Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO

Especialização em Desenvolvimento de Aplicações Web e Móveis Escaláveis

Turma 2019-2020

Big Data com Python

Profa. Dra. Jaqueline Brigladori Pugliesi

jbpugliesi@gmail.com

Tópicos

- Inteligência Artificial
- Data Mining / Mineração de Dados
- Aprendizado de Máquina
- Redes Neurais

24/10/2020

Inteligência Artificial

O que é inteligência?

O que é inteligência?

"Nossas mentes contêm processos que nos capacitam a solucionar problemas que consideramos difíceis. Inteligência é o nome que damos a qualquer um destes processos que ainda não compreendemos" – Marvin Minsky

"De aorcdo com uma pqsieusa de uma uinrvesriddae ignlsea, não ipomtra em qaul odrem as lrteas de uma plravaa etãso, a úncia csioa iprotmatne é que a piremria e útmlia lrteas etejasm no lgaur crteo. O rseto pdoe ser uma ttaol bçguana que vcoê pdoe anida ler sem pobrlmea. Itso é poqrue nós não lmeos cdaa lrtea isladoa, mas a plravaa cmoo um tdoo"

Snietlao que a ntsialarepuocdide é um fônnmeeo itaesstnerne e iuensátoivnqel, cumonteme euddsato no detmnarteapo de nsoglooifirieua, odne iamcsestntnenee pssuqoaiems fmoôneens ismpncieovírenes, mas de eraándxtrioiro iamctpo madniul.

Saliento que a neuroplasticidade é um fenômeno interessante e inquestionável, comumente estudado no departamento de neurofisiologia, onde incessantemente pesquisamos fenômenos incompreensíveis, mas de extraordinário impacto mundial.

35T3 P3QU3N0 T3XTO 53RV3 4P3N45 P4R4 M05TR4R COMO NO554 C4B3Ç4 CONS3GU3 F4Z3R CO1545 1MPR3551ON4ANT35! R3P4R3 N155O!

NO COM3ÇO 35T4V4 M310 COMPL1C4DO, M45 N3ST4 L1NH4 SU4 M3NT3 V41 D3C1FR4NDO O CÓD1GO QU453 4UTOM4T1C4M3NT3, S3M PR3C1S4R P3N54R MU1TO, C3RTO?

POD3 F1C4R B3M ORGULHO5O D155O!

SU4 C4P4C1D4D3 M3R3C3!

P4R4BÉN5!

Teste

Quantas letras "F" tem o seguinte texto?

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS

O que é Inteligência Artificial?

O que é Inteligência Artificial?

- “Uma área de pesquisa que investiga formas de habilitar o computador a realizar tarefas nas quais, até o momento, o ser humano tem um melhor desempenho”. (Elaine Rich)
- “Conjunto de Técnicas para a construção de máquinas inteligentes capazes de resolver problemas complexos.” (Nilson)
- “Tecnologia de Processamento de Informação que envolve processos de raciocínio, aprendizado e percepção.” (Winston)
- “Ramo da Ciência da Computação dedicado à automação do comportamento inteligente” (Luger e Stubble)

Abordagens de IA

Sistemas que pensam como humanos	**Sistemas que pensam racionalmente**
Sistemas que atuam como humanos	**Sistemas que atuam racionalmente**

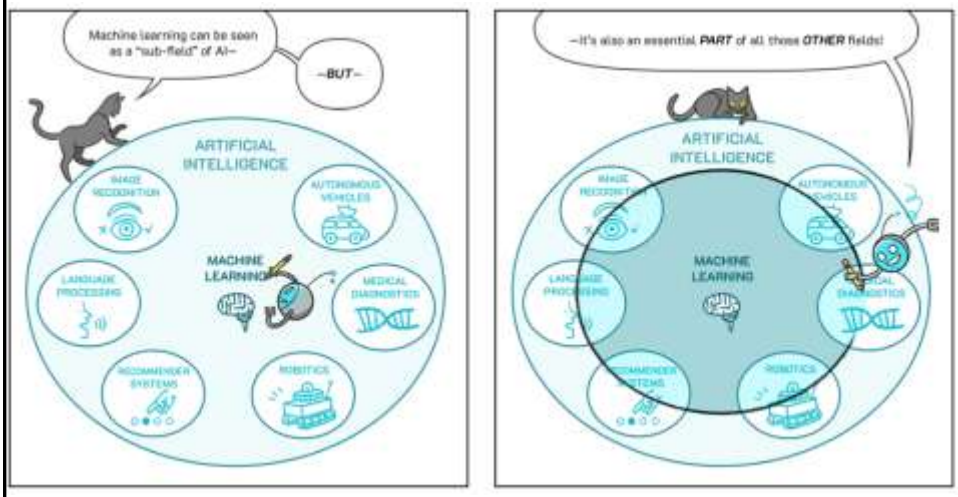
Estrutura

Inteligência Artificial

- Aprender a partir de experiências
- Aprender a partir de dados

Visão

- Descoberta de Conhecimento em Bases de Dados
- Aprendizado de Máquina
- Mineração de Dados
- Redes Neurais
- Deep Learning
- Ciência de Dados



Fonte: https://cloud.google.com/products/ai/ml-comic-1

Data Mining

Definição de Data Mining

- Data Mining (DM) refere-se ao processo de extrair conhecimento de bases de dados, ou seja, trabalhar com grandes quantidades de dados com o objetivo de extrair significado e descobrir novos conhecimentos.

Data Mining

- Processo de extração de conhecimento de Bases de Dados.
- Definição formal (Fayyad,96)
  - Processo não trivial de identificação de padrões:
    - válidos;
    - novos;
    - potencialmente úteis;
    - compreensíveis.
- Área multidisciplinar.

