



Data Science Academy

Machine Learning



Seja bem-vindo!





Data Science Academy

Algoritmos de Machine Learning e Modelos Preditivos



Data Science Academy

Valor = Dados + Análise



Modelos Descritivos

- Quantos clientes perdemos nos últimos 3 meses?
- As fraudes aumentaram ou diminuiram no último ano?





Modelo Preditivo é uma função matemática que, aplicada a uma massa de dados, consegue identificar padrões ocultos e prever o que poderá ocorrer



Aprendizagem
Supervisionada

Aprendizagem Não
Supervisionada

Métodos
Baseados em
Instância

Métodos
Probabilísticos

Métodos
Baseados em
Procura

Métodos
Baseados em
Otimização





A construção de bons modelos preditivos implica o domínio de um conjunto de metodologias e conceitos sem os quais a sua qualidade poderá ser afetada





DataDriven Business







Data Science Academy

Exemplo de Utilização de Modelos Preditivos



Data Science Academy



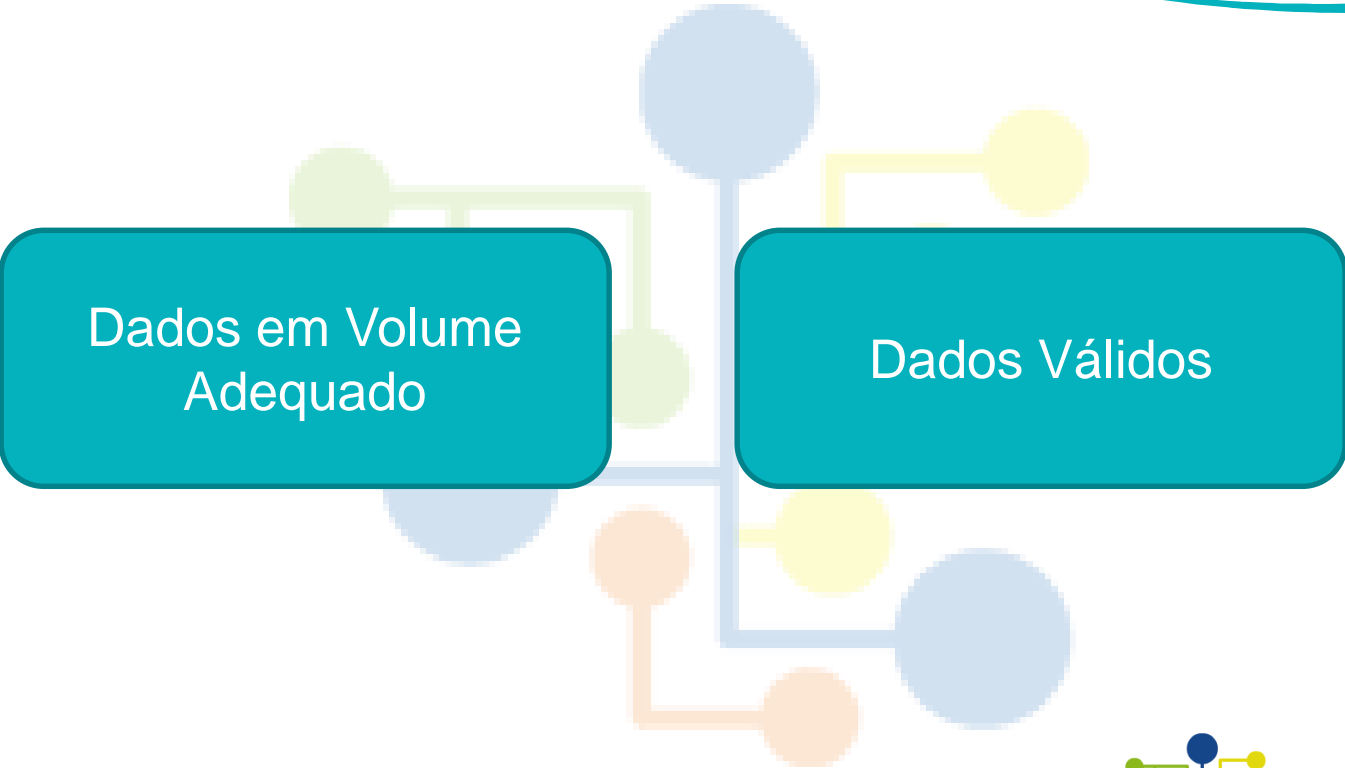
Suponhamos uma operadora de telefonia móvel. Um dos seus principais problemas de empresas deste segmento é a taxa de desconexão ou churn rate





Agregando ao modelo regras de negócio, como agrupar clientes por rentabilidade, a operadora pode fazer ofertas diferenciadas para evitar a desconexão.

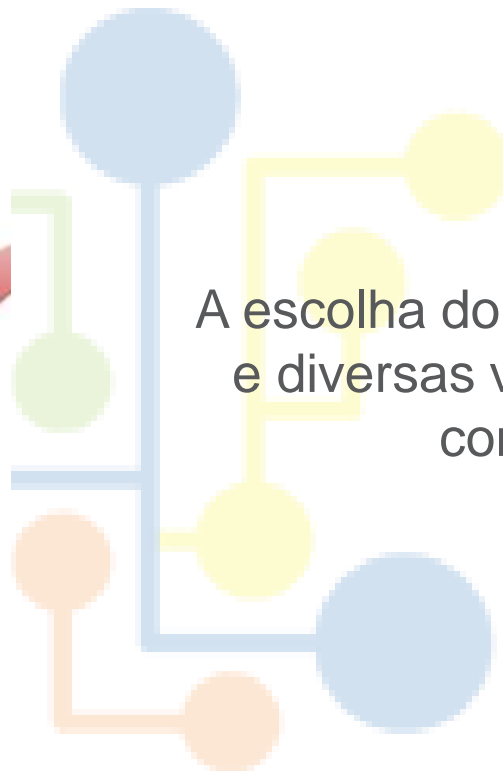




Dados em Volume
Adequado

Dados Válidos





A escolha do modelo é importante
e diversas variáveis devem ser
consideradas





Criar iniciativas de Big Data Analytics, não é simplesmente adquirir tecnologias



Identifique com a maior precisão possível o problema de negócio. Quanto mais precisa a pergunta, mais precisa será a resposta e portanto maior o valor da resposta.



Não superestime o valor da predição. Mesmo em uma sociedade cada vez mais data-driven, a intuição muitas vezes é necessária.



Tenha dados em volume e qualidade adequados. Sem qualidade, o volume não tem valor.



Não subestime o desafio da implementação. Não basta ter apenas a tecnologia, é necessário expertise (conhecimento do negócio, tecnologia, modelagem) para fazer a coisa acontecer.







Data Science Academy

O que é um Modelo Preditivo?



Data Science Academy

Modelo Preditivo é uma função matemática que, aplicada a uma massa de dados, consegue identificar padrões ocultos e prever o que poderá ocorrer





Modelo Preditivo



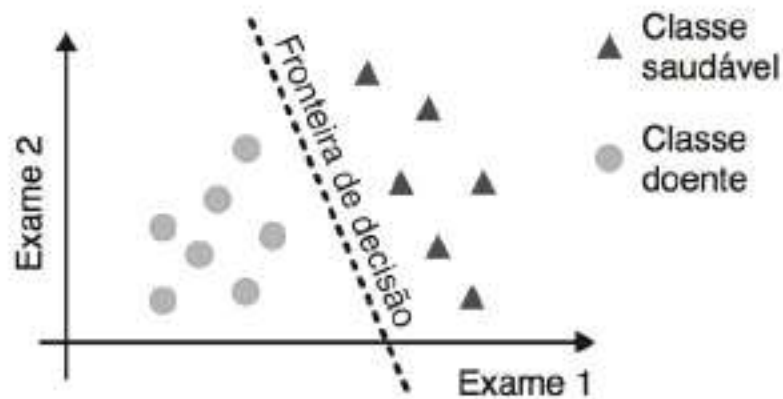
$$\mathbf{D} = \{(\mathbf{x}_i, f(\mathbf{x}_i)), i = 1, \dots, n\}$$

f = função desconhecida

\hat{f} = aproximação da função desconhecida



Classificação



Mas o que é um processo estocástico?



Classificação

Variáveis Preditoras

Espécie	Tamanho (Petal)	Largura (Petal)	Tamanho (Sepal)	Largura (Sepal)
Setosa	5.1	3.5	1.4	0.2
Setosa	4.9	3.0	1.4	0.2
Versicolor	7.0	3.2	4.7	1.4
Versicolor	6.4	3.2	4.5	1.5
Virgínica	6.3	3.3	6.0	2.5
Virgínica	5.8	2.7	5.1	1.9

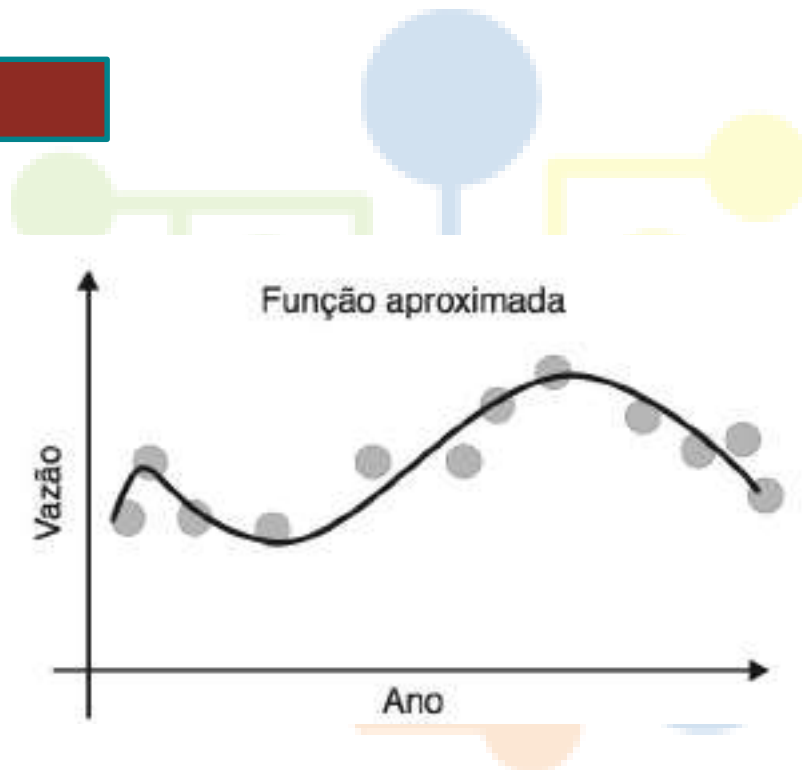
Classe



O objetivo do aprendizado de máquina é aprender a aproximação da função f que melhor representa a relação entre os atributos de entrada (chamadas variáveis preditoras) com a variável target



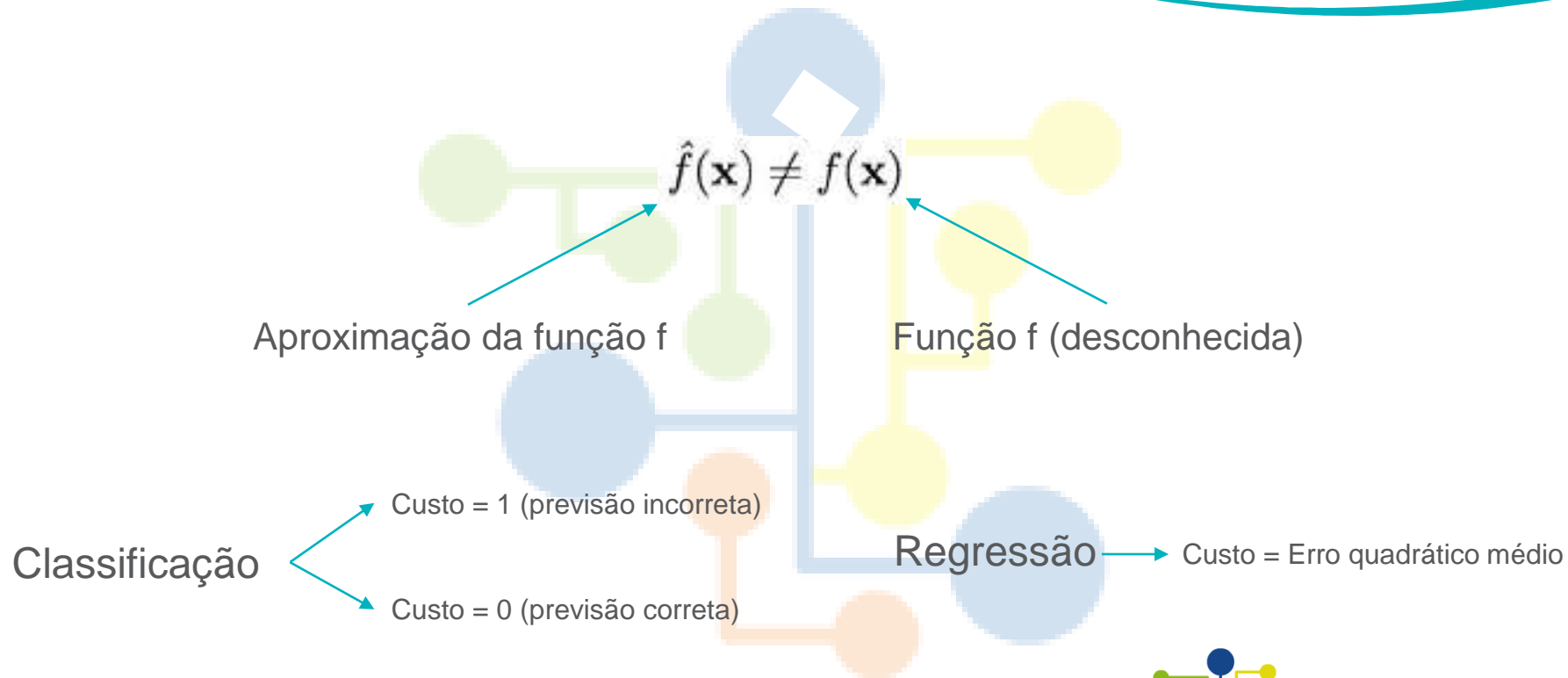
Regressão



Regressão

Mortalidade	Fertilidade	Agricultura	Educação	Renda
21.2	70.2	17.2	13	9.2
24.2	86.1	44.3	9	81.3
23.2	91.5	34.9	4	98.7
21.3	87.2	34.2	8	36.1
23.6	78.9	42.9	14	6.8





Modelos Generativos

- Naive Bayes

Modelos Discriminativos

- Árvores de Decisão
- Redes Neurais
- KNN



Métodos Baseados em Instância

Métodos Probabilísticos

Métodos Baseados em Procura

Métodos Baseados em Otimização



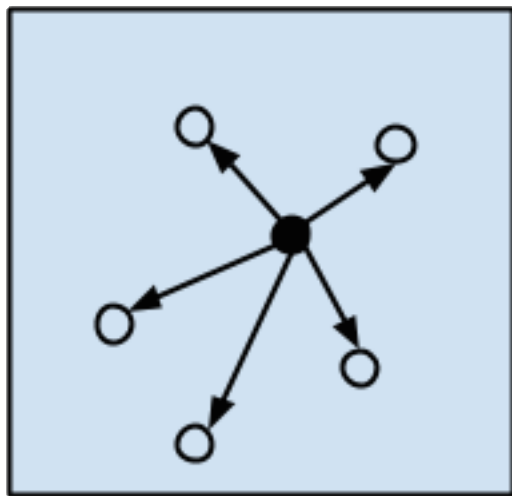


Data Science Academy

Métodos Baseados em Instâncias (Instance Based)

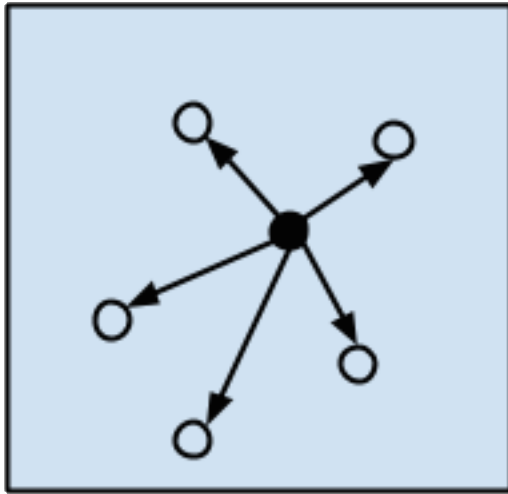


Data Science Academy



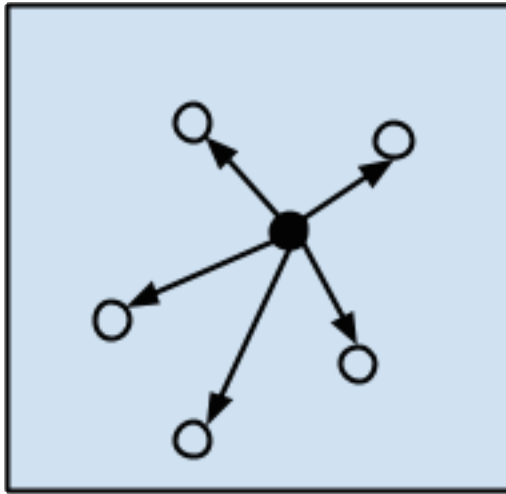
Métodos de Aprendizagem Baseado em Instâncias





A aprendizagem consiste
somente em armazenar os
exemplos de treinamento





O conceito base por trás deste método é que os dados tendem a estar concentrados em uma mesma região no espaço de entrada



$$d(x_i, x_j) = \sqrt{\sum_{p=1}^n (a_p(x_i) - a_p(x_j))^2}$$

Onde $a_p(x_i)$ é o valor do p-ésimo atributo da instância x



Métodos de aprendizagem baseados em instâncias assumem que as instâncias podem ser representadas como pontos em um espaço Euclidiano



Outras Medidas de Distância

- Correlação de Pearson – Coeficiente de correlação usado em estatística. Muito usado em bioinformática.
- Similaridade de Cosseno – Cosseno do ângulo entre os vetores. Usado para classificação de textos e outros dados de alta dimensão.
- Distância de edição – Usado para medir distância entre strings. Usado em classificação de textos e bioinformática.



