

PROCESSAMENTO DE LINGUAGEM NATURAL

Aprendizado de Máquina

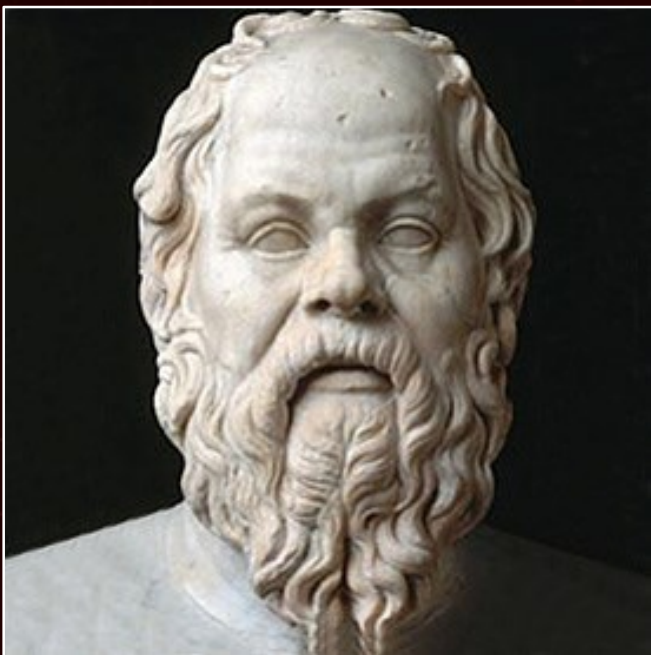


TÓPICOS

1. **Aprendizado**
2. **Paradigmas de AM**
3. **Classificação**
4. **Avaliação de Classificadores**



DEFINIÇÃO DE APRENDIZADO



Sócrates: Aprender é recordar
(Diálogos de Platão)

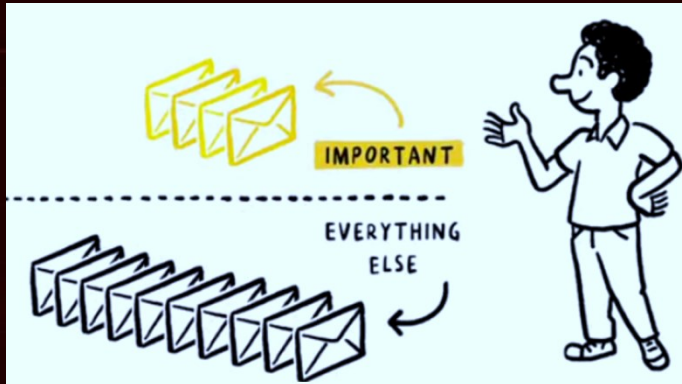
Definição clássica (Mitchell, 1997)

“Um programa de computador é dito **aprender** a partir de uma experiência E com respeito a alguma classe de tarefas T e medida de desempenho P , se seu desempenho em tarefas de T , medido por P , melhora com a experiência E .”

Tom Mitchell (1997)

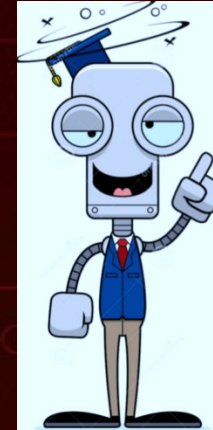
PARADIGMAS DE AM

- **Supervisionado**
- **Por reforço**
- **Não-supervisionado**
- **Semissupervisionado**



GUIADO POR “PROFESSOR” EXTERNO

- *Professor* possui conhecimento sobre a tarefa
- Representado por conjuntos de pares (x, d)
- Algoritmo de AM gera modelo que busca **reproduzir comportamento do professor**
- Parâmetros do modelo são ajustados por apresentações sucessivas dos pares (x, d) : fase de *treinamento*
- Após o treinamento, o desempenho do sistema deve ser testado com dados não-vistos: fase de teste