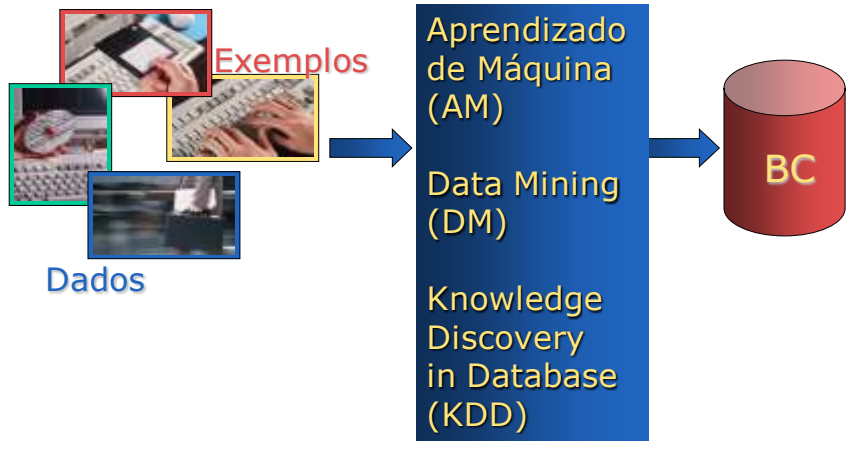
Alguns Casos de Sucesso

NIKE

WAL MART

Processo de Data Mining

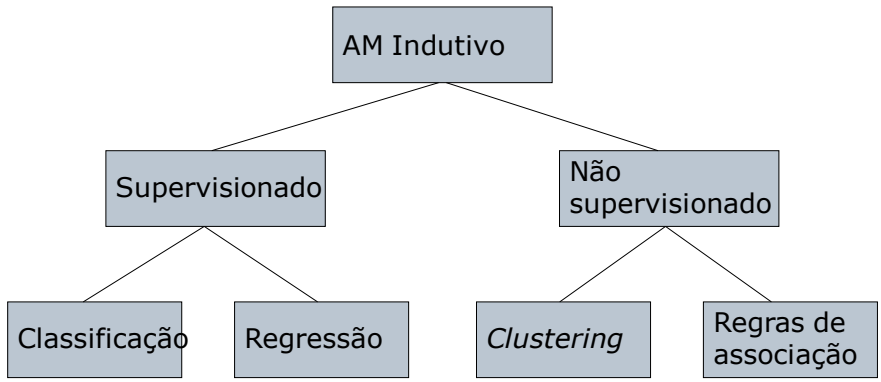
Aquisição de Conhecimento de Dados



Tarefas de Data Mining

- Preditivas
  - Classificação
  - Regressão
- Descritivas
  - Regras de associação
  - Sumarização
  - Clustering
  - etc.

Hierarquia



Aprendizado de Máquina

Aprendizado de Máquina

- Pode ser utilizado como meio para vencer um dos maiores problemas de Sistemas de IA - o gargalo da aquisição de conhecimento.
- Sub-área da Inteligência Artificial que pesquisa métodos computacionais relacionados à aquisição de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente.

Sistemas de Aprendizado de Máquina

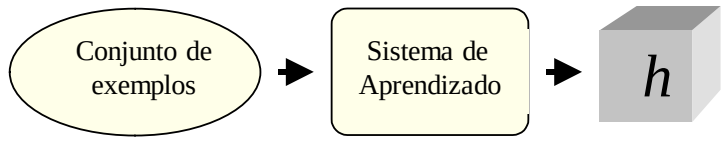
Modo de Aprendizado	Paradigmas de Aprendizado	Linguagens de Descrição	Formas de Aprendizado
• Supervisionado • Não Supervisionado	• Simbólico • Estatístico • Instance-Based • Conexionista • Genético	• Instâncias ou Exemplos • Conceitos Aprendidos ou Hipóteses • Teoria de Domínio ou Conhecimento de Fundo	• Incremental • Não Incremental

Indução

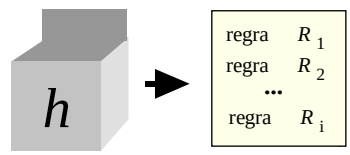
- Abordagem mais comum de AM
- Inferência lógica que permite obter conclusões com base em exemplos fornecidos → exemplos específicos são generalizados
- Possibilita:
  - Derivar conhecimento novo
  - Predizer eventos futuros
- Cuidado: nem sempre o conhecimento adquirido é verdadeiro

Conhecimento Adquirido (Hipótese h)