Dataset de Treino

Dataset de Teste

(x_1, c_1)

(x_2, c_2)

(x_3, c_3)

.

.

.

(x_n, c_n)

$(x_z, ?)$

Variável
Preditora

Variável
Target



Dataset de Treino

Dataset de Teste

(x_1, c_1)

(x_2, c_2)

(x_3, c_3)

.

.

.

(x_n, c_n)

d_1

d_2

d_3

d_n

$(x_z, ?)$

Variável
Preditora

Variável
Target

Qual a similaridade entre x_1 e x_z ?
Distância Euclidiana



Dataset de Treino

Dataset de Teste

(x_1, c_1)

$(x_2, \mathbf{c_2})$

(x_3, c_3)

.

.

.

(x_n, c_n)

d_1

d_2

d_3

d_n

$(x_z, \mathbf{c_2})$

Variável
Preditora

Variável
Target

Qual a similaridade entre x_1 e x_z ?
Distância Euclidiana



Métodos Baseados em Instância

Constroem aproximações da função alvo para cada instância de teste diferente.



Métodos Baseados em Instância

Podem utilizar representações mais complexas e simbólicas para as instâncias



Métodos Baseados em Instância

Os métodos de aprendizagem baseados em instâncias são métodos não paramétricos.



Métodos Baseados em Instância

Uma desvantagem é o alto custo para classificação. Toda computação ocorre no momento da classificação.



Outras Características:

- Ao contrário das outras abordagens, não ocorre a construção de um modelo de classificação explícito.
- Novos exemplos são classificados com base na comparação direta e similaridade aos exemplos de treinamento.
- Treinamento pode ser fácil, apenas memoriza exemplos.
- Teste pode ser caro pois requer comparação com todos os exemplos de treinamento.
- Métodos baseados em instância favorecem a similaridade global e não a simplicidade do conceito.





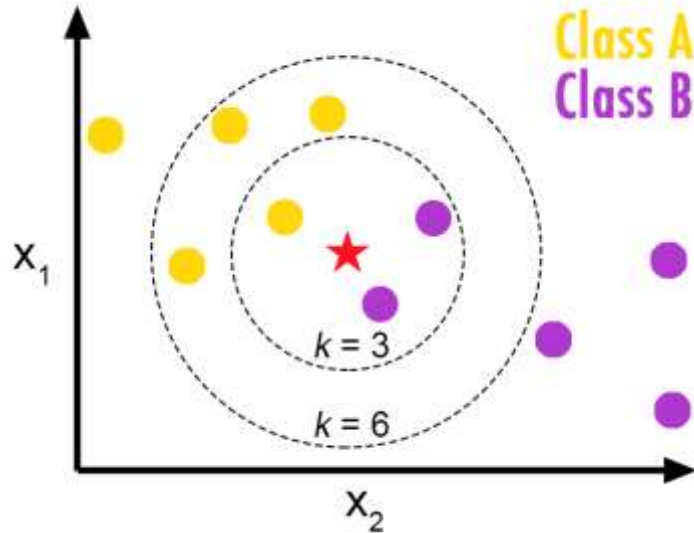
Objetos relacionados ao mesmo conceito são semelhantes entre si



Regra do KNN

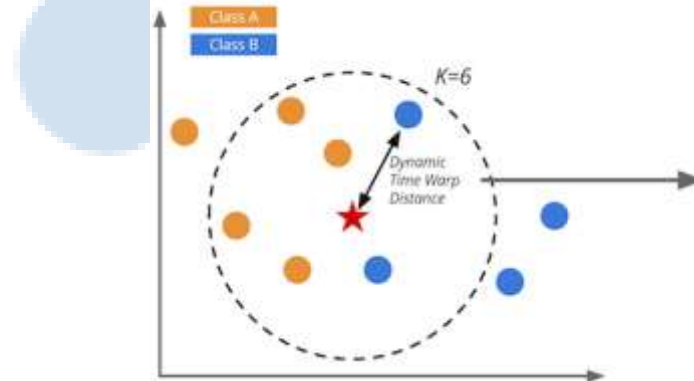
Classificar xz atribuindo a ele o rótulo representado mais frequentemente dentre as k amostras mais próximas e utilizando um esquema de votação.

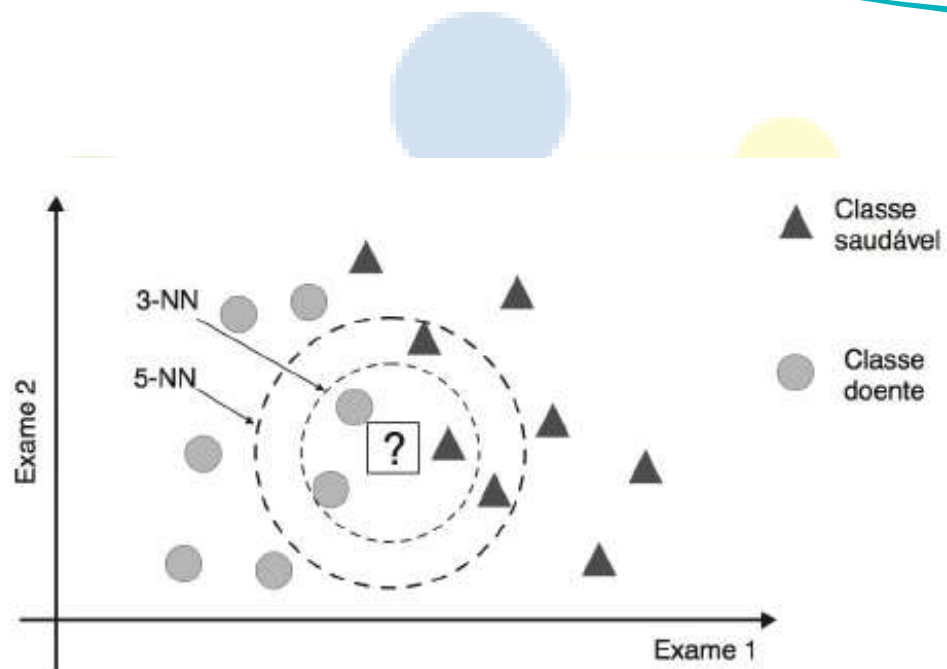




Regra de Classificação

Função que Calcula a Distância
entre as duas Instâncias







Data Science Academy

Métodos Probabilísticos



Data Science Academy

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Métodos Probabilísticos



Os métodos probabilísticos bayesianos assumem que a probabilidade de um evento A, que pode ser uma classe, dado em um evento B, poder o conjunto de valores dos atributos de entrada, não depender apenas da relação entre A e B, mas também da probabilidade de observar A independentemente de B



$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Métodos Probabilísticos



Os Métodos Probabilísticos são relevantes por dois motivos:

1. Fornecem algoritmos de aprendizagem práticos:

- Aprendizagem Naïve Bayes
- Aprendizagem de Redes Bayesianas
- Combinam conhecimento a priori com os dados observados



Os Métodos Probabilísticos são relevantes por dois motivos:

2. Fornecem uma estrutura conceitual útil:

Fornece a “norma de ouro” (regra do menor erro possível) para avaliar outros algoritmos de aprendizagem.



Cada exemplo de treinamento pode decrementar ou incrementar a probabilidade de uma hipótese ser correta



Conhecimento a priori pode ser combinado com os dados observados para determinar a probabilidade de uma hipótese.

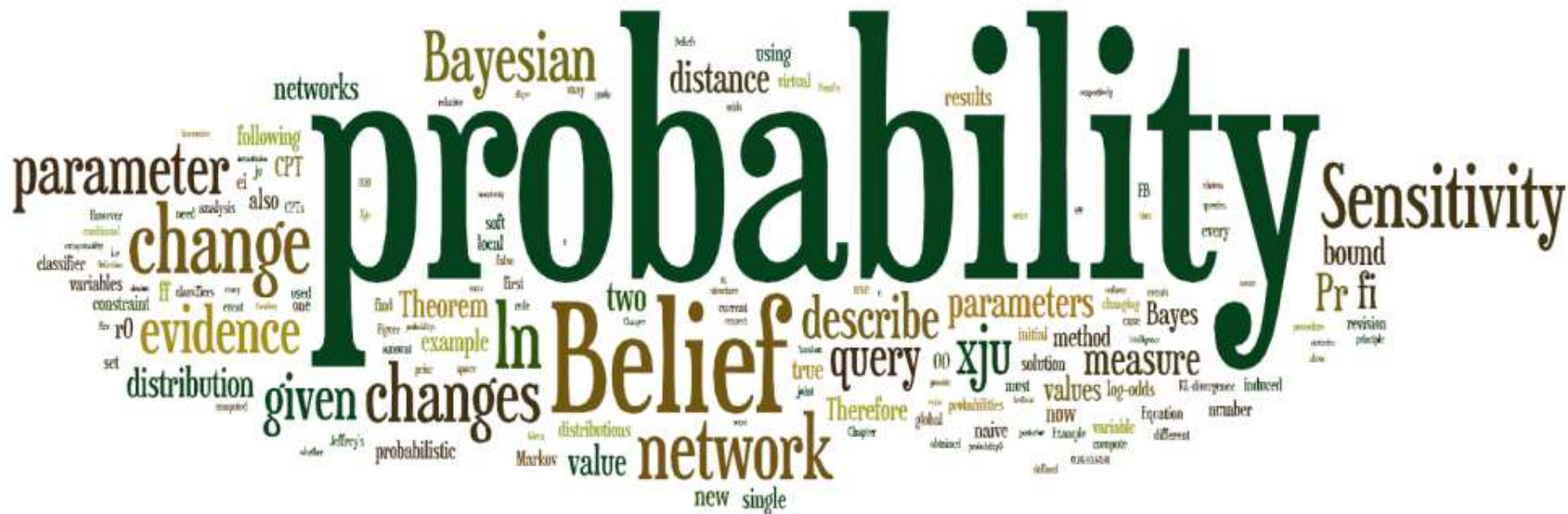


Métodos Bayesianos podem acomodar hipóteses que fazem
predições probabilísticas
(Ex: Este paciente tem uma chance de 93% de se recuperar)



Novas instâncias podem ser classificadas combinando a probabilidade de múltiplas hipóteses ponderadas pelas suas probabilidades.





$P(\text{Doença} = \text{presente}) = 0.08$

$P(\text{Doença} = \text{ausente}) = 0.92$

$P(\text{Teste} = \text{positivo} \mid \text{Doença} = \text{presente}) = 0.75$

$P(\text{Teste} = \text{negativo} \mid \text{Doença} = \text{ausente}) = 0.96$

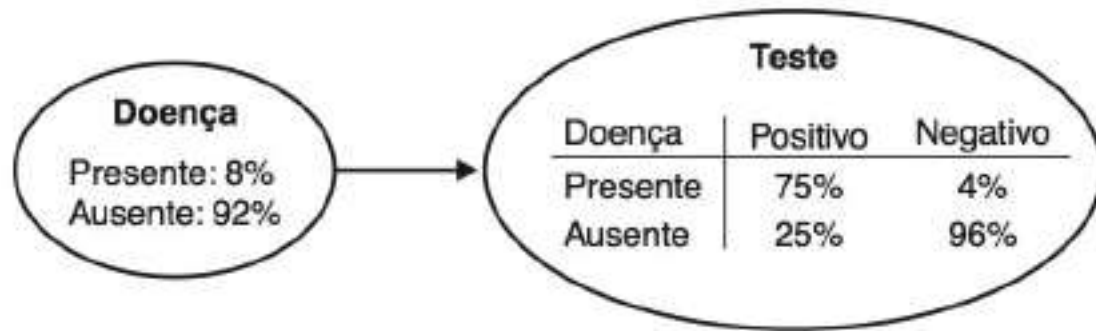


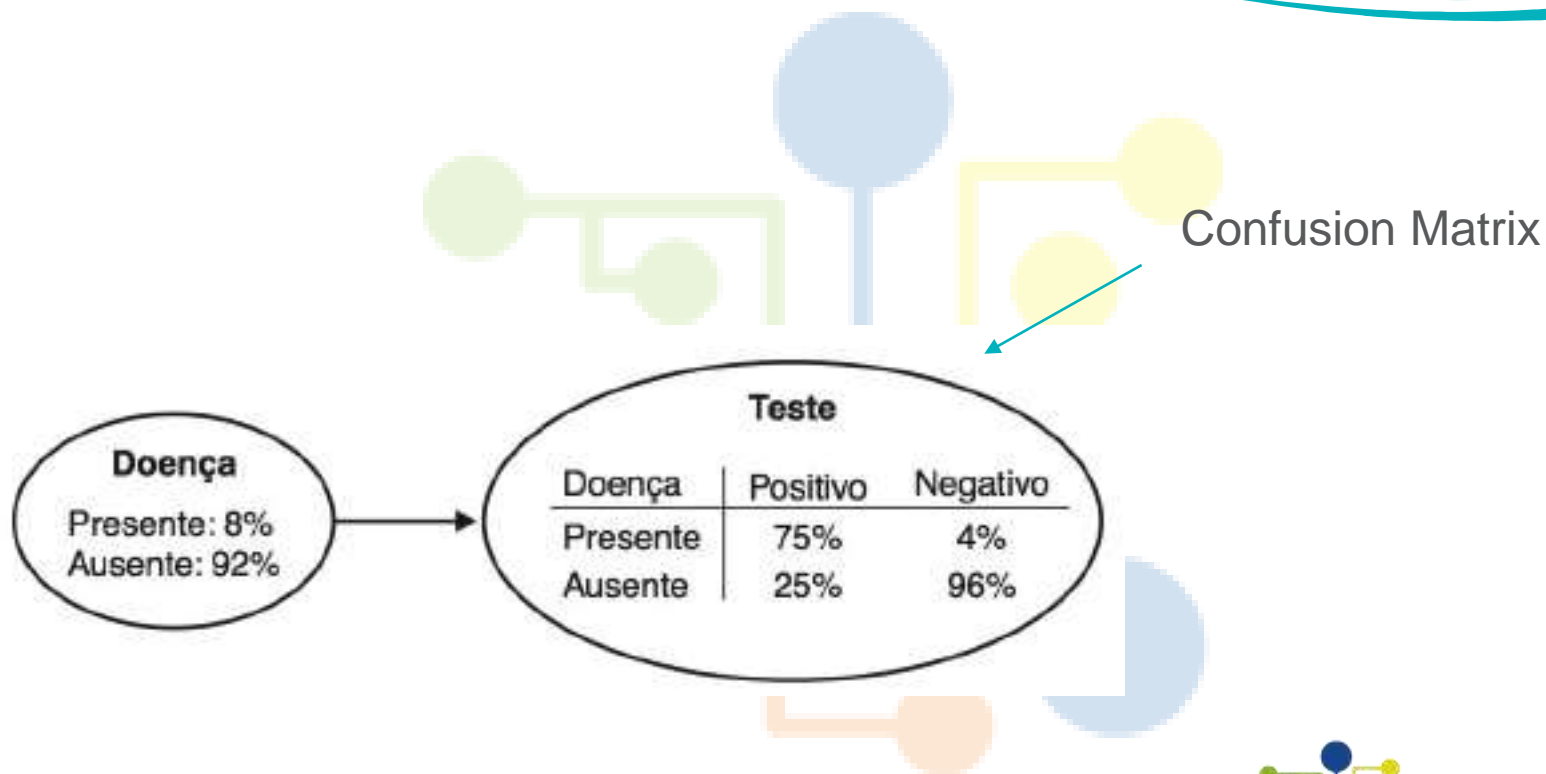
$$P(\text{Doença} = \text{presente}) = 0.08$$

$$P(\text{Doença} = \text{ausente}) = 0.92$$

$$P(\text{Teste} = \text{positivo} \mid \text{Doença} = \text{presente}) = 0.75$$

$$P(\text{Teste} = \text{negativo} \mid \text{Doença} = \text{ausente}) = 0.96$$





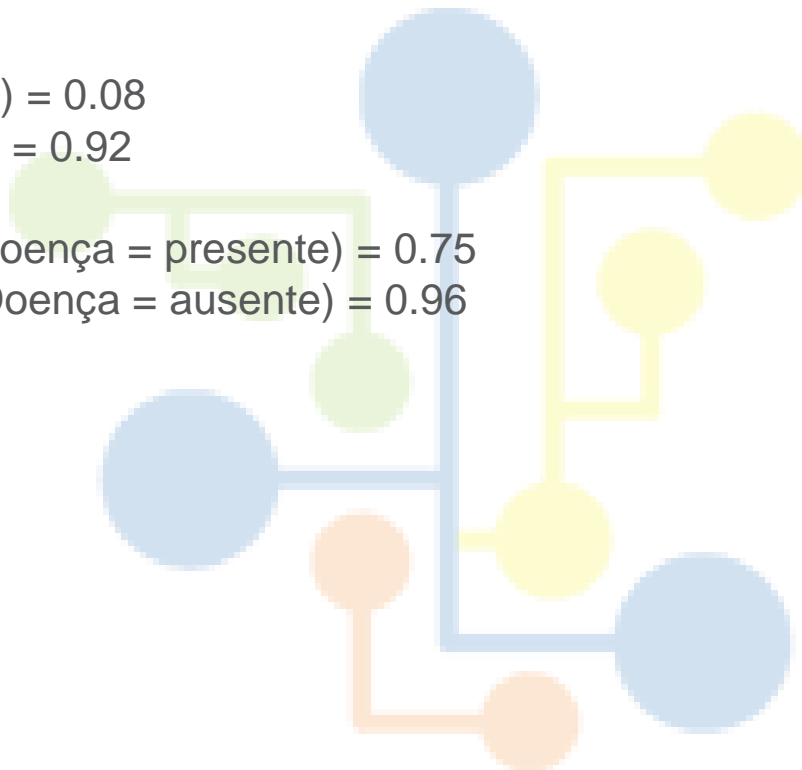
$$P(\text{Doença} = \text{presente}) = 0.08$$

$$P(\text{Doença} = \text{ausente}) = 0.92$$

$$P(\text{Teste} = \text{positivo} \mid \text{Doença} = \text{presente}) = 0.75$$

$$P(\text{Teste} = \text{negativo} \mid \text{Doença} = \text{ausente}) = 0.96$$

$$P(A) = P(A|B) \times P(B)$$



P(Teste = positivo) =

$$\begin{aligned} &= P(\text{Teste} = \text{positivo} \mid \text{Doença} = \text{presente}) \times P(\text{Doença} = \text{presente}) \\ &+ P(\text{Teste} = \text{positivo} \mid \text{Doença} = \text{ausente}) \times P(\text{Doença} = \text{ausente}) \\ &= 0.75 \times 0.08 \times 0.04 \times 0.92 = 0.0968 \end{aligned}$$

P(Teste = negativo) =

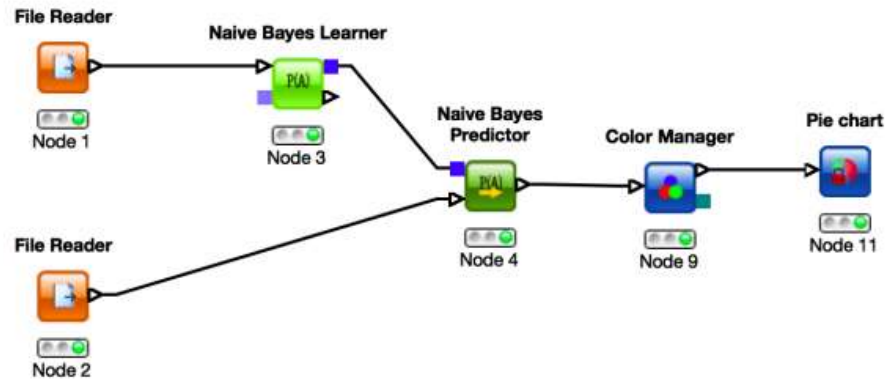
$$\begin{aligned} &= P(\text{Teste} = \text{negativo} \mid \text{Doença} = \text{presente}) \times P(\text{Doença} = \text{presente}) \\ &+ P(\text{Teste} = \text{negativo} \mid \text{Doença} = \text{ausente}) \times P(\text{Doença} = \text{ausente}) \\ &= 0.25 \times 0.08 \times 0.96 \times 0.92 = 0.9032 \end{aligned}$$



Naive Bayes



Naive Bayes



O classificador Naïve Bayes é baseado na suposição simplificadora de que os valores dos atributos são condicionalmente independentes dado o valor alvo.





