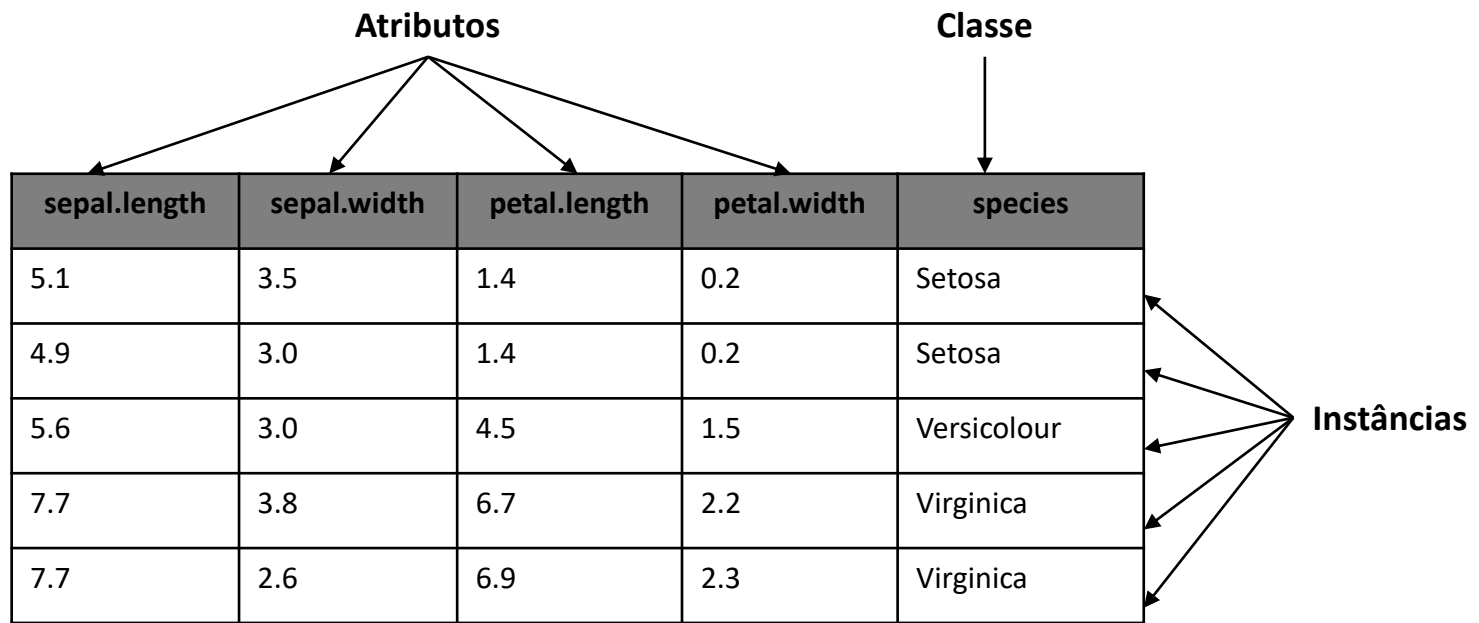
---

- Subárea da Inteligência Artificial
- Supervisionada e não supervisionada
- Tarefas:
  - Classificação
  - Regressão
  - Agrupamento
  - Associação
  - Detecção de Anomalias



# Classificação

- Tarefa supervisionada de aprender uma *função alvo*  $f$  que mapeie cada conjunto de atributos  $x$  para um dos rótulos de classes  $y$  pré-determinados.

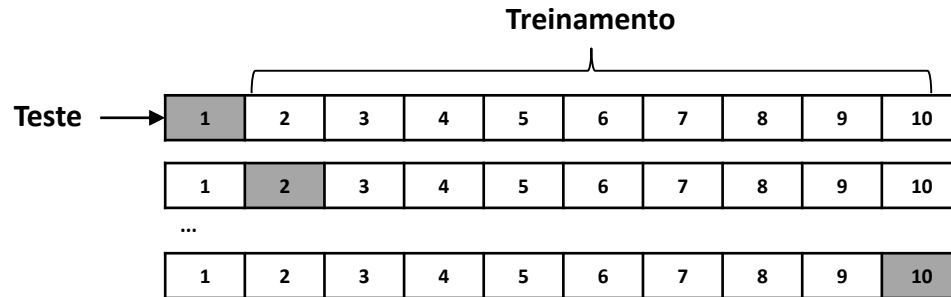


Atributos				Classe
sepal.length	sepal.width	petal.length	petal.width	species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
5.6	3.0	4.5	1.5	Versicolour
7.7	3.8	6.7	2.2	Virginica
7.7	2.6	6.9	2.3	Virginica

Instâncias

# Classificação

- Partição do conjunto de dados
  - **Divisão *hould-out***
    - Normalmente  $p = 2/3$  e  $(1 - p) = 1/3$
  - **Cross-validation**
    - Divisão do conjunto em  $K$  subconjuntos, com  $N$  elementos.