- *h* vista como classificador

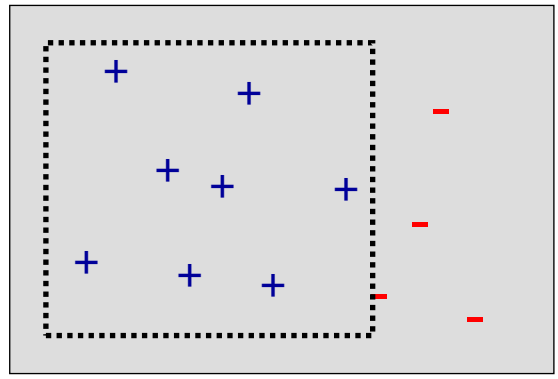


- *h* vista como conjunto de regras



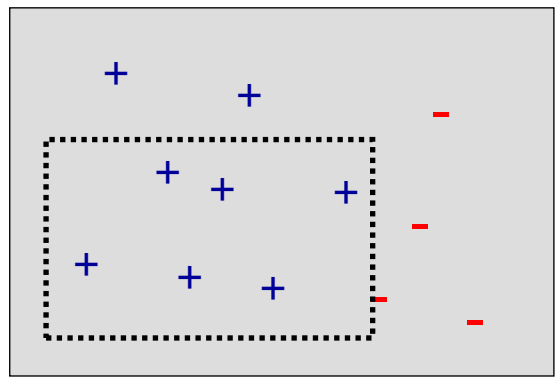
Consistência e Completude

h consistente e completa



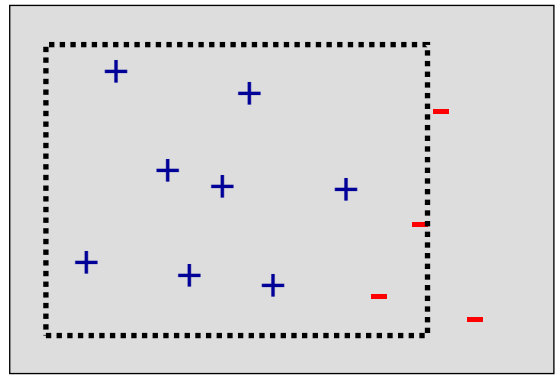
Consistência e Completude

h consistente e incompleta



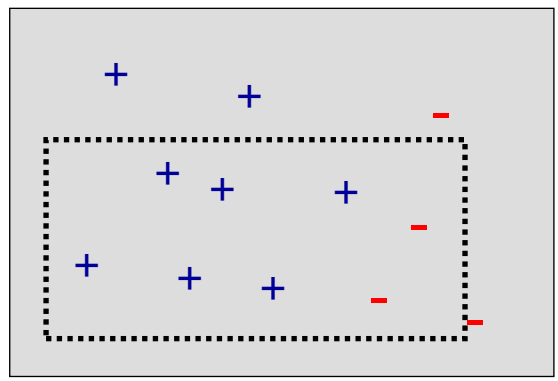
Consistência e Completude

h inconsistente e completa

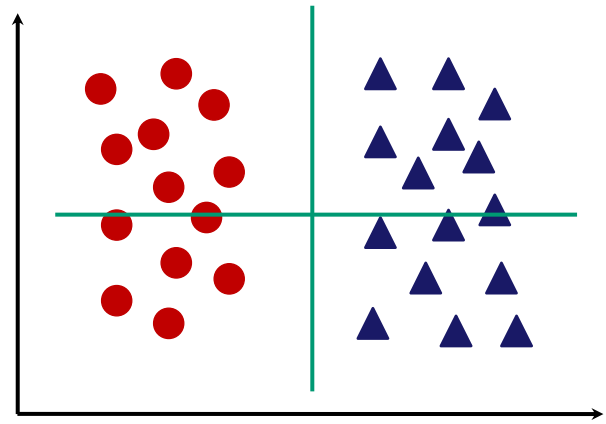


Consistência e Completude

h inconsistente e incompleta



Dados



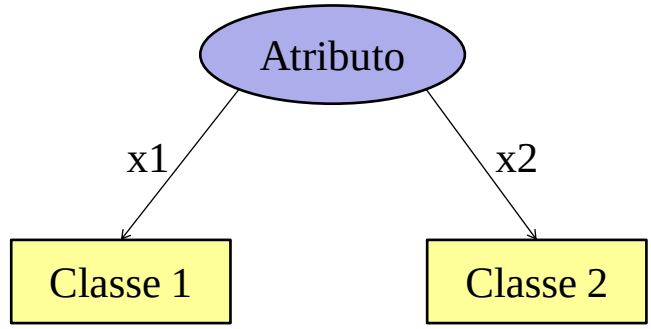
Algoritmos de AM

- Regressão linear
- SVM (Support Vector Machine)
- KNN (K-vizinhos mais próximos)
- Regressão Logística
- Árvore de decisão
- K-Means
- Floresta aleatória
- Baías ingénuas
- Algoritmos de redução dimensional
- Algoritmos de aumento de gradiente

Árvore de Decisão

Árvore de decisão é uma forma simples de representação que classifica exemplos de uma base de dados em um número finito de classes. (QUINLAN, 1993)

Árvore de Decisão



Árvores de Decisão

Entropia

Medida que expressa a impureza ("bagunça") do conjunto de dados, nesse caso, do conjunto de treinamento

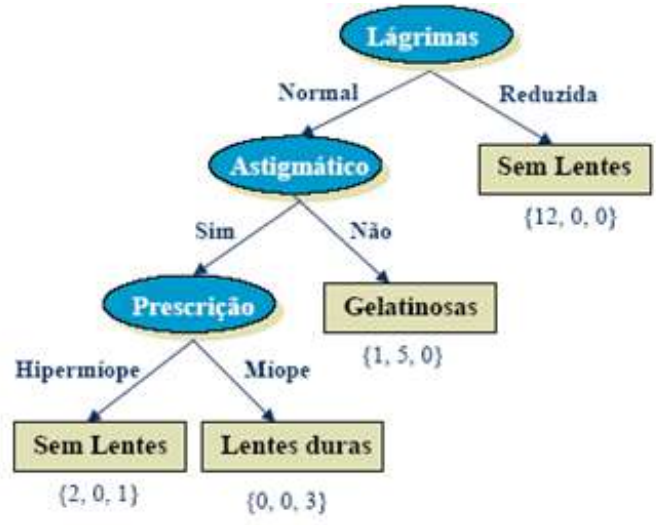
$$Entropia(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Ganho de Informação

Redução esperada na entropia causada pela partição do conjunto de treinamento de acordo com um determinado atributo

$$\text{Ganho de Informação}(S, A) = \text{Entropia}(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} \text{Entropia}(S_v)$$