É possível obter classificadores bayesianos de complexidade crescente, que consideram diferentes graus de dependências entre atributos





Data Science Academy

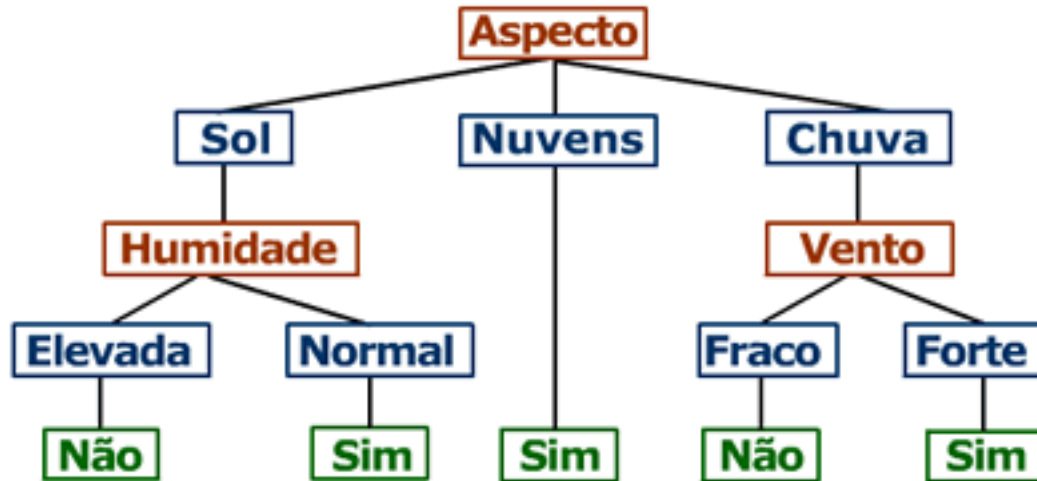
Métodos Baseados em Procura



Data Science Academy

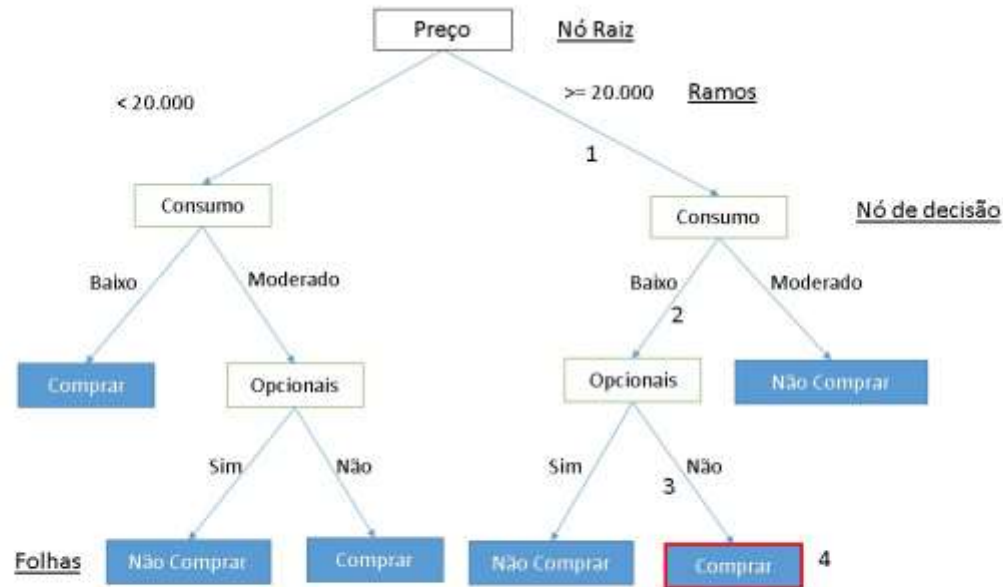
Árvores de Decisão





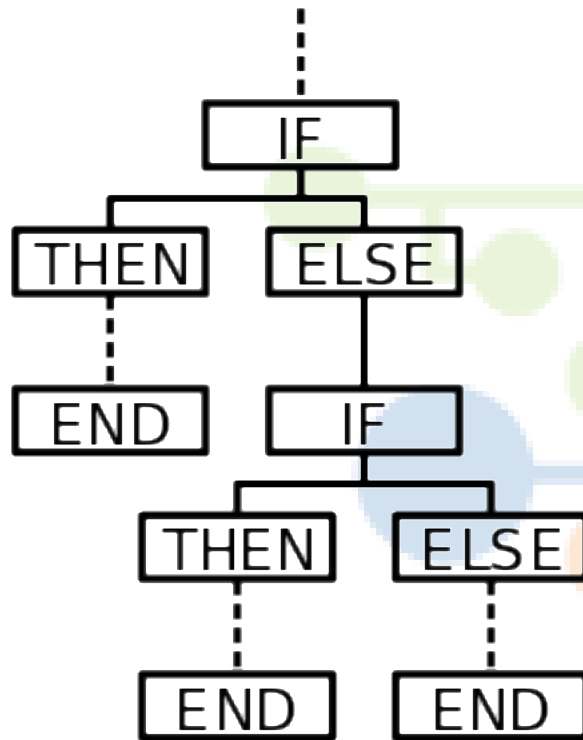
Árvores de Decisão





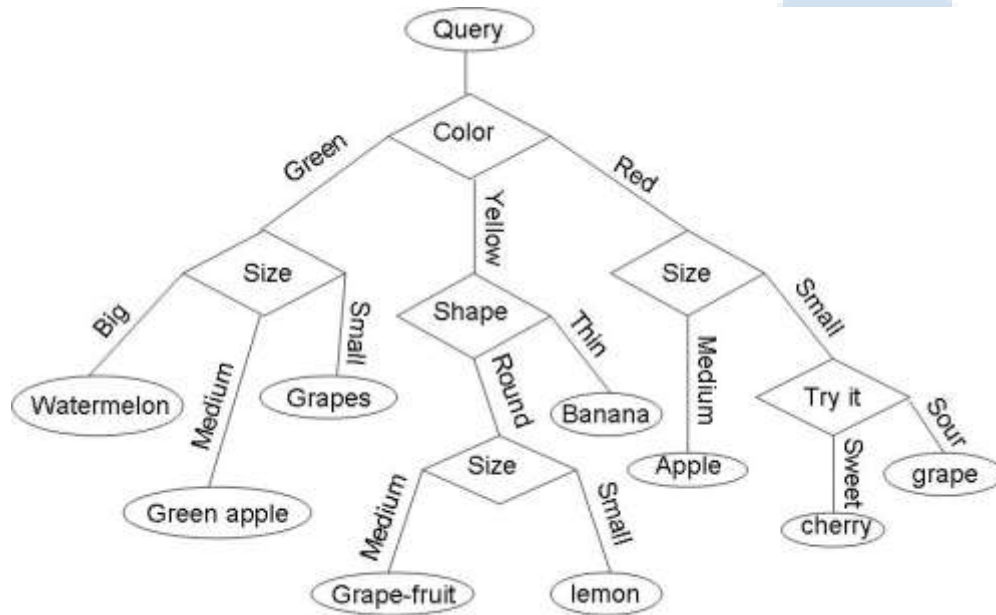
Árvores de Decisão





Árvores de decisão também podem ser representadas como conjuntos de regras SE-ENTÃO (IF-THEN)





Árvores de decisão classificam instâncias ordenando as árvores acima (ou abaixo), a partir da raiz até alguma folha

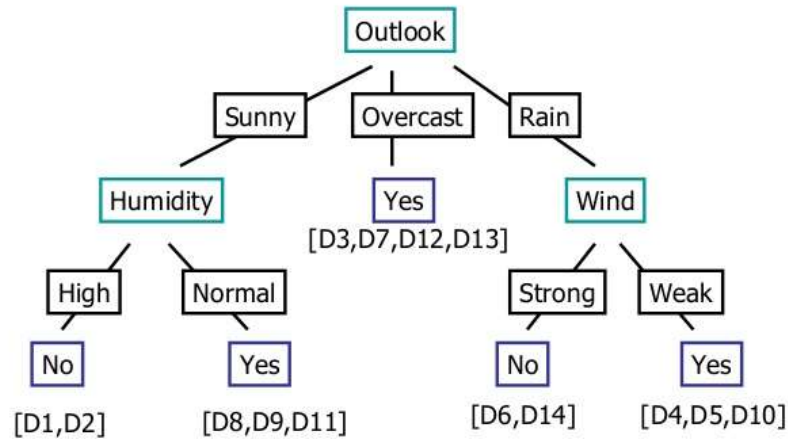




ID3 (Quinlan, 1986)

C4.5 (Quinlan, 1993)

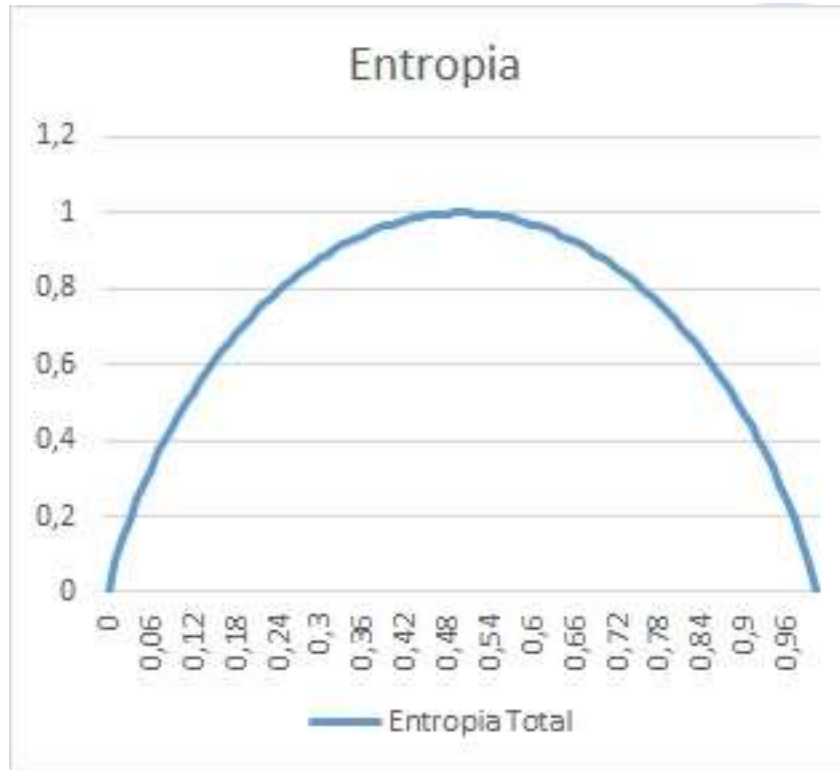




Algoritmo ID3

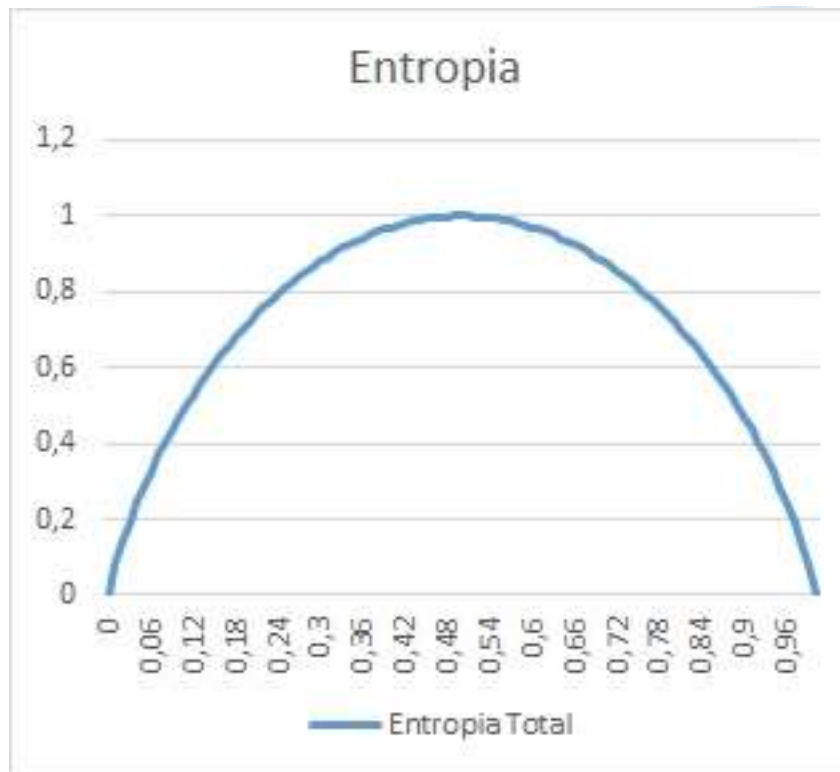
Entropia





A Física usa o termo entropia para descrever a quantidade de desordem associada a um sistema





A incerteza ou impureza em um nó pode ser medida através da Entropia

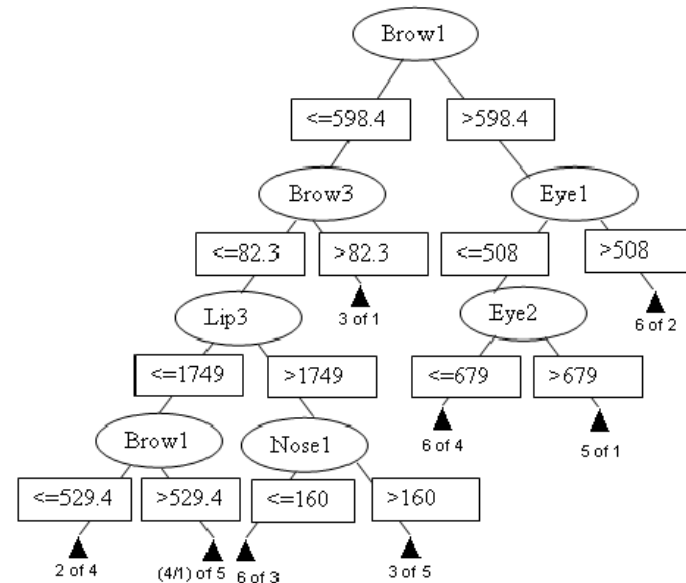
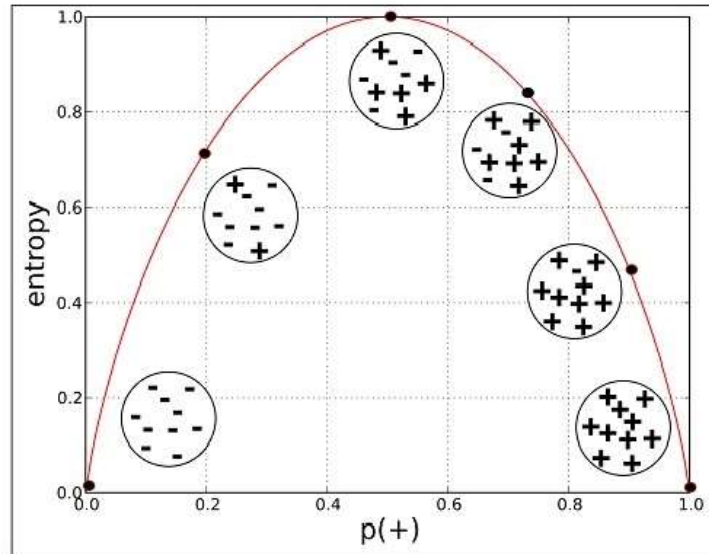
Se todos os exemplos são da mesma classe, então a entropia assume valor mínimo