PARADIGMAS DE AM

- **Supervisionado**
- Por reforço
- Não-supervisionado
- Semissupervisionado

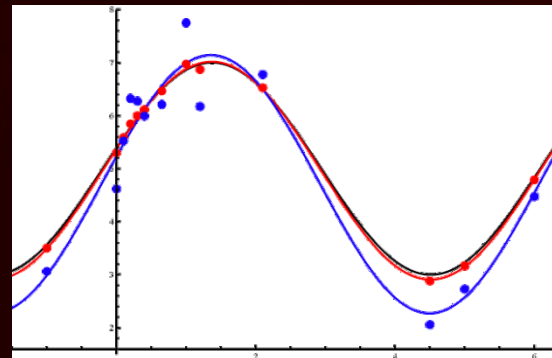


CLASSIFICAÇÃO DE PADRÕES

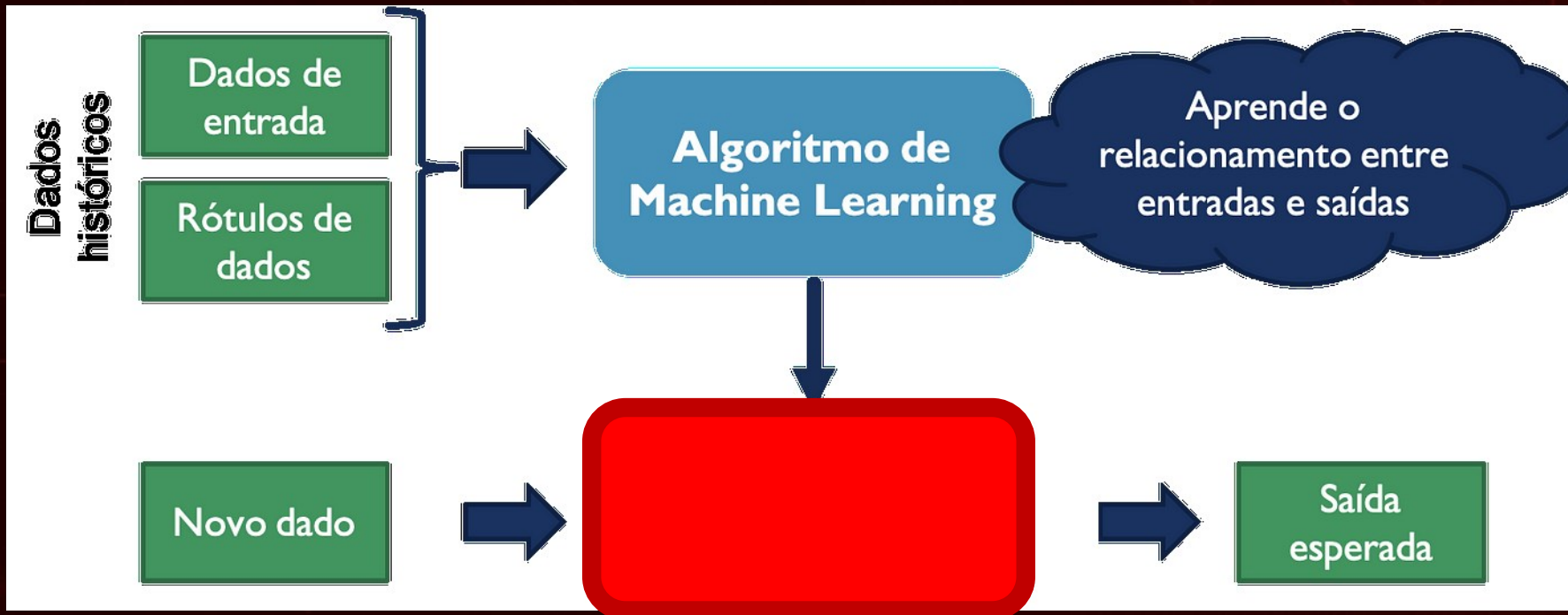
➤ Classificar objetos

REGRESSÃO

➤ Previsão de valores contínuos



CLASSIFICAÇÃO



CLASSIFICAÇÃO – REPRESENTAÇÃO

Modelos Matemáticos

- Regressão Linear/Logística, Redes Neurais Artificiais, Máquinas de Vetores de Suporte

Modelos simbólicos

- Árvores de Decisão, Regras de decisão, Redes Semânticas

Modelos “Lazy”

- K-NN, Raciocínio Baseado em Casos (CRB)