Árvore de Decisão



Idade	Prescrição	Astigmático	Lágrimas	Lentes
Jovem	Miope	Nao	Normal	Lentes_gelatinosas
Jovem	Miope	Nao	Reduzida	Sem_lentes
Jovem	Miope	Sim	Normal	Lentes_duras
Jovem	Miope	Sim	Reduzida	Sem_lentes
Jovem	Hipermiope	Nao	Normal	Lentes_gelatinosas
Jovem	Hipermiope	Nao	Reduzida	Sem_lentes
Jovem	Hipermiope	Sim	Normal	Lentes_duras
Jovem	Hipermiope	Sim	Reduzida	Sem_lentes
Media	Miope	Nao	Normal	Lentes_gelatinosas
Media	Miope	Nao	Reduzida	Sem_lentes
Media	Miope	Sim	Normal	Lentes_duras
Media	Miope	Sim	Reduzida	Sem_lentes
Media	Hipermiope	Nao	Normal	Lentes_gelatinosas
Media	Hipermiope	Nao	Reduzida	Sem_lentes
Media	Hipermiope	Sim	Normal	Sem_lentes
Media	Hipermiope	Sim	Reduzida	Sem_lentes
Senior	Miope	Nao	Normal	Sem_lentes
Senior	Miope	Nao	Reduzida	Sem_lentes
Senior	Miope	Sim	Normal	Lentes_duras
Senior	Miope	Sim	Reduzida	Sem_lentes
Senior	Hipermiope	Nao	Normal	Lentes_gelatinosas
Senior	Hipermiope	Nao	Reduzida	Sem_lentes
Senior	Hipermiope	Sim	Normal	Sem_lentes
Senior	Hipermiope	Sim	Reduzida	Sem_lentes

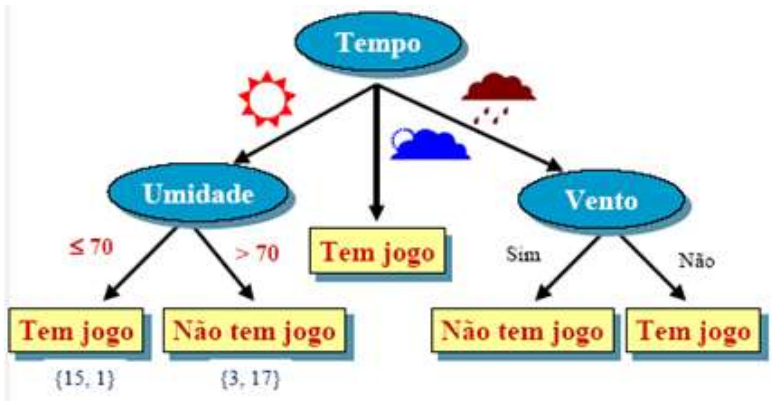
Atributos com valores contínuos

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

Atributos com valores discretos

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Árvores de Decisão



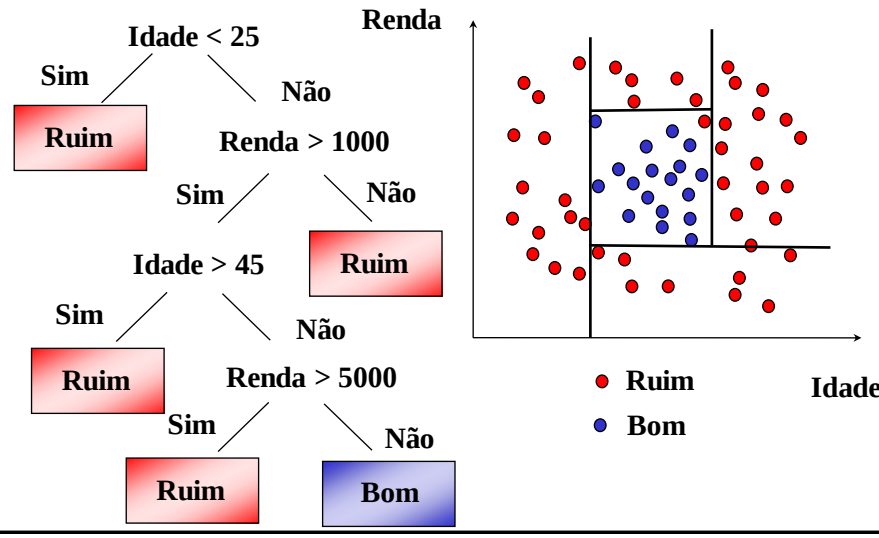
Indução a partir de Dados

- Algoritmos de AM induzem modelos de dados que podem ser usados para a predição.
- Indução de uma árvore de decisão usando um conjunto de dados artificial.
- Esse conjunto de dados possui apenas dois atributos (renda e idade) e duas classes (ruim e bom).

Dados sobre Crédito Bancário

Idade	Renda	Classe
20	2000	Ruim
30	5100	Bom
60	5000	Ruim
40	6000	Bom
...	...	...