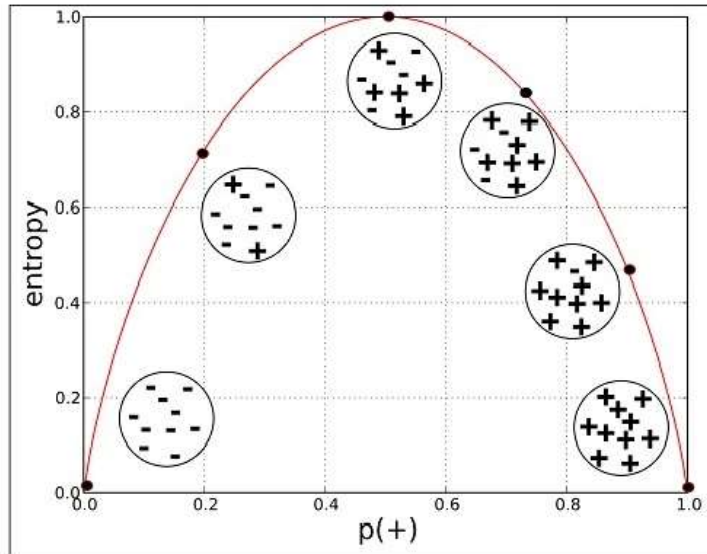
Se todas as classes têm o mesmo número de exemplos então a entropia assume o valor máximo



O algoritmo ID3 segue os seguintes passos:

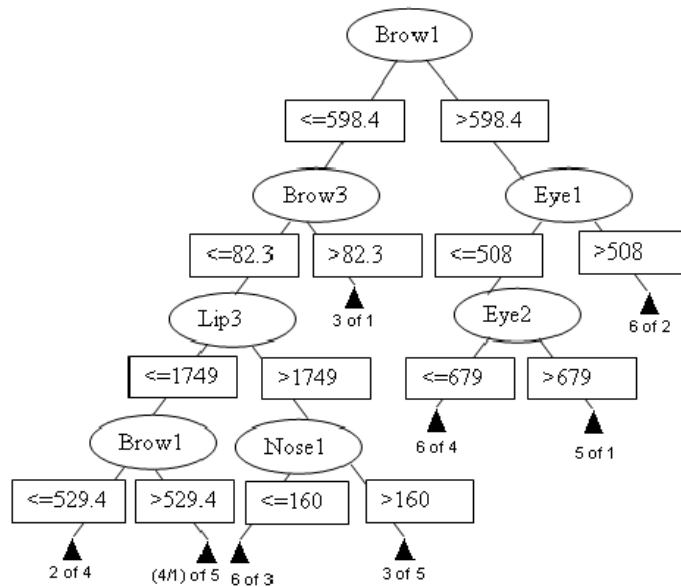






$$Entropia(S) = \sum p_i \log_2 p_i$$





Ganho de Informação



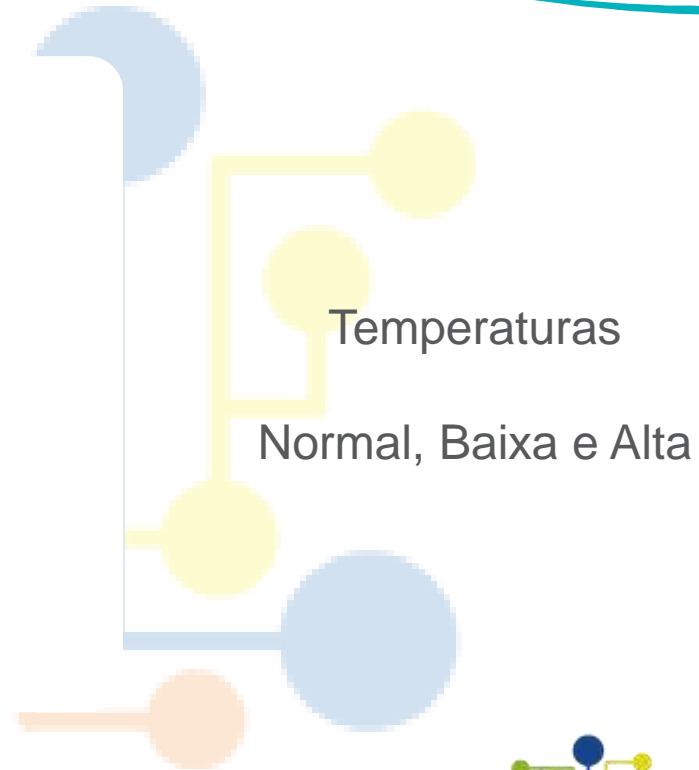
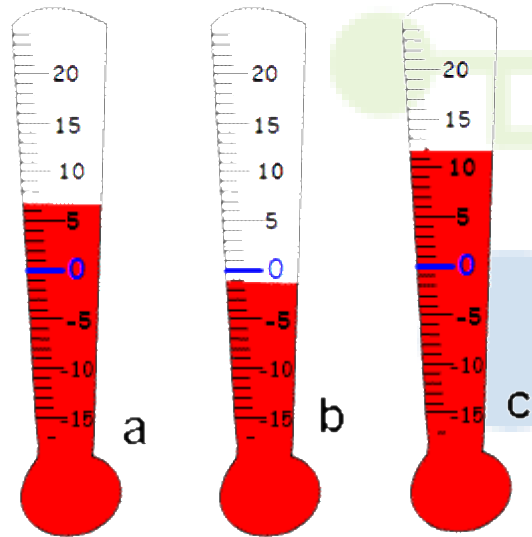
Entropia mede o nível de certeza que temos sobre um evento



Entropia mede o nível de certeza que temos sobre um evento

Ganho de Informação mede a efetividade de um atributo em classificar um conjunto de treinamento







Data Science Academy

Métodos Baseados em Otimização



Data Science Academy

Redes Neurais Artificiais

SVM
(Support Vector Machines)



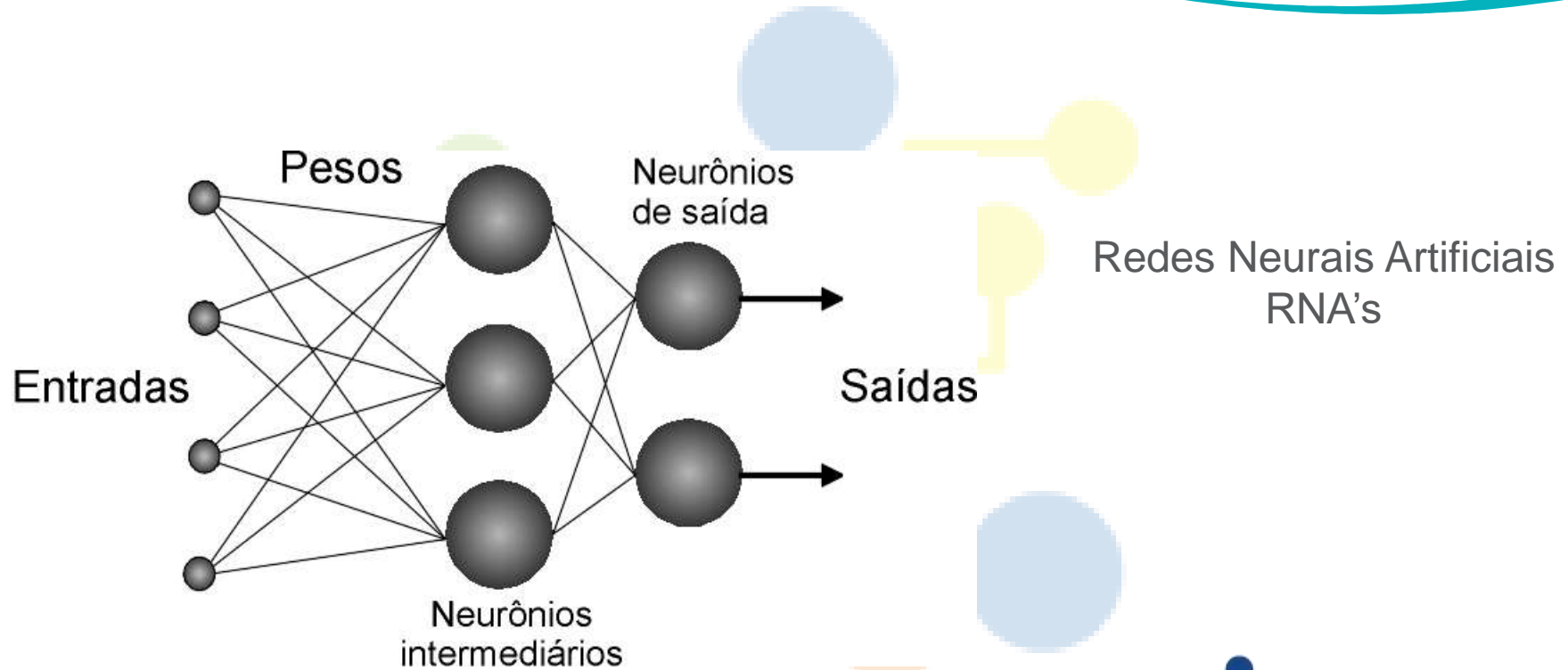


Data Science Academy

Redes Neurais Artificiais

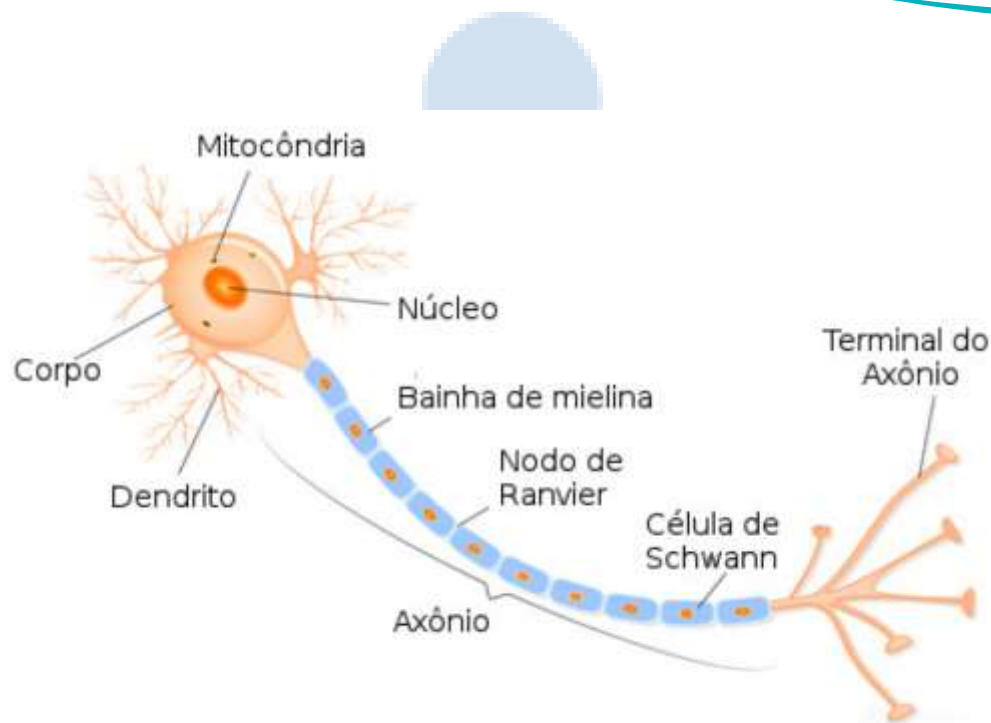


Data Science Academy





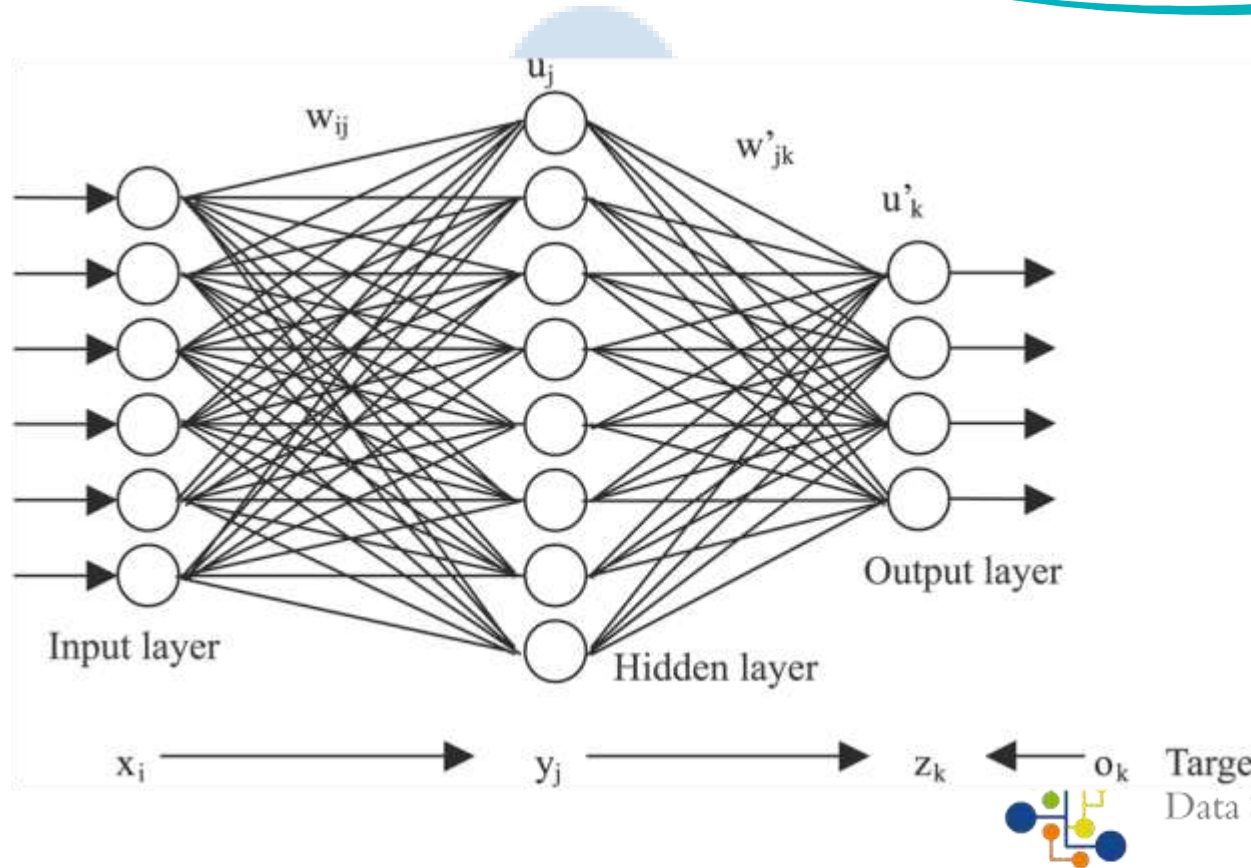


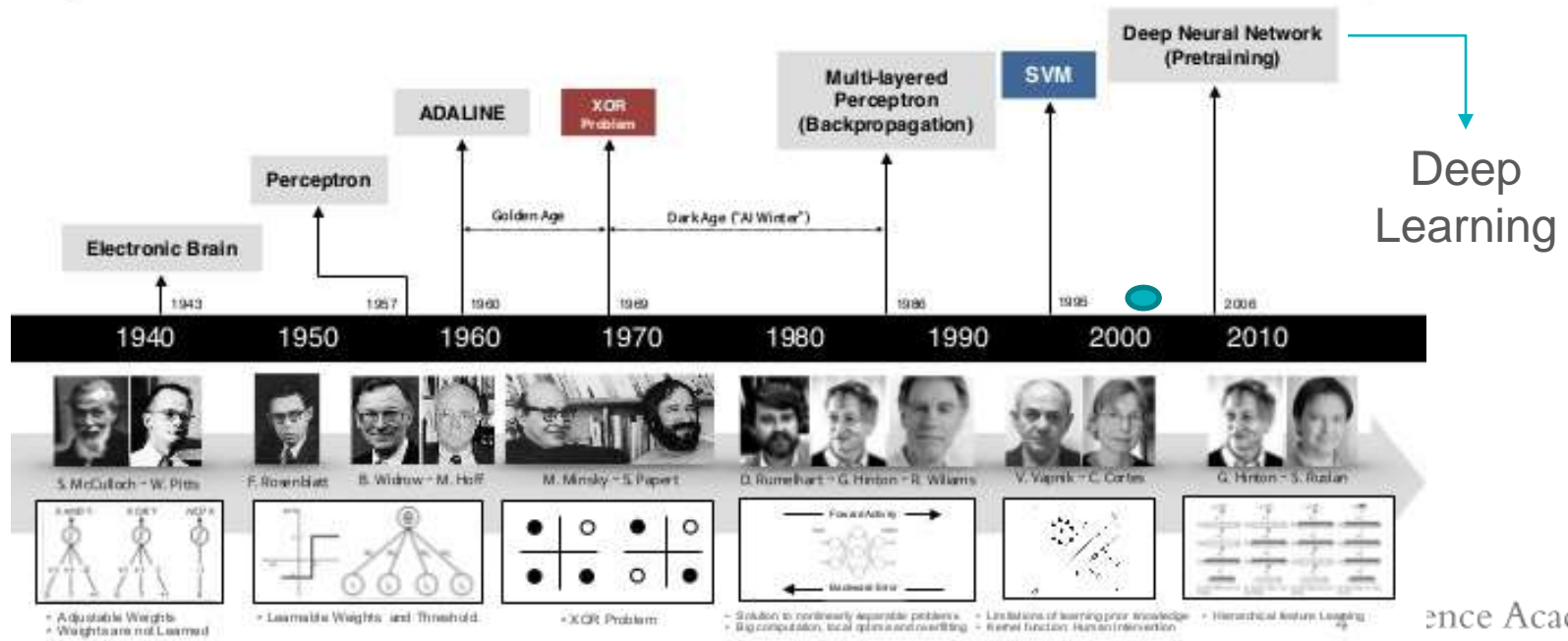


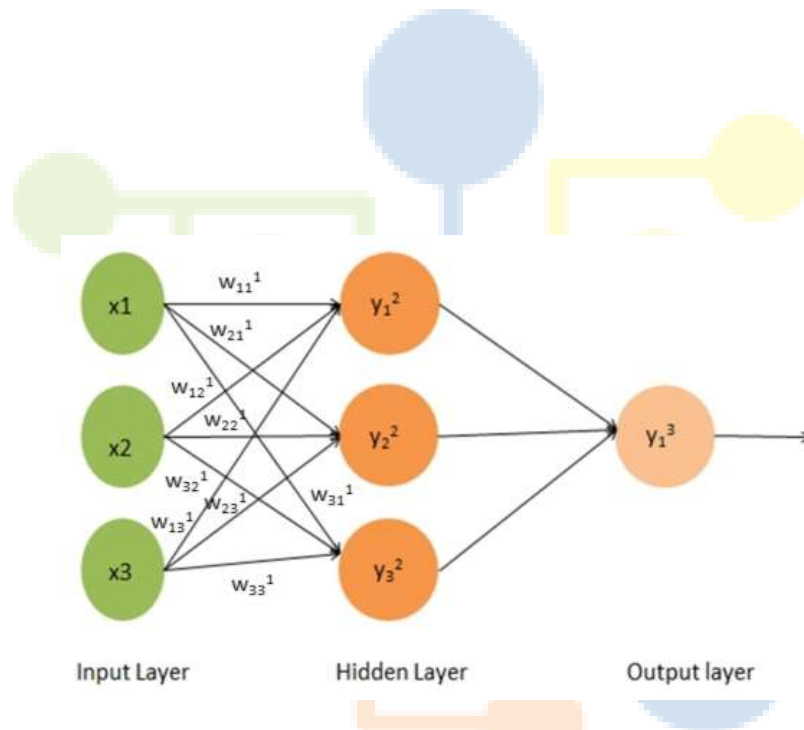
Fórmula do Neurônio Artificial proposto por McCulloch e Pitts em 1943.