Modelos Probabilísticos

- Naïve Bayes, Redes Bayesianas, Misturas Gaussianas, Modelos de Markov

UMA REPRESENTAÇÃO PARA O CONHECIMENTO ADQUIRIDO

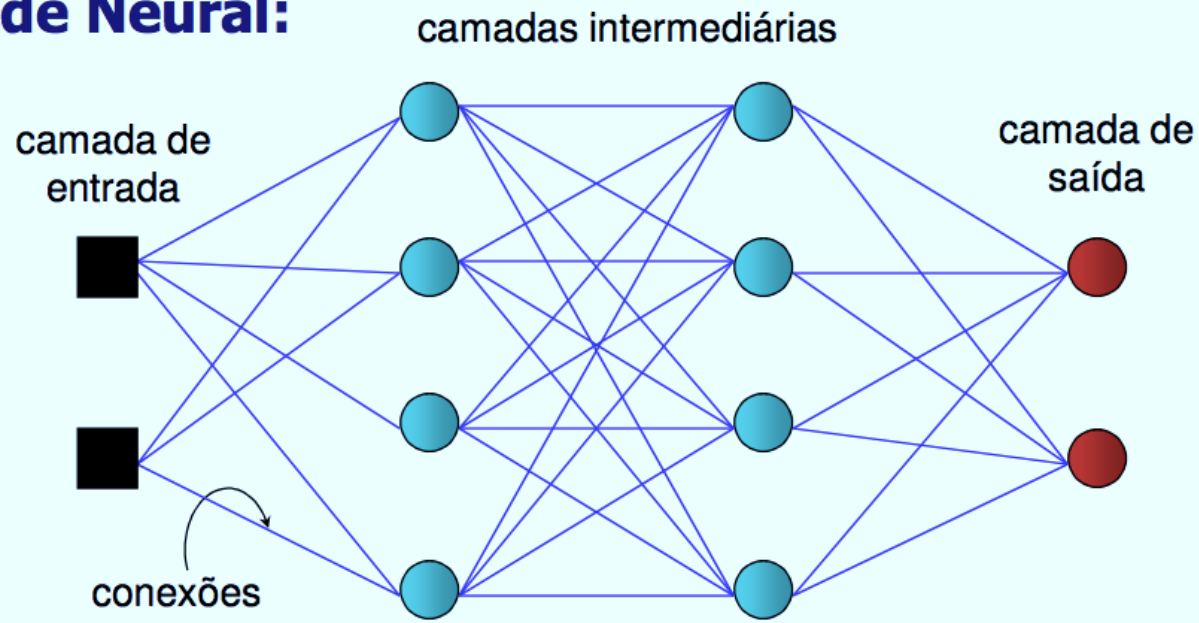
- Modelo de representação do conhecimento

CLASSIFICAÇÃO – REPRESENTAÇÃO

UMA REPRESENTAÇÃO PARA O CONHECIMENTO ADQUIRIDO

- Modelo de representação do conhecimento

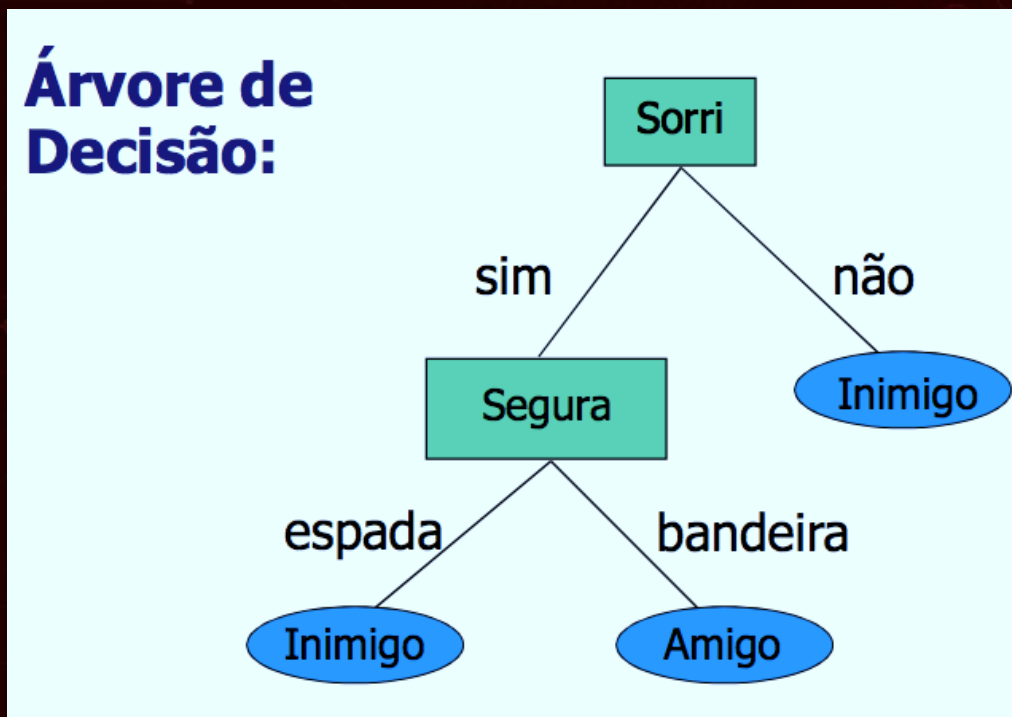
Rede Neural:



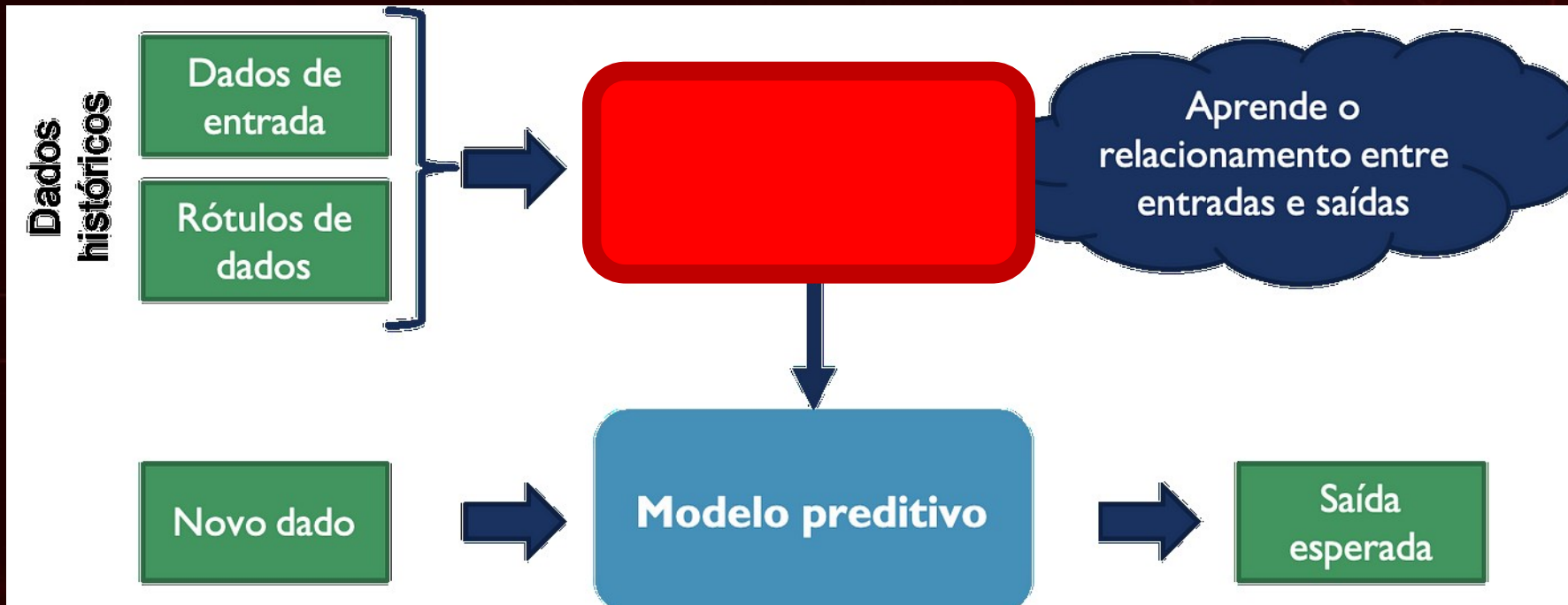
CLASSIFICAÇÃO – REPRESENTAÇÃO

UMA REPRESENTAÇÃO PARA O CONHECIMENTO ADQUIRIDO

- Modelo de representação do conhecimento



CLASSIFICAÇÃO



CLASSIFICAÇÃO – TÉCNICA DE APRENDIZADO

Algoritmos Baseados em Gradiente

- Regressão linear/logística, redes neurais...

Algoritmos baseados em Programação Dinâmica

- HMMs...

Algoritmos baseados em Divisão e Conquista

- Indução de árvores e regras de decisão

Algoritmos baseados em Probabilidades

- Naïve Bayes, Redes Bayesianas...