- **Bootstrap**
  - Conjunto de treinamento gerado a partir de  $N$  sorteios aleatórios.
  - Repetido várias vezes a fim de estimar a média de desempenho.

# Classificação

- Construção do modelo

Conjunto de treinamento

sepal.length	sepal.width	petal.length	petal.width	species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
5.6	3.0	4.5	1.5	Versicolour
6.4	3.2	4.5	1.5	Versicolour
7.7	3.8	6.7	2.2	Virginica
7.7	2.6	6.9	2.3	Virginica
6.7	2.5	5.8	1.8	Virginica

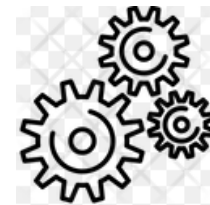
Conjunto de teste

sepal.length	sepal.width	petal.length	petal.width	species
5.0	3.4	1.6	0.4	?
5.6	3.0	4.5	1.5	?
7.7	2.8	6.7	2.0	?

Indução

Dedução

Aplicação dos Algoritmos



Avaliação dos Resultados

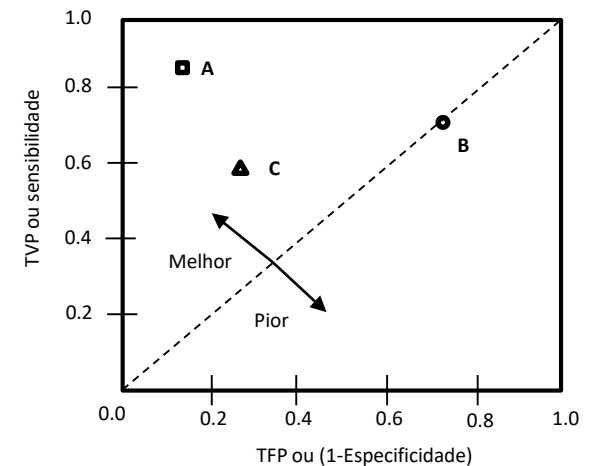
# Classificação

- Eficiência do classificador
  - **Accuracy**
    - Percentual de classificação correta.
  - **Confusion Matrix**
    - Matriz que relaciona classes desejadas com as preditas.
  - **Curva ROC**
    - Ferramenta gráfica para avaliar múltiplos classificadores.

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

$$TVP = \frac{VP}{VP + FN}$$

$$TFP = \frac{FP}{FP + FN}$$





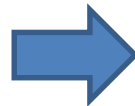
# Regressão

- Tarefa análoga à Classificação, formada por valores numéricos.
- A regressão pode ser: Linear ou Não Linear
  - Função linear
  - Função polinomial
- Regressão Linear Múltipla
  - $X_1, X_2, \dots, X_k \rightarrow$  várias variáveis independentes
  - $Y \rightarrow$  variável dependente (função linear das variáveis  $X_i$ )

# Regressão

- Exemplo:

X (experiência)	Y (salário)
03	30
08	57
09	64
13	72
03	36
06	43
11	59
21	90
01	20
16	83



$$\bar{x} = 9,1 \text{ e } \bar{y} = 55,4$$

$$\beta = \frac{(3-9,1)(30-55,4) + (8-9,1)(57-55,4) + \dots + (16-9,1)(83-55,4)}{(3-9,1)^2 + (8-9,1)^2 + \dots + (16-9,1)^2} = 3,7$$

$$\alpha = 55,4 - (3,7)(9,1) = 21,7$$

$$Y = 21,7 + 3,7 * X$$

# Agrupamento

- Tarefa de aprendizado de máquina não supervisionada.
- Busca reunir instâncias com atributos comuns em grupos, que, podem ser classificados posteriormente.
- Medidas de similaridade
  - Maximizar similaridade intra-grupo
  - Minimizar dissimilaridade inter-grupo

# Agrupamento

- Construção do modelo

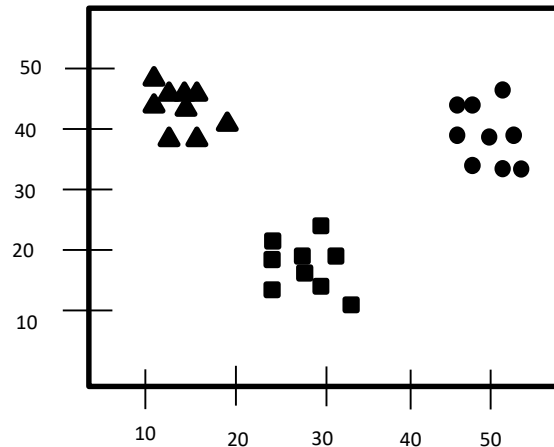
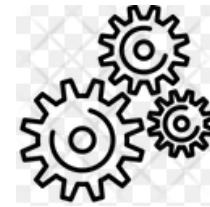
Conjunto de dados

sepal.length	sepal.width	petal.length	petal.width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
5.6	3.0	4.5	1.5
6.4	3.2	4.5	1.5
7.7	3.8	6.7	2.2
7.7	2.6	6.9	2.3
6.7	2.5	5.8	1.8
5.0	3.4	1.6	0.4
5.6	3.0	4.5	1.5
7.7	2.8	6.7	2.0