$$y = f(x) \left(\sum_{j=1}^n w_j x_j - \alpha \right)$$

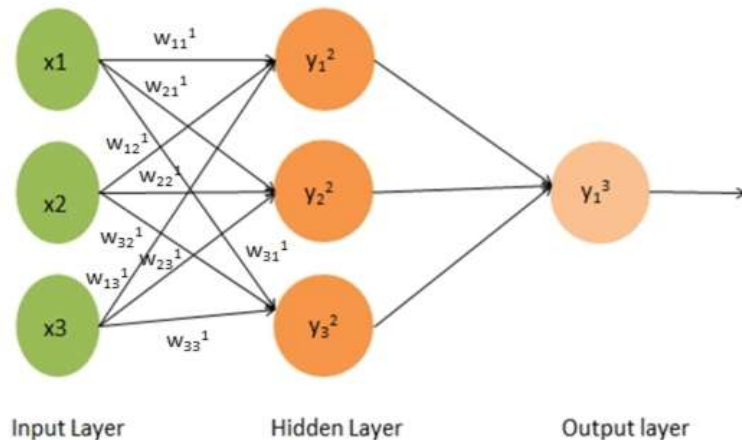








$$g(w_{11}x_1 + w_{12}x_2 + w_{13}x_3)$$



função semi-linear:

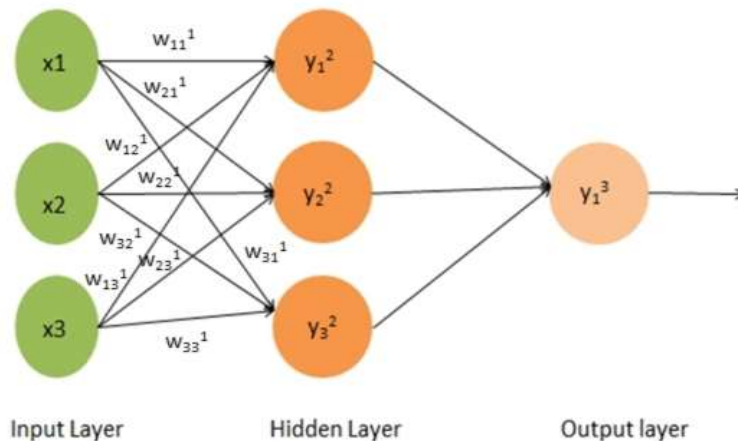
$$f(x) = \begin{cases} 0 & \text{se } x < \alpha_{min} \\ mx + l & \text{se } \alpha_{min} \leq x \leq \alpha_{max} \\ f_{max} & \text{se } x > \alpha_{max} \end{cases}$$

função sigmoidal:

$$f(x) = \frac{f_{max}}{1 + e^{-x}}$$



$$g(w_{110}x_0 + w_{111}x_1 + w_{112}x_2 + w_{113}x_3)$$



Processos de Aprendizado das Redes Neurais

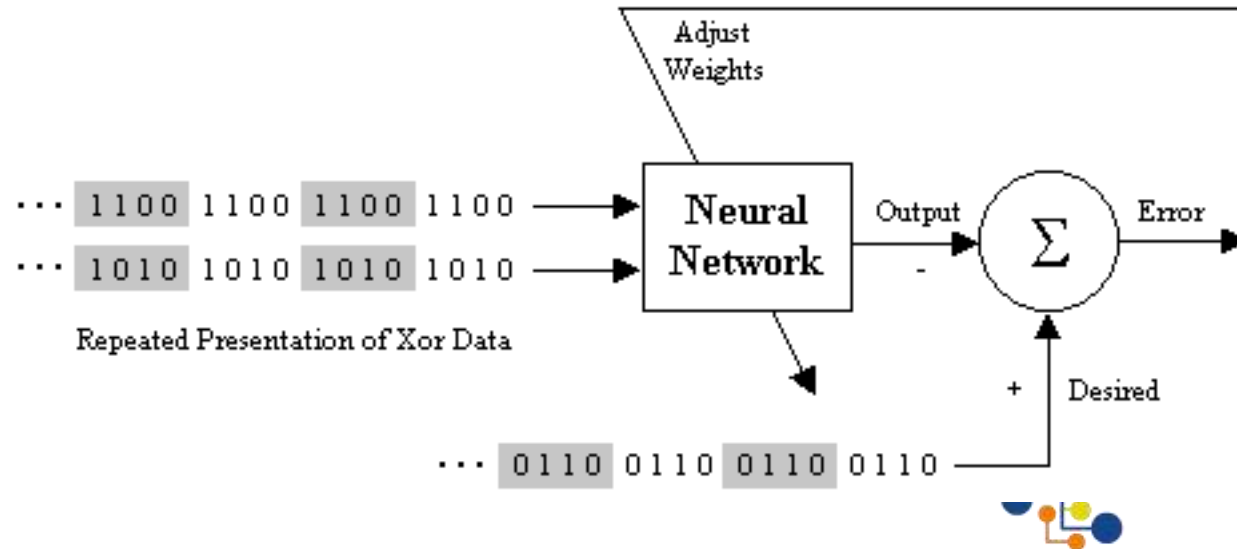


Processos de Aprendizado das Redes Neurais

- **Aprendizado Supervisionado**, quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada;
- **Aprendizado Não Supervisionado** (auto-organização), quando não existe uma agente externo indicando a resposta desejada para os padrões de entrada;
- **Aprendizado Por Reforço**, quando um crítico externo avalia a resposta fornecida pela rede.



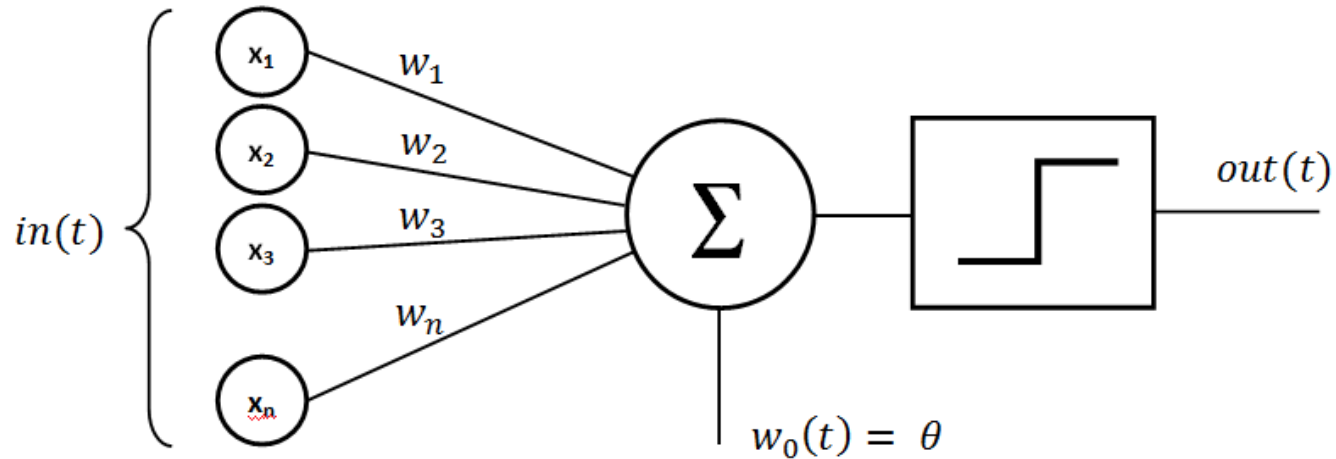
Processos de Aprendizado das Redes Neurais



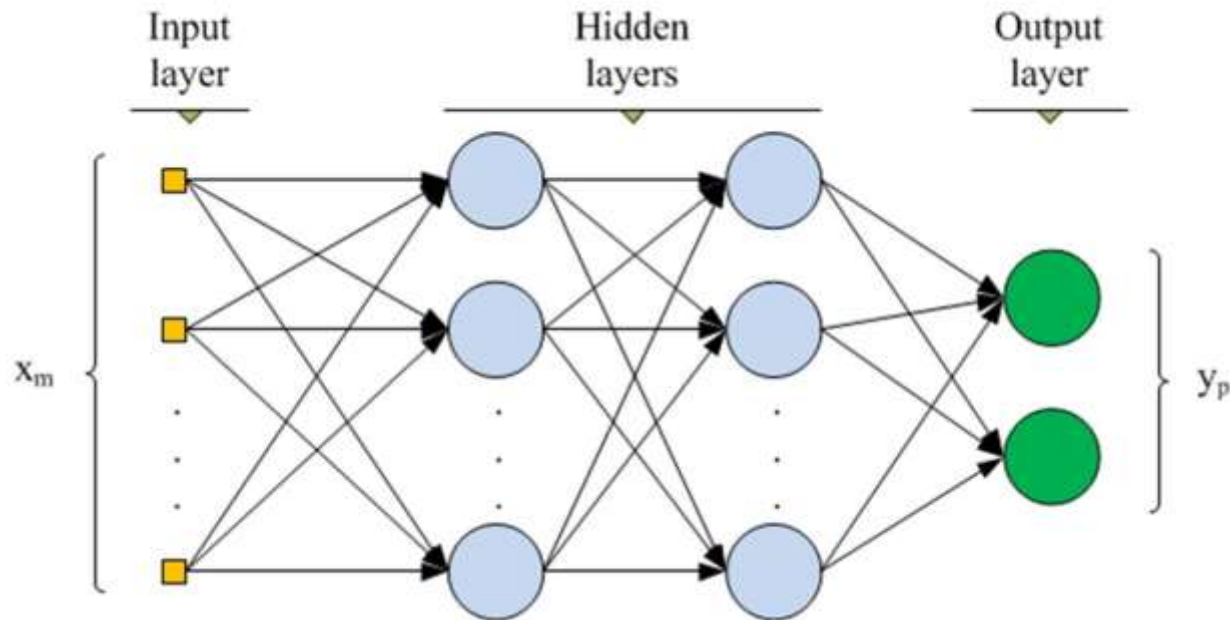
Principais Modelos de Redes Neurais



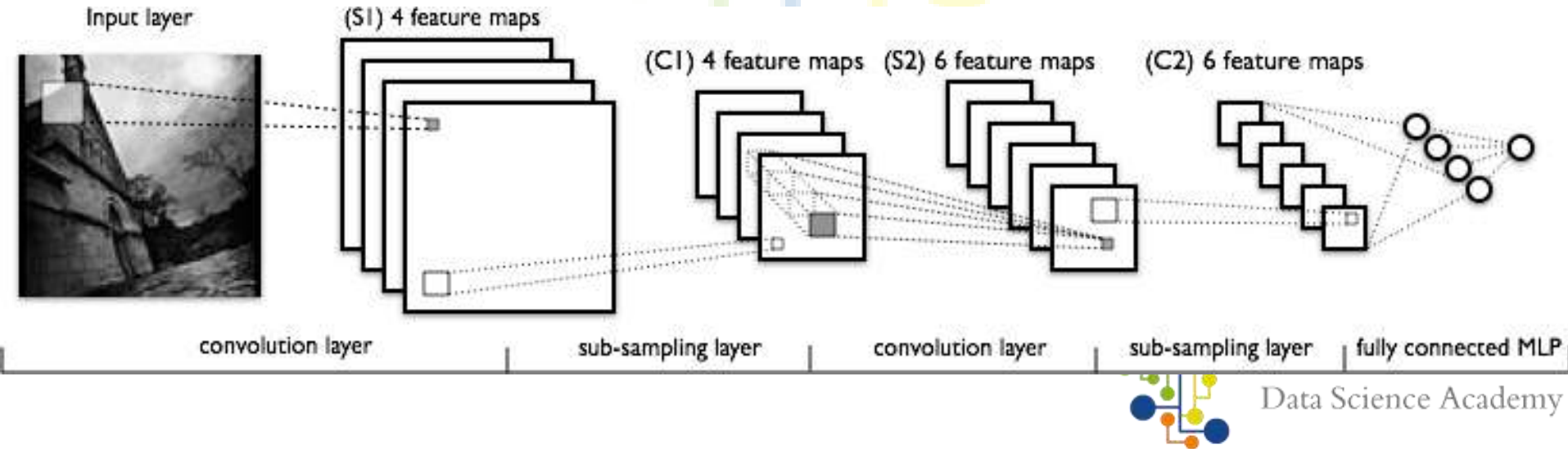
Perceptron



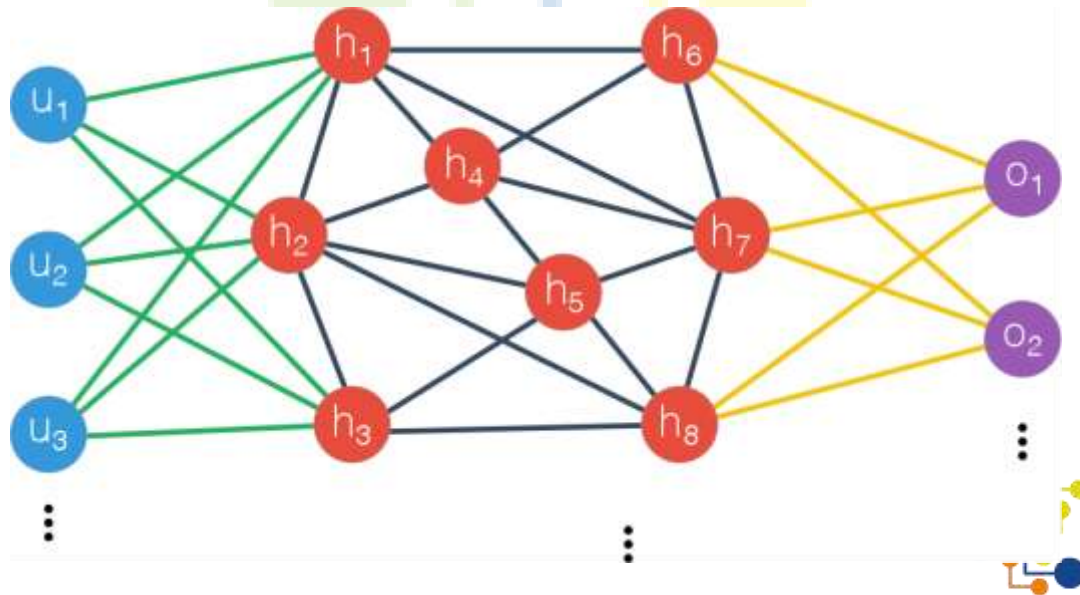
Perceptron de Múltiplas Camadas



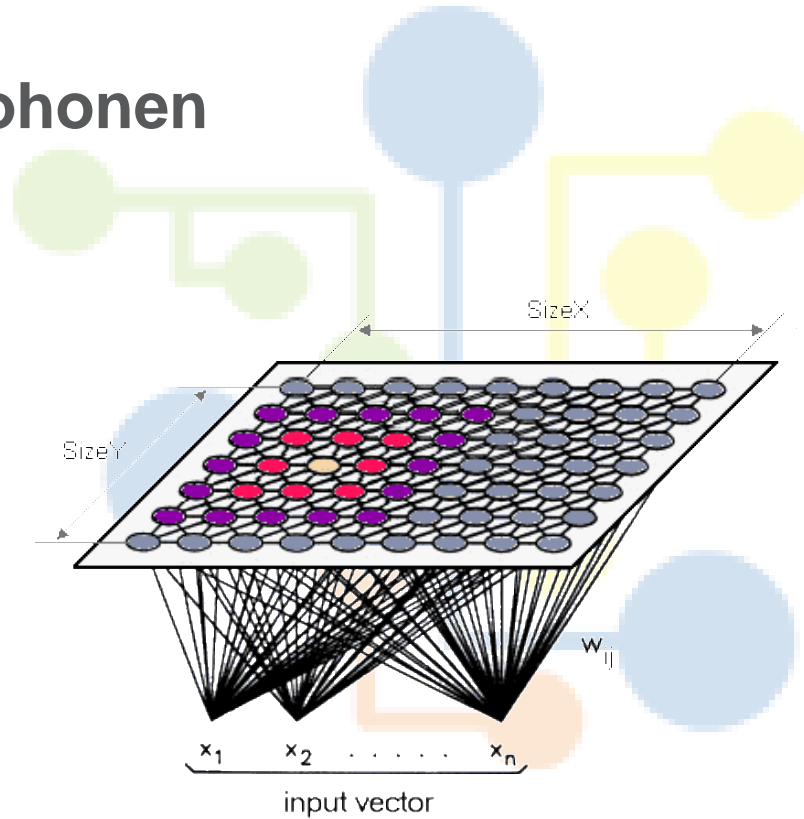
Redes Neurais Convulocionais (CNN – Convolutional Neural Networks)