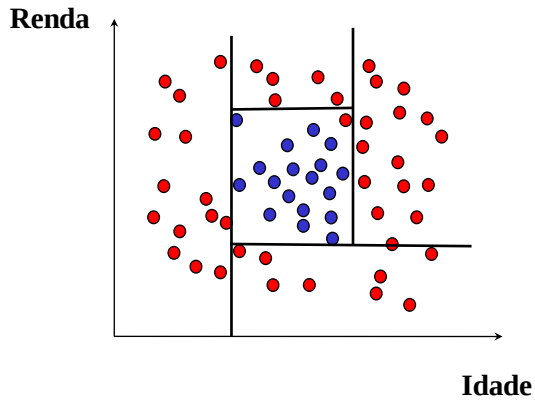
Árvore de Decisão



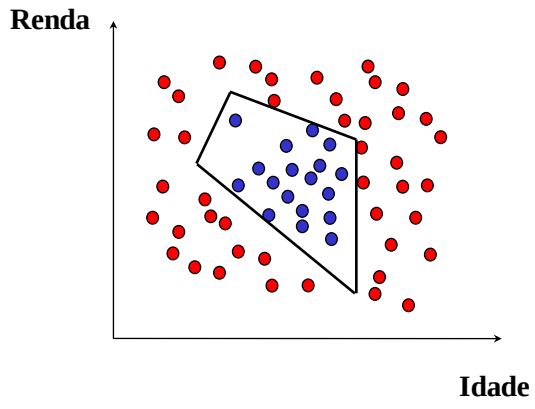
Erro

- Principais fatores:
  - Informação dos atributos
  - Adaptação do algoritmo de aprendizado aos dados
  - Distribuição dos casos futuros
  - Quantidade de dados

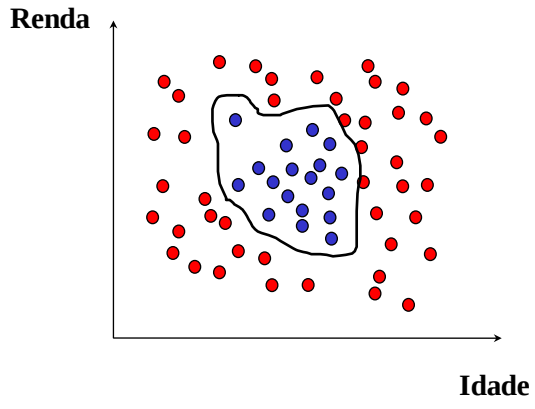
... Uma Possível H1



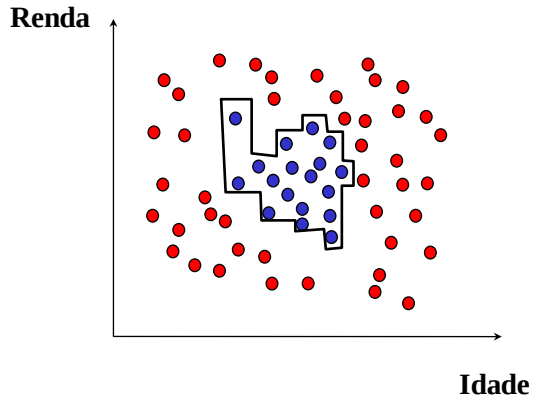
... Outra Possível H2



... Outra Possível H3



... Outra Possível H4



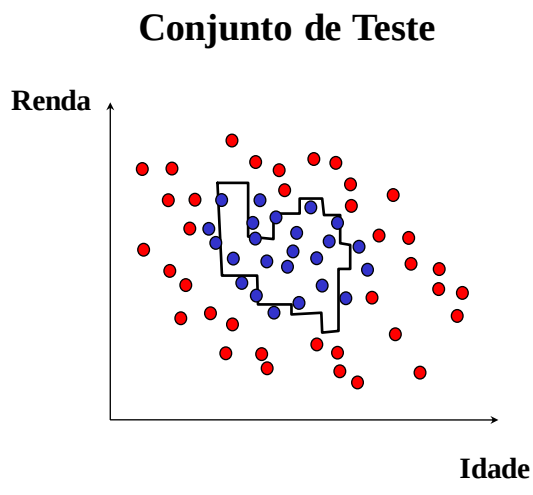
Qual a Melhor H?
Não Esquecer o Erro...

Medidas:

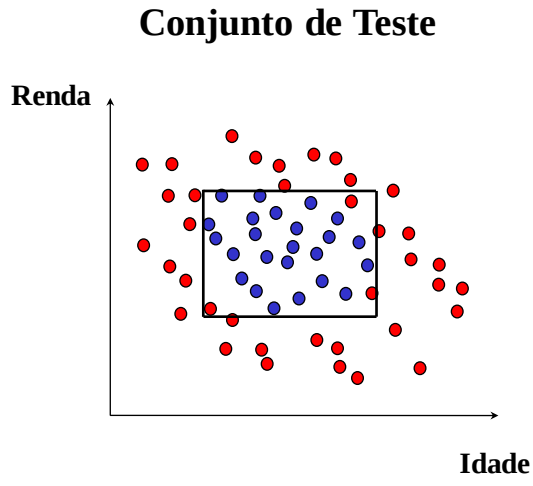
$$\text{Precisão} = \frac{\text{corretamente classificados}}{\text{total de exemplos}}$$

$$\text{Erro} = \frac{\text{incorretamente classificados}}{\text{total de exemplos}}$$