Definição da  
medida de  
similaridade



Aplicação dos  
Algoritmos



Avaliação  
dos Resultados

## Regras de Associação

- Consiste em encontrar conjuntos de itens que ocorram simultaneamente de forma frequente.
- Princípio da antimonotonicidade:
  - “Um  $k$ -itemset somente pode ser frequente se todos os seus  $(k-1)$  itemsets forem frequentes”.

TID	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cerveja, Refrigerante}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Refrigerante}



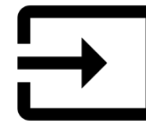
# Regras de Associação

- Construção do modelo:

Conjunto de dados

TID	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	0	1	0	1	1	0	0
2	1	0	1	1	1	0	0
3	0	1	0	1	1	0	0
4	1	1	0	1	1	0	0
5	0	0	1	0	0	0	0
6	0	0	0	0	1	0	0
7	0	0	0	1	0	0	0
8	0	0	0	0	0	0	1
9	0	0	0	0	0	1	1
10	0	0	0	0	0	1	0

Definição da  
suporte e  
confiança



## A) itens frequentes

### 1 – item sets

Café	0,3
Pão	0,5
Manteiga	0,5

### 2 – itens sets

Café, Pão	0,3
Café, Manteiga	0,3
Pão, Manteiga	0,4

### 3 – itens sets

Café, Pão, Manteiga	0,3
---------------------	-----



## b) Regras de associação

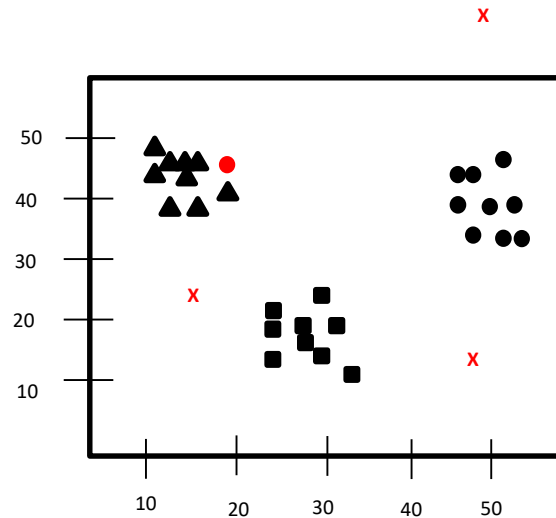
- Se café Então pão. -> Conf = 1,0
- Se café Então manteiga -> Conf = 1,0
- Se pão Então manteiga -> Conf = 0,8
- Se manteiga Então pão -> Conf = 0,8
- Se café, pão Então manteiga. -> Conf = 1,0
- Se café, manteiga Então pão -> Conf = 1,0
- Se café Então pão, manteiga -> Conf = 1,0



Avaliação  
dos Resultados

# Detecção de Anomalias

- Tarefa que destaca objetos que possuem atributos que desviam significativamente dos valores típicos.
- Anomalias, inconsistências e ruídos



# ALGORITMOS DE APRENDIZADO DE MÁQUINA

---

Tarefas	Técnicas	Algoritmos
Classificação	Bayes	NaiveBayes, Redes Bayesianas
	Árvore de Decisão	J48, Random Forest
	Redes Neurais	MLP, RBF, SVM
Agrupamento	Protótipo	K-means, k-medoides
	Grafos	Hierárquico
	Difuso	Fuzzy c-means
	Densidade	DBSCAN
Regressão		Linear, Logística
Regras de Associação		Apriori, FP Growth
Detecção de anomalias	Estatísticos, proximidade, redes neurais	



## Linguagens



## Ferramentas - IDEs



Rodeo



## Ferramentas de Análise



## Infraestrutura



# AVALIAÇÃO DOS RESULTADOS

---



- O que gostaríamos de alcançar com aprendizado de máquina?
- Muitas vezes não é possível medir perfeitamente a meta final
- Não se pode oferecer uma única métrica de avaliação para qualquer problema

# DINÂMICA

---

- Utilize as bases de dados do Weka para analisar uma técnica de classificação ou de agrupamento e descreva suas considerações.