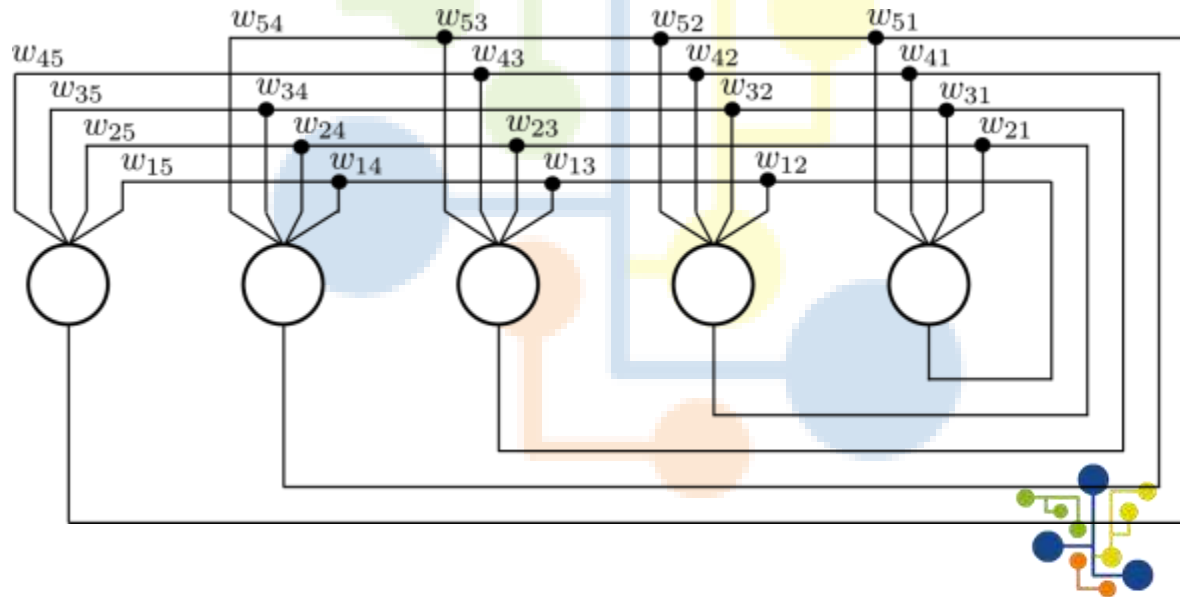
Redes Neurais Recorrentes (RNN – Recurrent Neural Networks)



Rede de Kohonen



Rede de Hopfield



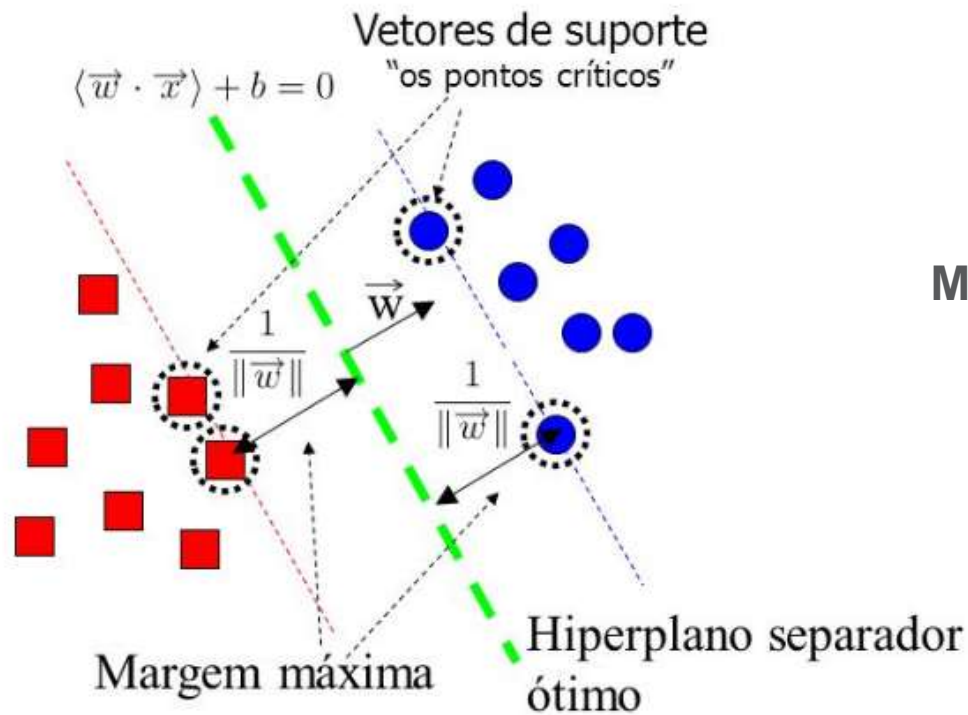


Data Science Academy

Máquinas de Vetores de Suporte

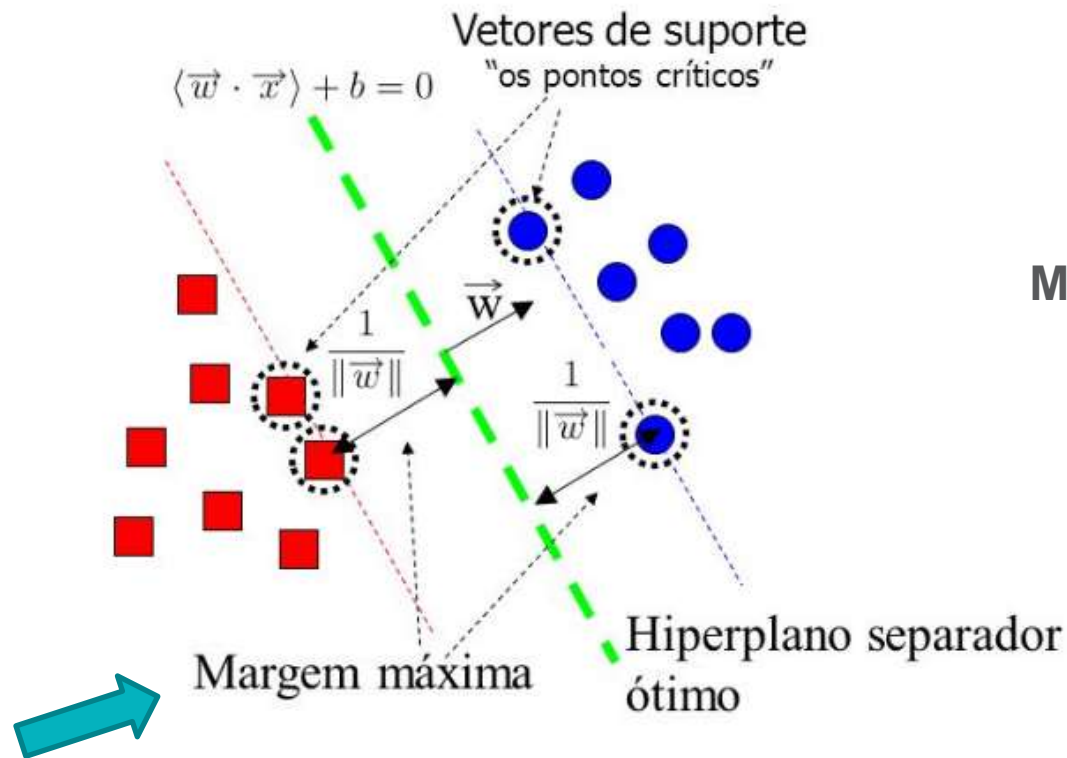


Data Science Academy



Máquina de Vetores de Suporte

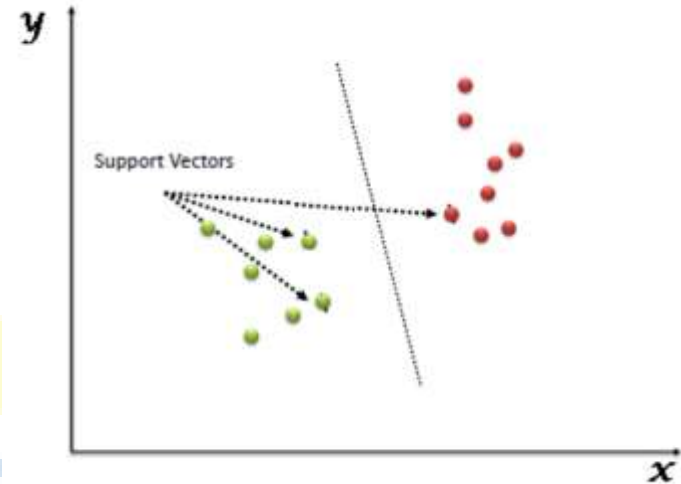




Máquina de Vetores de Suporte



Support Vector Machine



Vetores de suporte são simplesmente as coordenadas de observação individual.

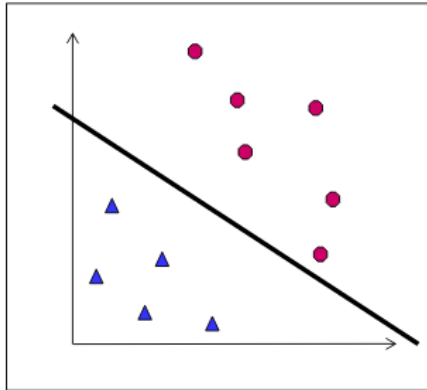
Support Vector Machine é uma fronteira (hiperplano) que melhor segrega as duas classes



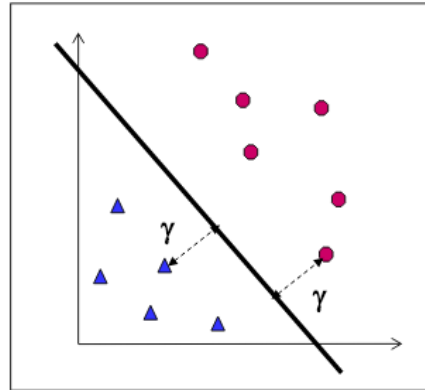
A SVM é uma técnica de aprendizado estatístico, baseada no princípio da Minimização do Risco Estrutural (SRM) e pode ser usada para resolver problemas de classificação



Rede Neural



SVMs



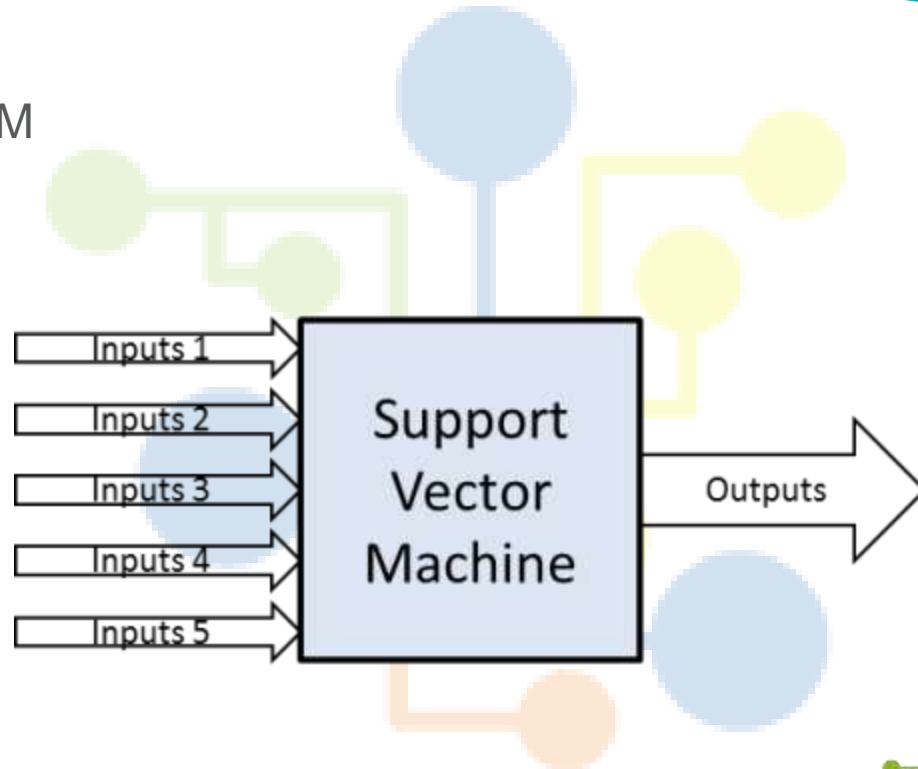
As máquinas de vetores de suporte são consideradas uma outra categoria das redes neurais



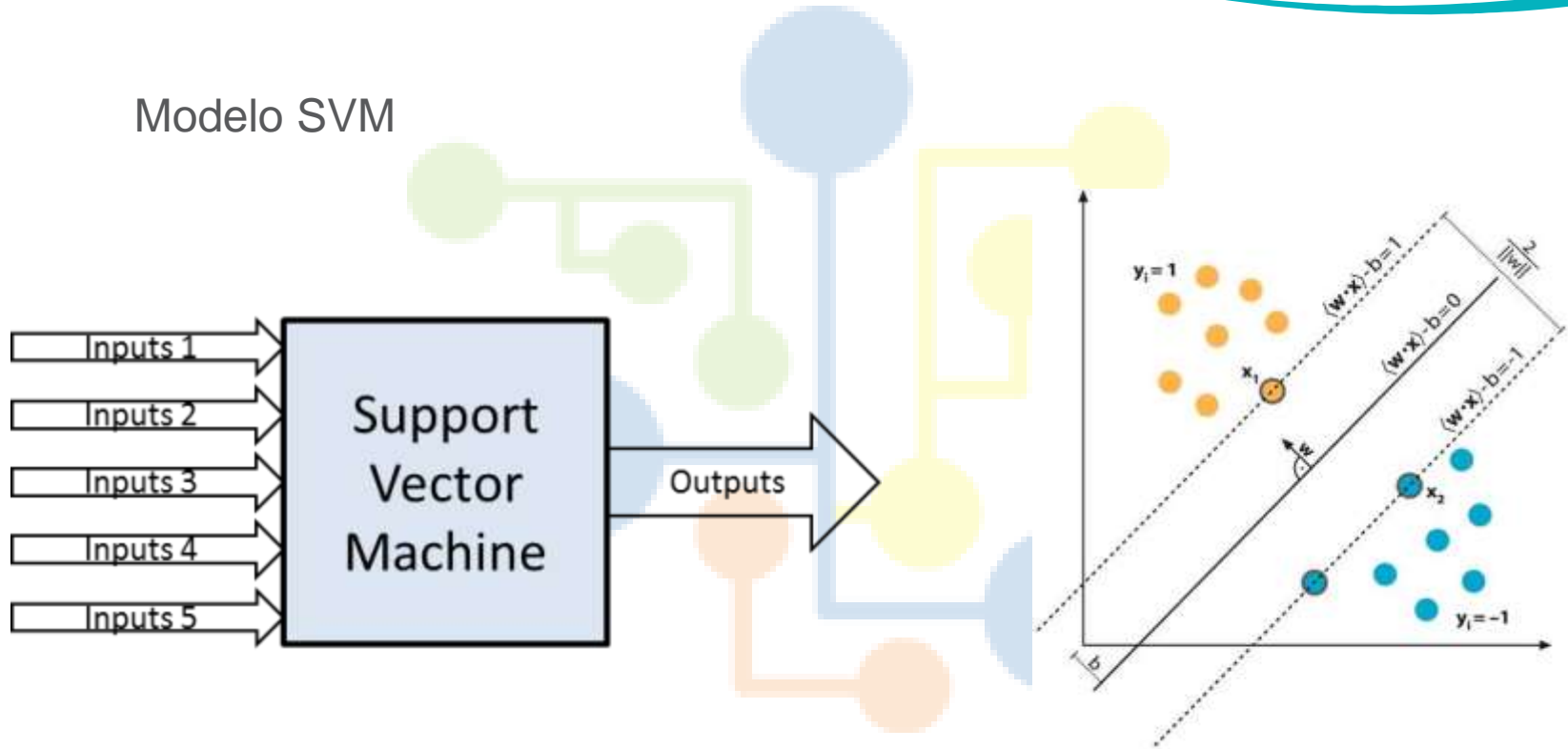
O uso de SVM's é capaz de resolver problemas de classificação de dados, gerando classificadores que apresentam bons resultados. Porém, esses classificadores possuem uma limitação de interpretabilidade, não sendo possível compreender a saída obtida