Erro de H4



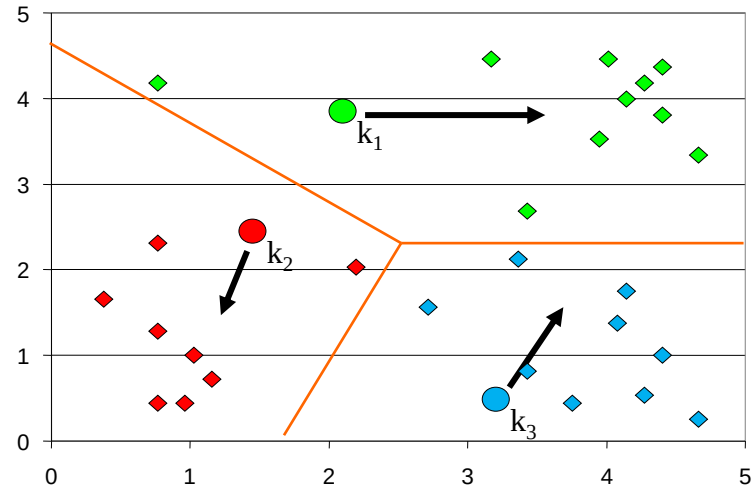
Erro de H1



Algoritmo k-NN

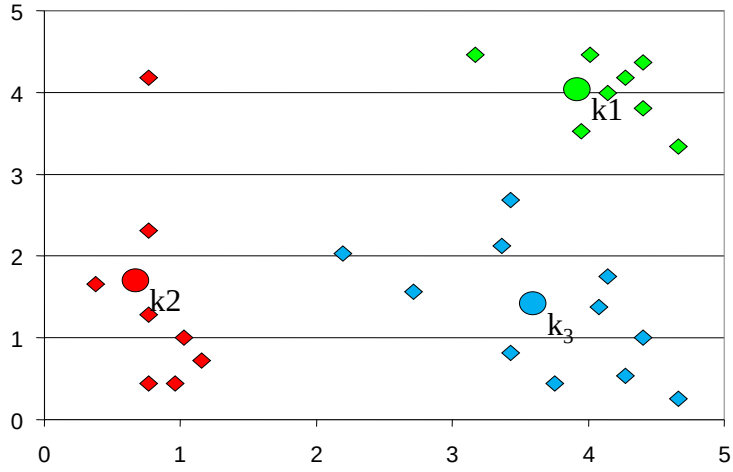
- k-Nearest Neighbor
- Baseado em distâncias, vizinhos, pesos, ...
- Ainda é um algoritmo de AM popular
- K muito grande x muito pequeno

K-means



Fonte: material de Profa. Dra. Solange Rezende

K-means



Fonte: material de Profa. Dra. Solange Rezende

Importante

- Conhecer os dados
- Escolhar do modelo
- Todo algoritmo indutivo tem um bias
- Desempenho de um algoritmo varia com o domínio
- Análise experimental é fundamental
- Overfitting
- Métodos de Amostragem

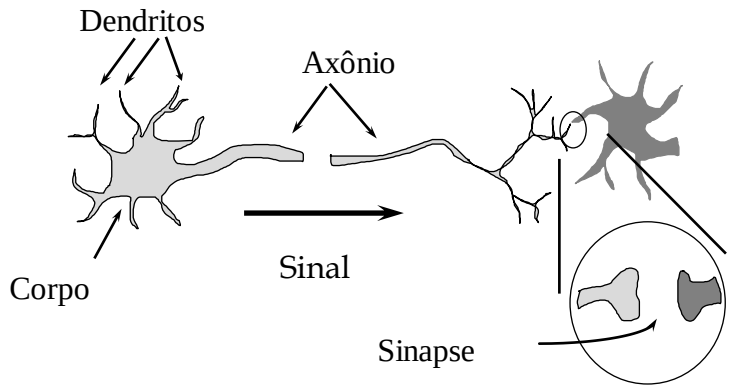
Problemas e Desafio

- Valores ausentes
- Erros
- Dados não estruturados
- Tipo de representação

Redes Neurais

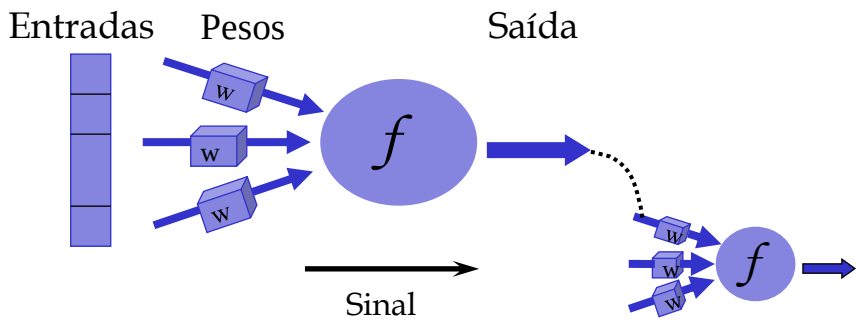
Neurônio Natural

- Um neurônio simplificado:



Neurônio artificial

- Modelo de um neurônio abstrato:



Áreas de aplicação

- Reconhecimento de padrões
- Reconhecimento de voz e imagem
- Análise estatística
- Agrupamento
- Otimização
- Memória associativa
- Controle (robôs e processos químicos)
- Diagnóstico médico
- Finanças
- Fusão de sensores
- Bioinformática

RNAs no setor de serviços

- Tem havido um grande número de pesquisas sobre a utilização de RNAs para previsão financeira
  - Tendências sugerem expansão destas aplicações
  - Ex. Cartão de crédito VISA utiliza RNAs para liberação de cartões e detecção de fraudes

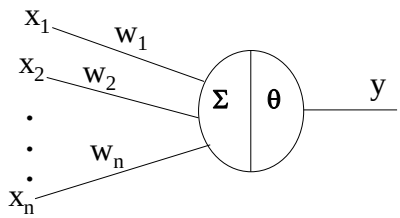
Conceitos básicos

- Principais aspectos das RNA
  - Arquitetura
    - Unidades de processamento (nós)
    - Conexões
    - Topologia
  - Aprendizado
    - Algoritmos
    - Paradigmas