Algoritmos baseados em Computação Evolutiva

- Aplicável a vários modelos

Um mecanismo de aprendizado

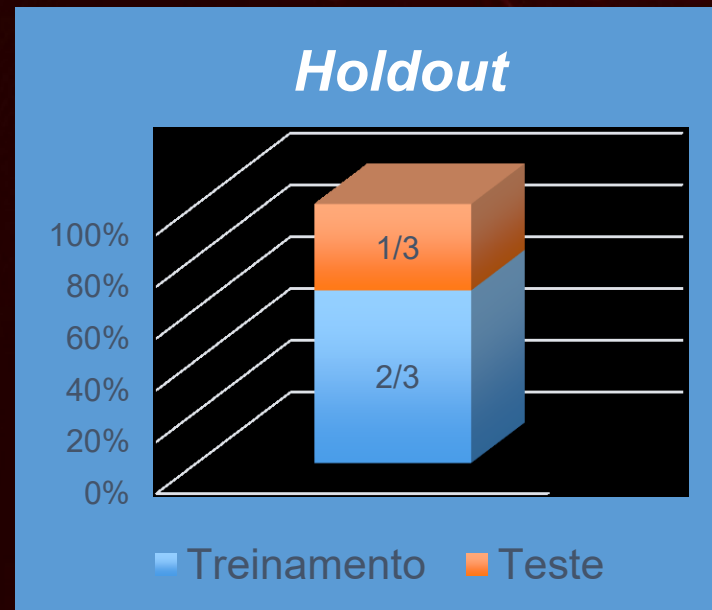
- Técnica de aprendizado

AVALIAÇÃO DE CLASSIFICADORES

- Espera-se de um classificador que ele apresente desempenho adequado para dados não vistos
 - Poder de generalização
- Para estimarmos de maneira correta o desempenho do modelo, precisamos seguir um protocolo bem definido
 - Ex.: *Holdout*, Reamostragem aleatória (*Random Subsampling*), Validação Cruzada (*Cross-Validation*), *Leave-one-out*, *Bootstrap*

HOLDOUT

- Técnica mais simples para divisão de dados
- Faz uma única partição (aleatória) da amostra em:
 - Treinamento
 - Teste
- **Atenção:** em problemas de classificação, recomenda-se que $p_{tr}(C_j) \approx p_{test}(C_j) \forall C_j \in Y$ (*holdout* estratificado)



MATRIZ DE CONFUSÃO

- Também chamada de Tabela de Contingência
 - Permite a extração de **diversas medidas** de desempenho preditivo
 - Usada para distinguir os tipos de erros
 - Usada para problemas binários ou multiclasse

Classe Prevista	Classe Verdadeira		
	A	B	C
A	25	40	20
B	25	10	5
C	20	5	10

Acurácia: $\frac{25 + 10 + 10}{25 + 40 + 20 + 25 + 10 + 5 + 20 + 5 + 10} = \frac{45}{100} = 0.45$ ou 45%

Valores fora da diagonal principal = erros

MATRIZ DE CONFUSÃO BINÁRIA

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva		
Negativa		

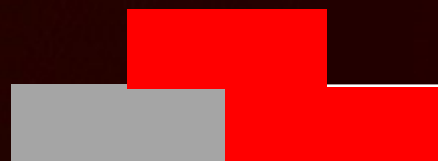


MATRIZ DE CONFUSÃO BINÁRIA

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva		
Negativa		

Erro:



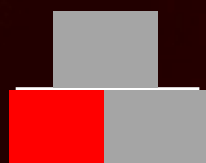
$$= (1 - \textit{acurácia})$$

MATRIZ DE CONFUSÃO BINÁRIA

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva		FP
Negativa		VN

Sensibilidade:
(TVP)
(Recall, Revocação, Benefício)




$$= (1 - TFN)$$

MATRIZ DE CONFUSÃO BINÁRIA

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

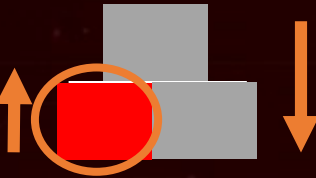
Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva		
Negativa	FN	VN

Precisão:
(Precision)



PRECISION X RECALL

Precisão:
(Precision)



Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva		
Negativa	FN	VN

O que acontece se um modelo classificar todos exemplos como sendo positivos?

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva		FP
Negativa		VN

Revocação:
(Recall)



MEDIDA F

MÉDIA HARMÔNICA DE *PRECISION* E *RECALL*

➤ Também conhecida como F_1 score ou F-score

$$F_1 = 2 \times \frac{\textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} = \frac{2}{\frac{1}{\textit{precision}} + \frac{1}{\textit{recall}}}$$

O QUE VIMOS?

- **Aprendizado**
- **Paradigmas de AM**
- **Classificação**
- **Avaliação de Classificadores**



PRÓXIMA VIDEOAULA

➤ **Aprendizado Bayesiano**