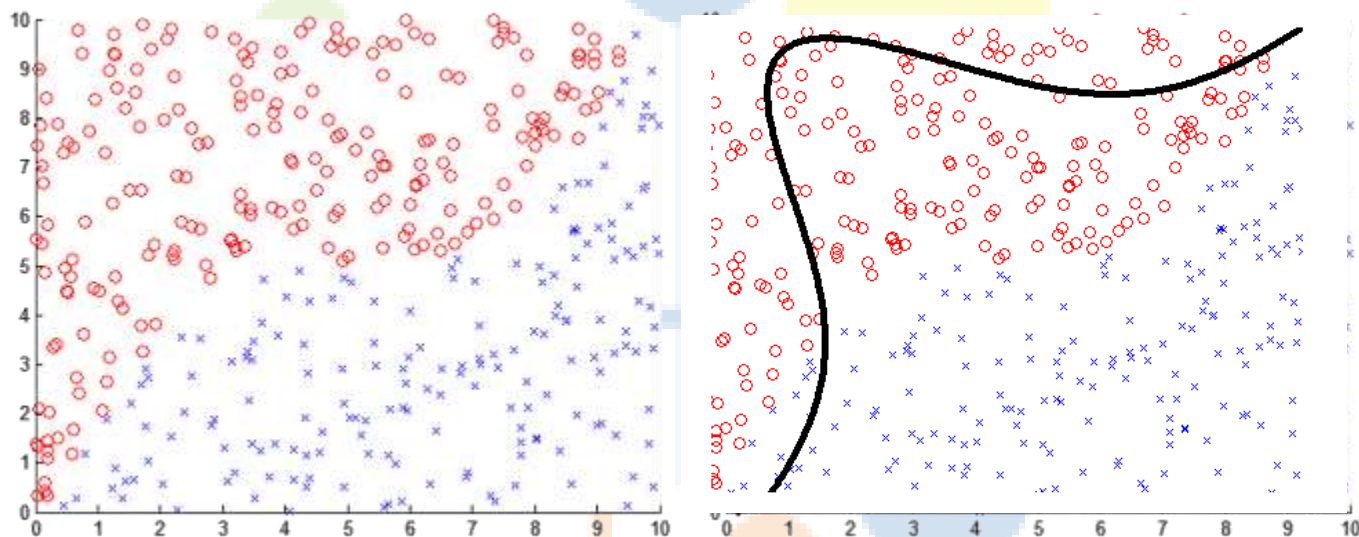
Modelo SVM



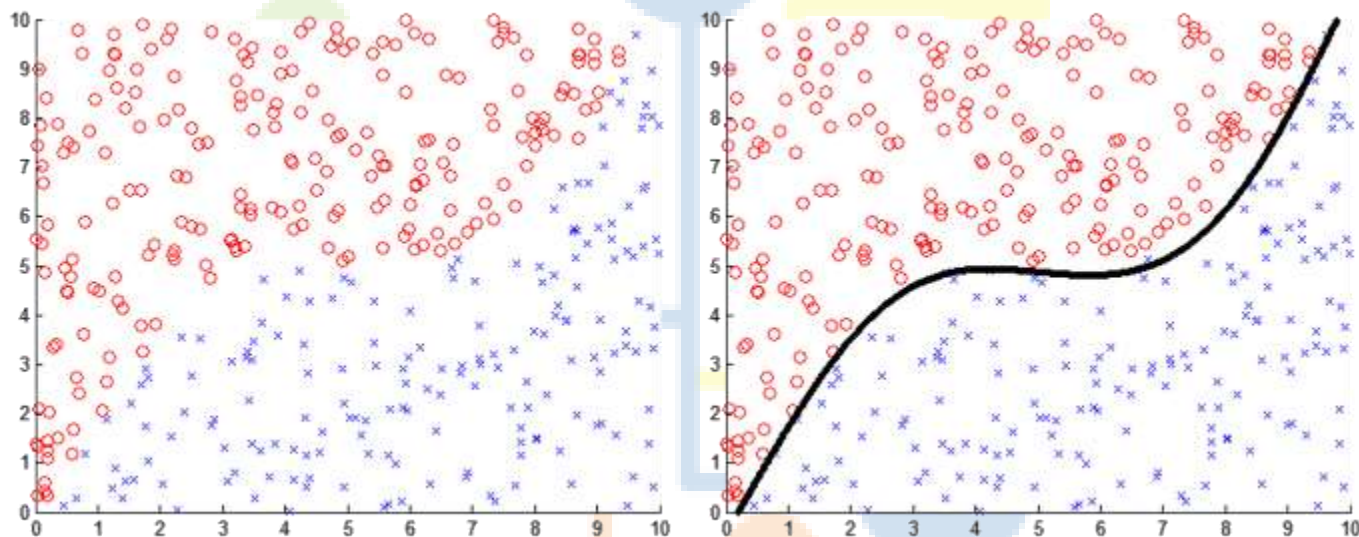
Modelo SVM



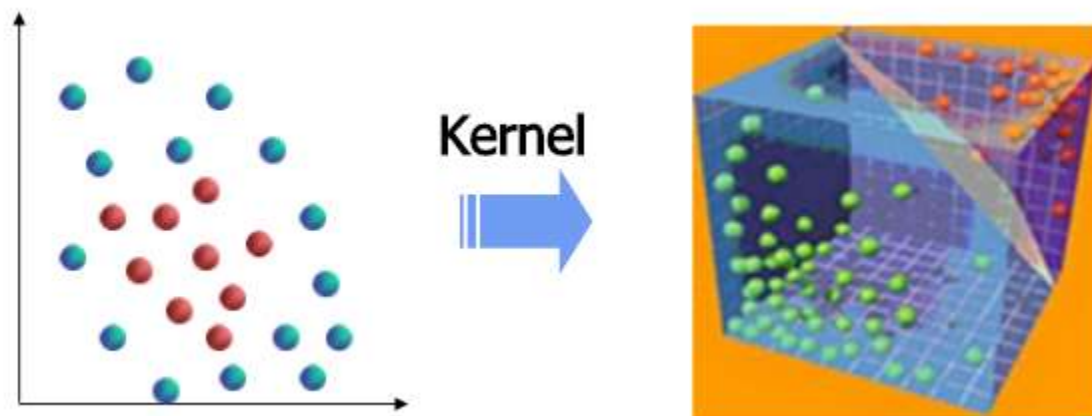
Modelo SVM com 2 Dimensões



Modelo SVM com 2 Dimensões



Modelo SVM com Múltiplas Dimensões





Boa capacidade de
generalização



Robustez em grandes
dimensões



Convexidade da função objetivo

Principais Características das
SVM's



Teoria bem definida



Outras características:

Em caso de outlier a SVM busca a melhor forma possível de classificação e, se necessário, desconsidera o outlier.

É um classificador criado para fornecer separação linear.

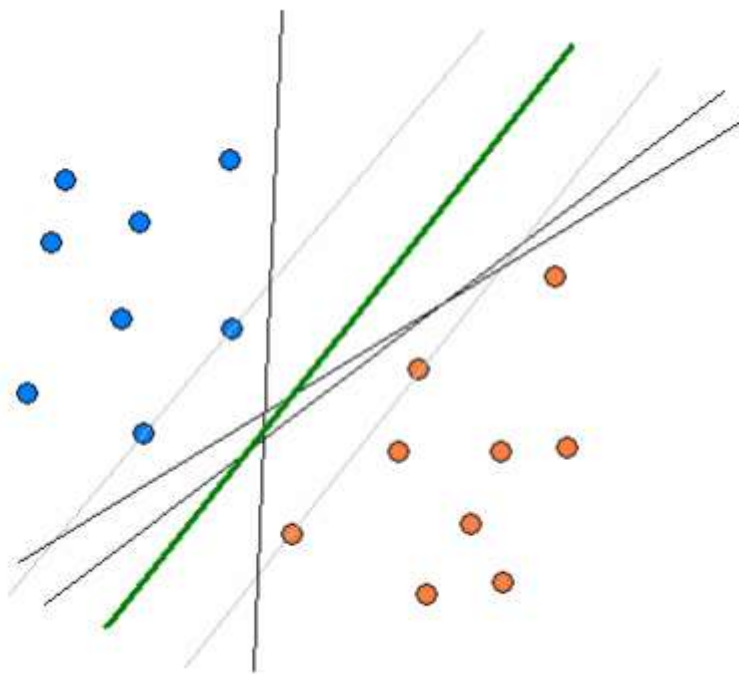
Funciona muito bem em domínios complicados, em que existe uma clara margem de separação.

Não funciona bem em conjuntos de dados muito grandes, pois o tempo de treinamento é muito custoso.

Não funciona bem em conjunto de dados com grande quantidade de ruídos.

Outras Características das SVM's





Por que a máquina de vetores de suporte é tão útil?





Data Science Academy

Clusterização



Data Science Academy

Supervised Learning



Unsupervised Learning



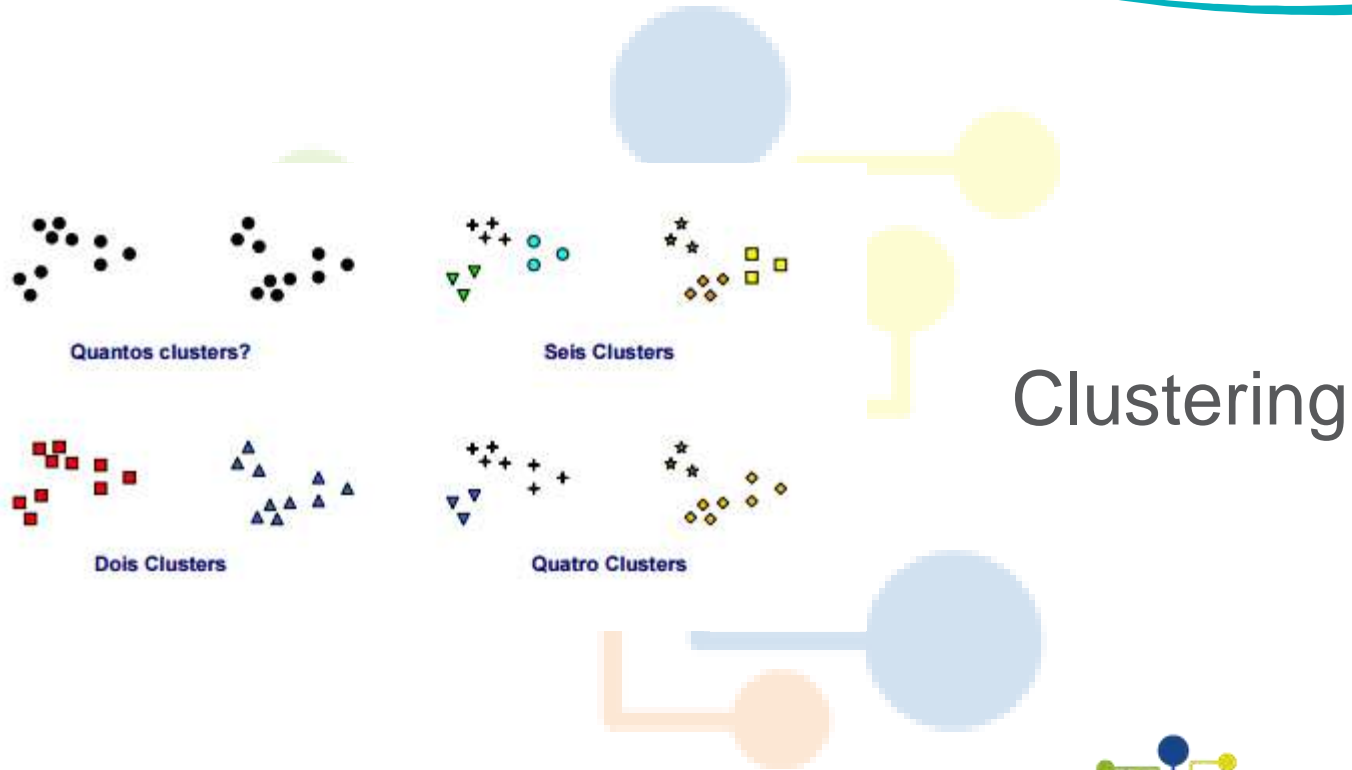


Clustering

$$X = \{X_1, X_2, \dots, X_n\}$$

$$C = \{C_1, C_2, \dots, C_k\}$$





Medidas de Distância

Distância euclidiana: considera a distância entre dois elementos X_i e X_j no espaço p-dimensional:

$$d(X_i, X_j) = \left[\sum_{l=1}^p (x_{il} - x_{jl})^2 \right]^{\frac{1}{2}}$$

Distância “city-block”: corresponde à soma das diferenças entre todos os p atributos de dois elementos X_i e X_j , não sendo indicada para os casos em que existe uma correlação entre tais atributos:

$$d(X_i, X_j) = \sum_{l=1}^p |x_{il} - x_{jl}|$$



Normalização \Rightarrow variáveis com mesmo peso.

- **Min-Max para um atributo f :**

$$s_f = \frac{x_{if} - \min_f}{\max_f - \min_f} \times (\text{novoMax} - \text{novoMin}) + \text{novoMin}$$

- **Z-score** $z_{if} = \frac{x_{if} - m_f}{\sigma_f}$

- **Desvio absoluto médio**

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

Clustering



Agrupamento (Clustering) é a tarefa de dividir a população ou pontos de dados em um número de grupos de tal forma que os pontos de dados nos mesmos grupos são mais semelhantes a outros pontos de dados no mesmo grupo do que aqueles em outros grupos. Em palavras simples, o objetivo é segregar grupos com traços semelhantes e atribuí-los em clusters.





Agrupar todos os clientes de sua locadora de automóveis em 10 grupos com base em seus hábitos de compra e usar uma estratégia separada para os clientes em cada um desses 10 grupos. Isso é Clusterização.