Unidades de processamento

- Função: receber entradas de conjunto de unidades A, computar função sobre entradas e enviar resultado para conjunto de unidades B
- Entrada total

$$u = \sum_{j=1}^{N} x_j w_j$$

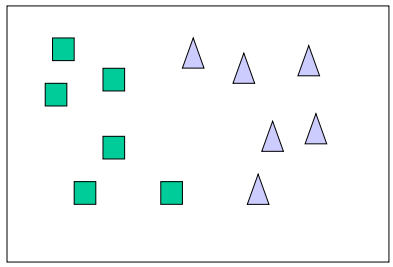


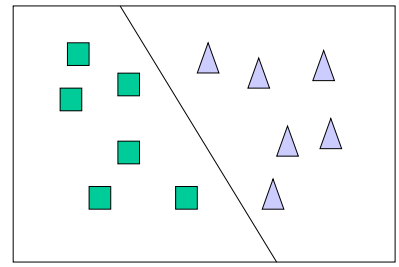
Perceptron

- Desenvolvida por Rosemblat, 1958
- Utiliza modelo de McCulloch-Pitts como neurônio
  - McCulloch-Pitts formularam matematicamente neurônios naturais
- Rede mais simples para classificação de padrões linearmente separáveis

Perceptron

Padrões linearmente separáveis





Perceptron

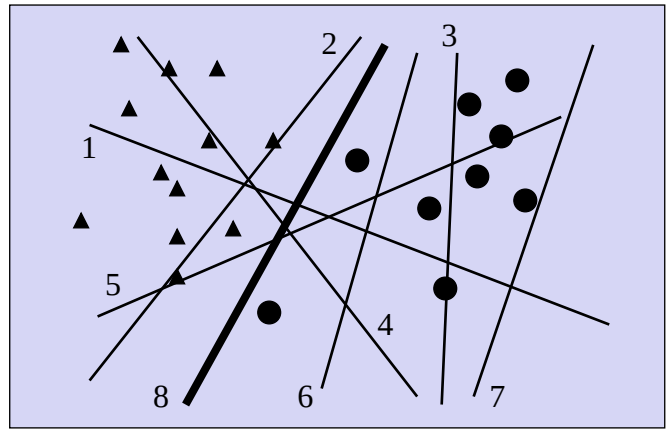
- Treinamento
  - Supervisionado
  - Correção de erro

$\Delta w_{ij} = \eta\, x_i\, (d_j - y_j)$ $\quad (d \neq y)$

$\Delta w_{ij} = 0$ $\quad (d = y)$

- Teorema de convergência: se é possível classificar um conjunto de entradas, uma rede Perceptron fará a classificação

Treinamento

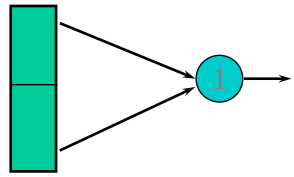


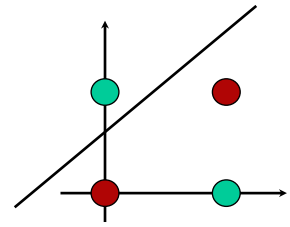
Problemas com Perceptron

- Redes de uma camada resolvem apenas problemas linearmente separáveis
- Grande número de aplicações importantes são não linearmente separáveis
  - Exemplo: paridade

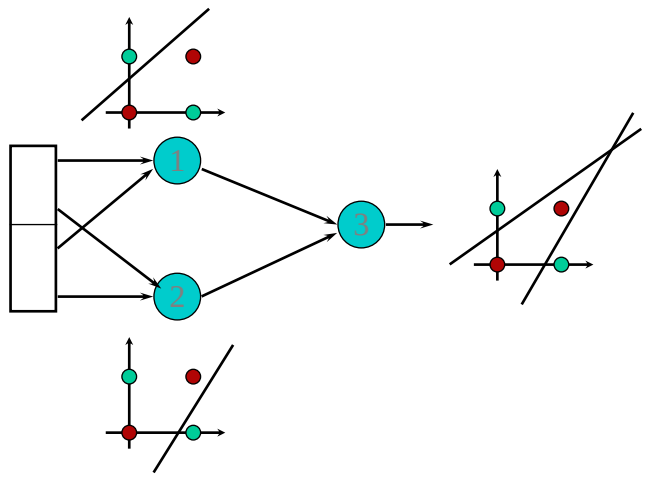
Problemas com Perceptron

$0, 0 \rightarrow 0$
$0, 1 \rightarrow 1$
$1, 0 \rightarrow 1$
$1, 1 \rightarrow 0$





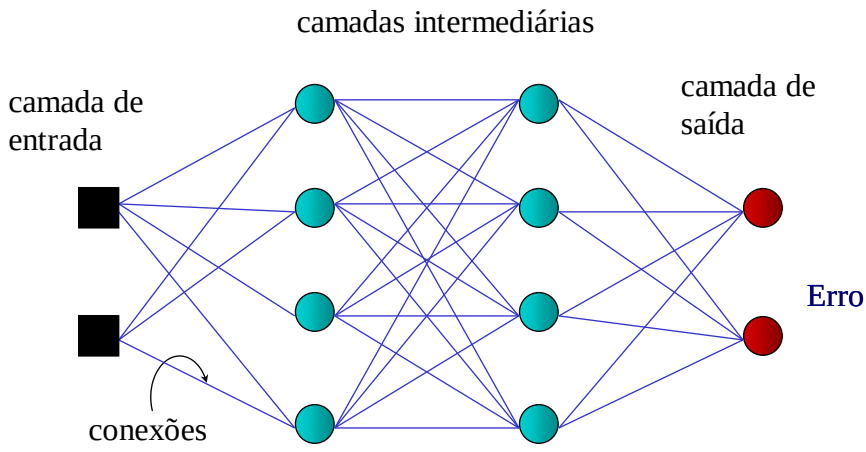
Problemas com Perceptron



Rede Multi-Layer Perceptron (MLP)

- Arquitetura de RNA mais utilizada
  - Possui uma ou mais camadas intermediárias de nós
- Grande Funcionalidade
  - Uma camada intermediária: qualquer função contínua ou Booleana
  - Duas camadas intermediárias: qualquer função
- Treinada com o algoritmo Backpropagation

Rede Multi-Layer Perceptron (MLP)



Exemplo

Seja o seguinte cadastro de pacientes:

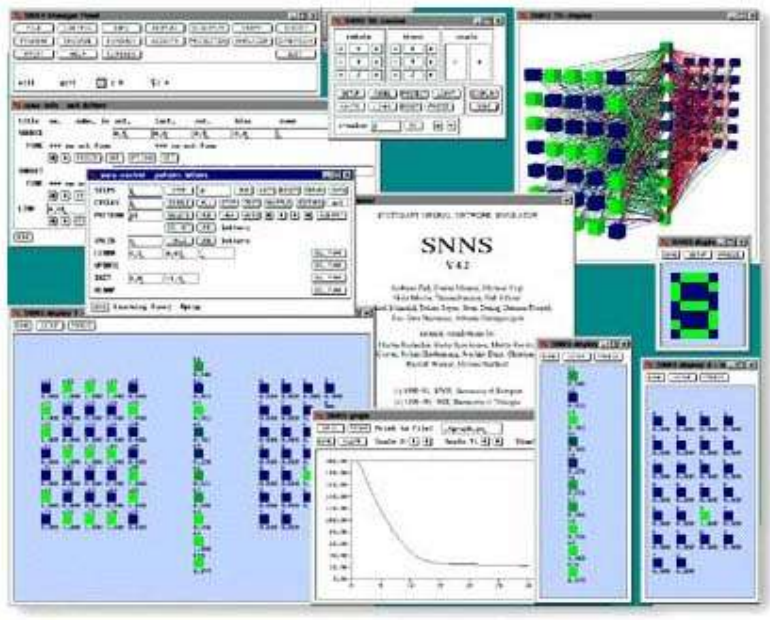
Nome	Febre	Enjôo	Manchas	Dores	Situação
João	sim	não	pequena	sim	saudável
Pedro	não	não	grande	não	doente
Maria	sim	sim	pequena	sim	saudável
José	sim	não	pequena	sim	doente
Ana	sim	não	pequena	sim	saudável
Leila	não	não	grande	sim	doente

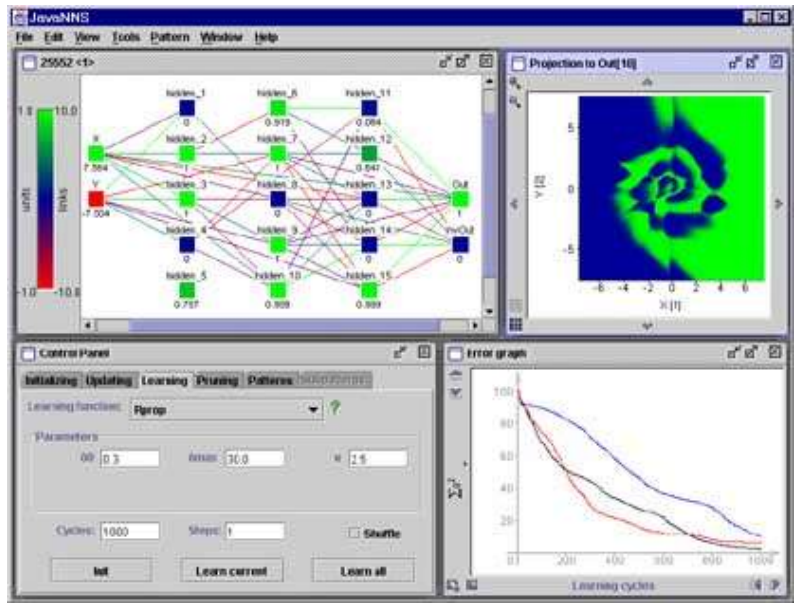
Exemplo

Nome	Febre	Enjôo	Manchas	Dores	Situação
João	1	0	1	1	+1
Pedro	0	0	0	0	-1
Maria	1	1	1	1	+1
Leila	0	0	0	1	-1

JavaNNS

- Stuttgart Neural Network Simulator (SNNS)
- Java Neural Network Simulator
  - http://www.ra.cs.uni-tuebingen.de/software/JavaNNS/





Neurônios camada oculta

- Alguns autores sugerem:
  - O número de neurônios escondidos deve estar entre o tamanho da camada de entrada e o da camada de saída.
  - O número de neurônios escondidos deve ser 2/3 do tamanho da camada de entrada, mais o tamanho da camada de saída.
  - O número de neurônios escondidos deve ser menor que o dobro do tamanho da camada de entrada.

Camadas - RN

- Exemplo:
  - Camada de entrada → 30
  - 1ª camada oculta → 18
  - 2ª camada oculta → 10
  - Camadas de saída → 2

Trabalho

Trabalho

- Selecionar do conjunto de dados
- Importação das bibliotecas
- Resumo do conjunto de dados
- Visualização de dados
- Separar um conjunto de dados de validação
- Configurar os testes para utilizar 10-fold cross validation
- Construir 5 modelos diferentes para prever as classes
- Selecionar o melhor modelo

Repositórios de dados

- UCI Machine Learning Repository
  http://archive.ics.uci.edu/ml/index.php

- Kaggle Datasets
  https://www.kaggle.com/datasets

- OpenML
  https://www.openml.org/

Entrega

- Arquivo .csv
- Arquivo do Jupyter Notebook (.ipynb)
- Arquivo com relatório da análise de dados

- Enviar por e-mail: profa.jaqueline@gmail.com

Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO

Final da Apresentação

Profa. Dra. Jaqueline Brigladori Pugliesi
jbpugliesi@gmail.com