Tipos de Clustering

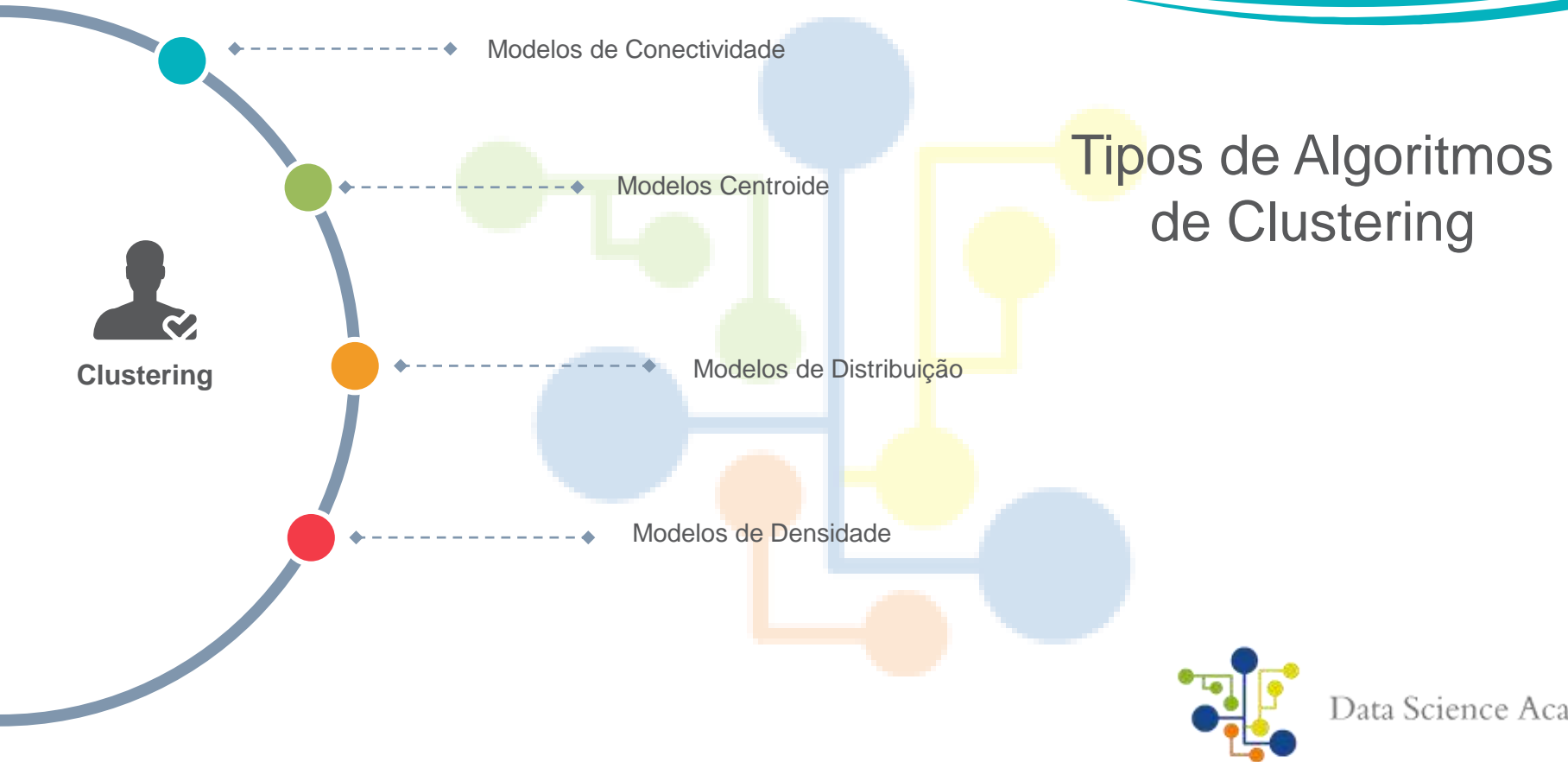
Hard Clustering

Soft Clustering



Tipos de Algoritmos de Clustering



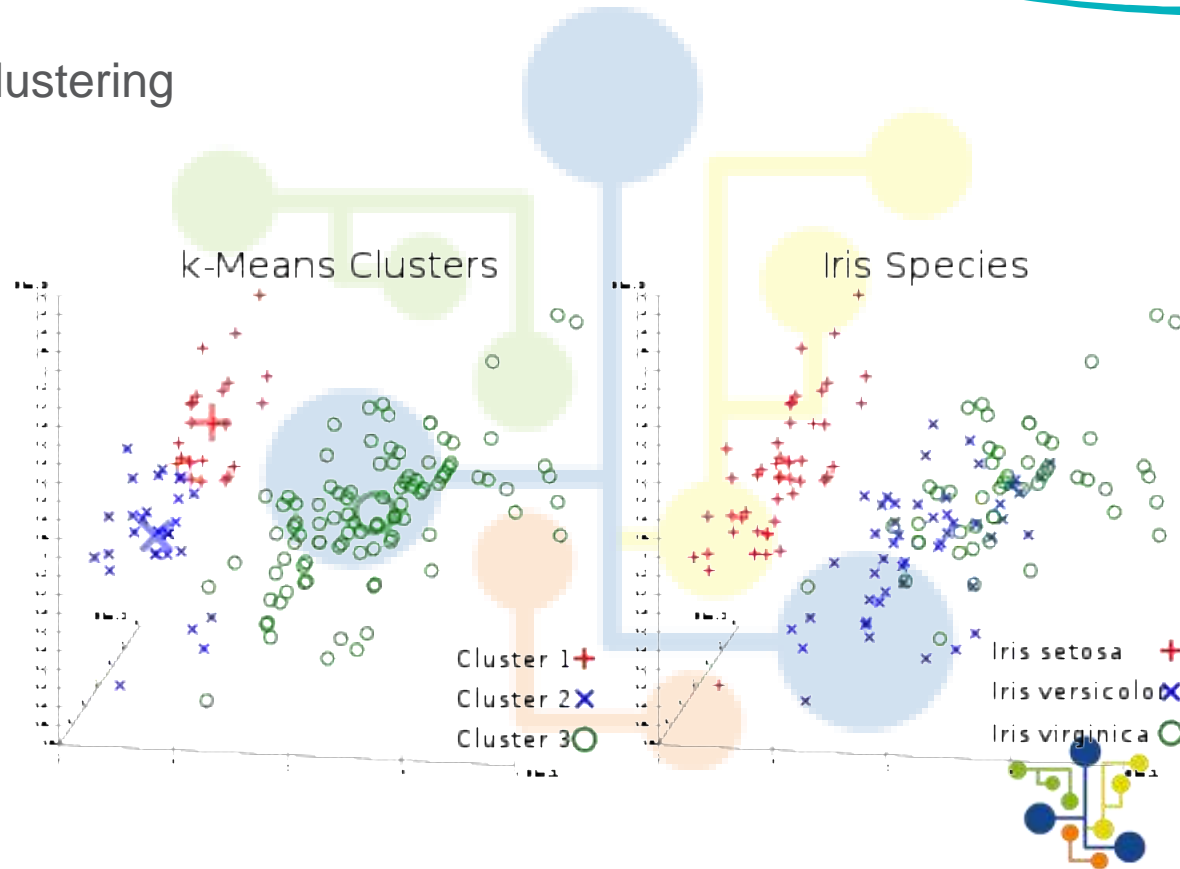


Métodos Utilizados para Clusterização

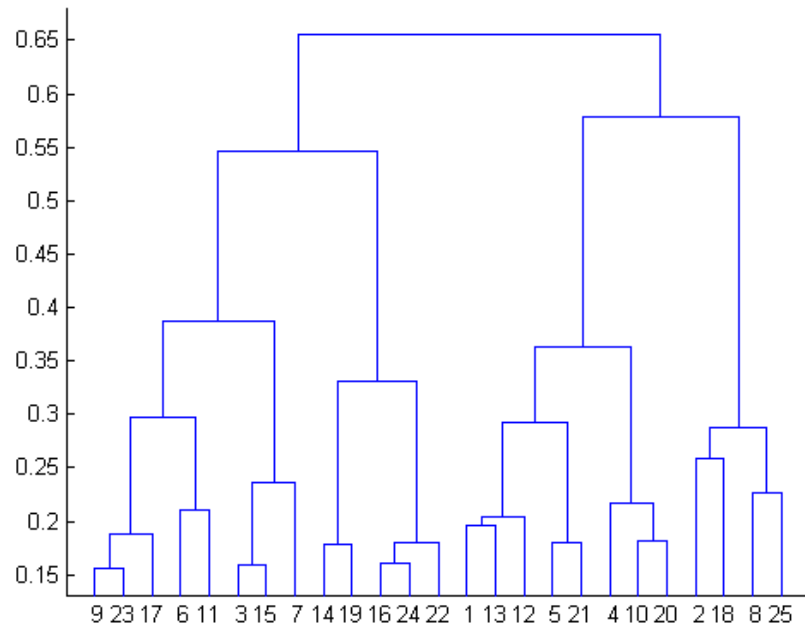
As heurísticas existentes para a solução de problemas de clusterização podem ser classificadas, de forma geral, em métodos hierárquicos e métodos de particionamento



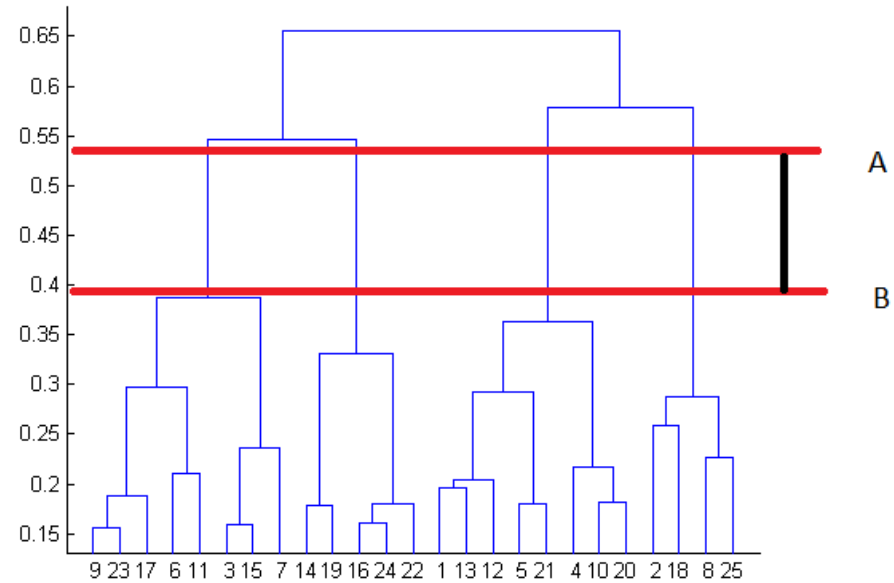
K-Means Clustering



Hierarchical Clustering



Hierarchical Clustering



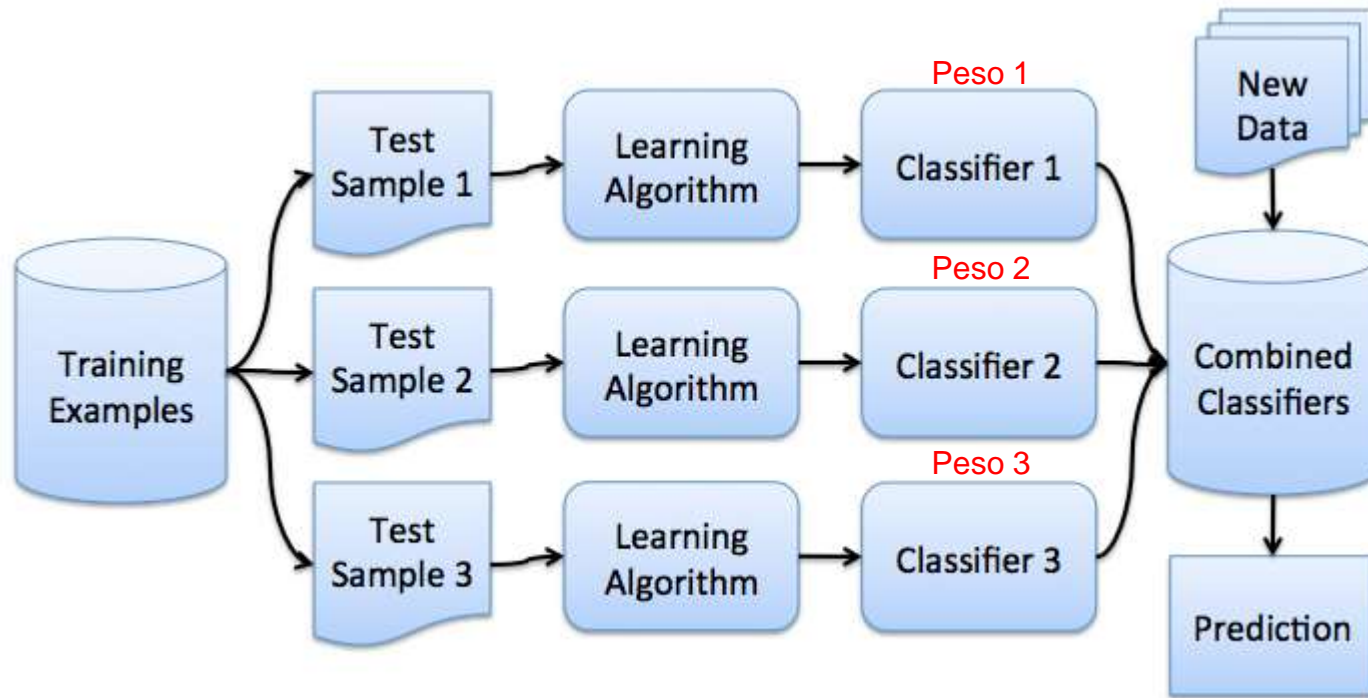


Data Science Academy

Métodos Ensemble



Data Science Academy





Métodos Ensemble

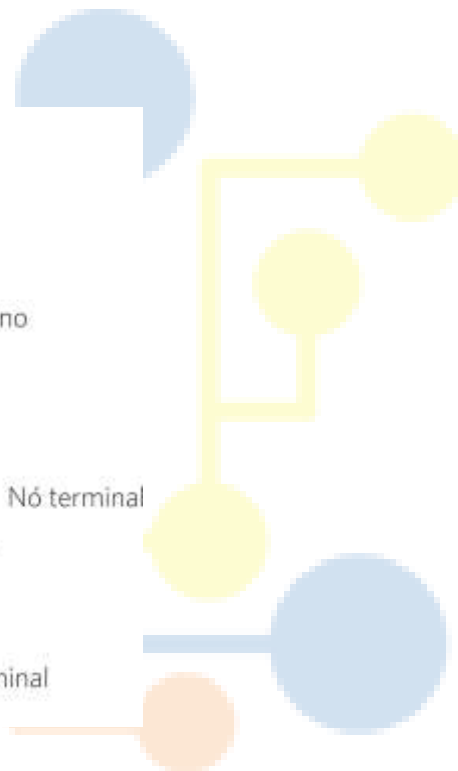
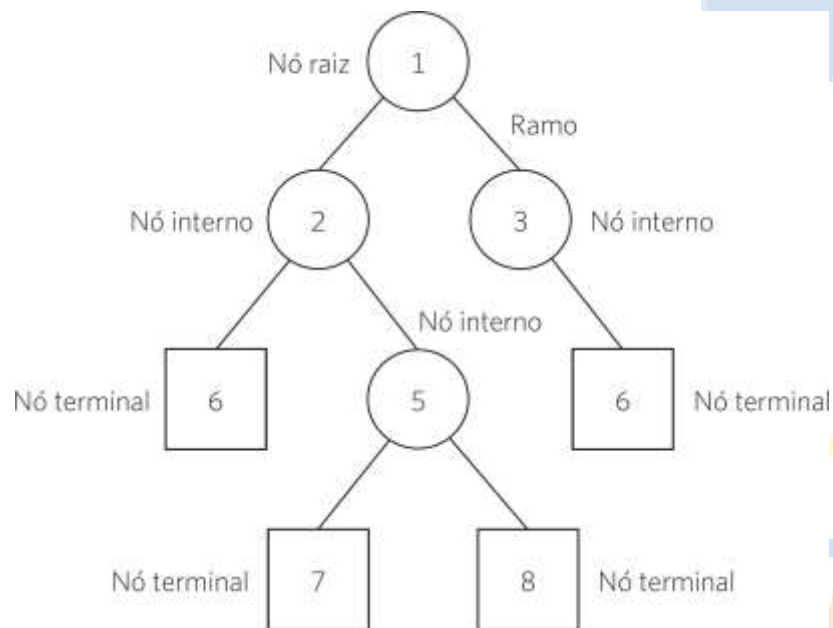


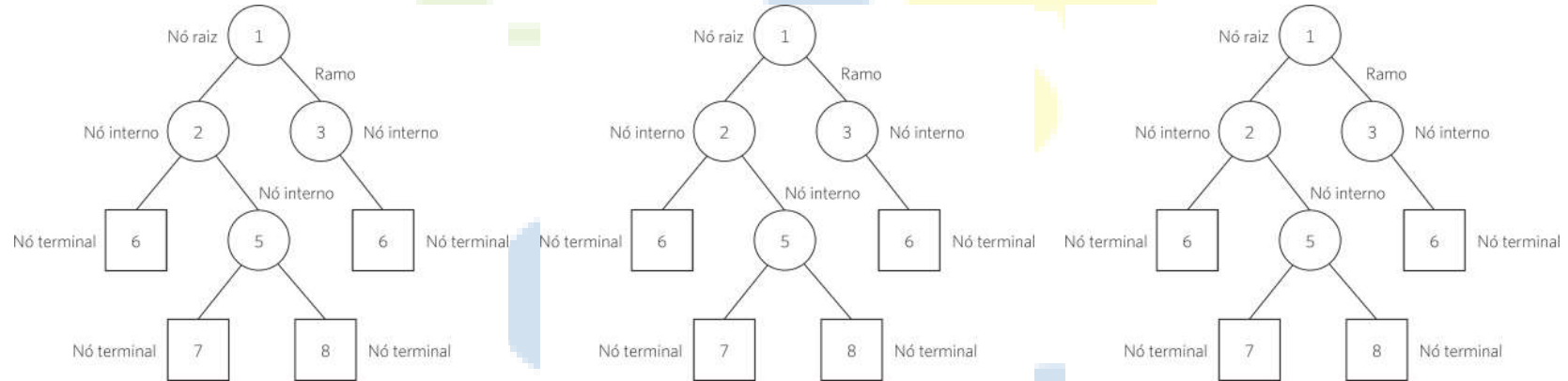


As duas respostas possíveis são:

- Cliente cancela
- Cliente não cancela





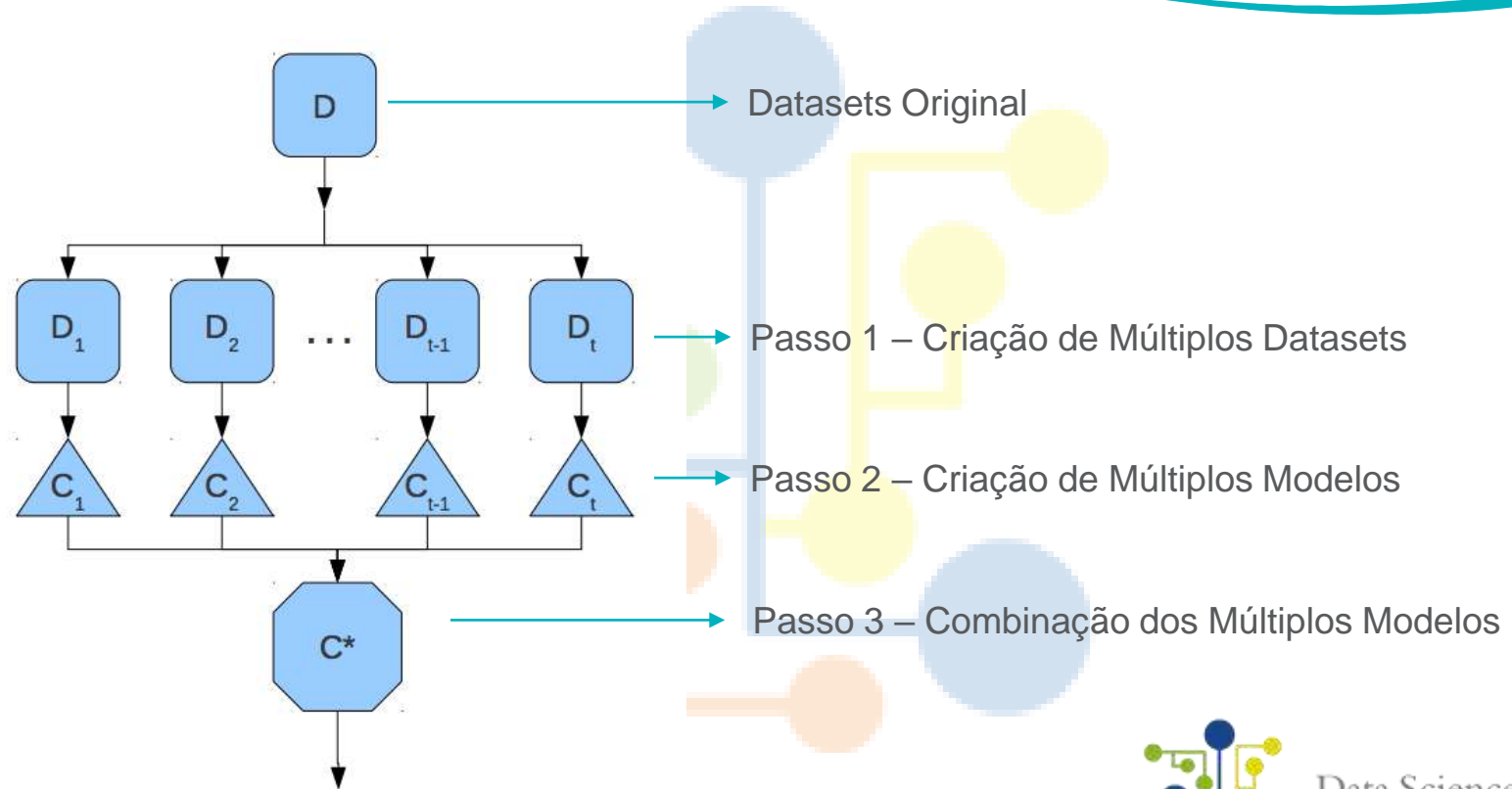


Resultado Final



Isso é Método Ensemble
Unimos as saídas de diferentes modelos para
encontrar a melhor resposta do problema





Estatísticas Simples

Ensemble Baseado em Modelos



Ensemble pode ser aplicado por você mesmo,
para combinar seus próprios modelos



Os algoritmos seguem duas abordagens principais para criar seu próprio ensemble

Bootstrap Aggregation ou Bagging

- Bagged CART
- Random Forest

Boosting

- C5.0
- Stochastic Gradient Boosting
- AdaBoost





Data Science Academy

Métodos Ensemble – Parte II



Data Science Academy

Combinação de Preditores



Bootstrap Aggregating
(Bagging)

Boosting

Voting



Bootstrap Aggregating (Bagging)

Para construção de múltiplos modelos (normalmente do mesmo tipo) a partir de diferentes subsets no dataset de treino.



Boosting

Para construção de múltiplos modelos (normalmente do mesmo tipo), onde cada modelo aprende a corrigir os erros gerados pelo modelo anterior, dentro da sequência de modelos criados.

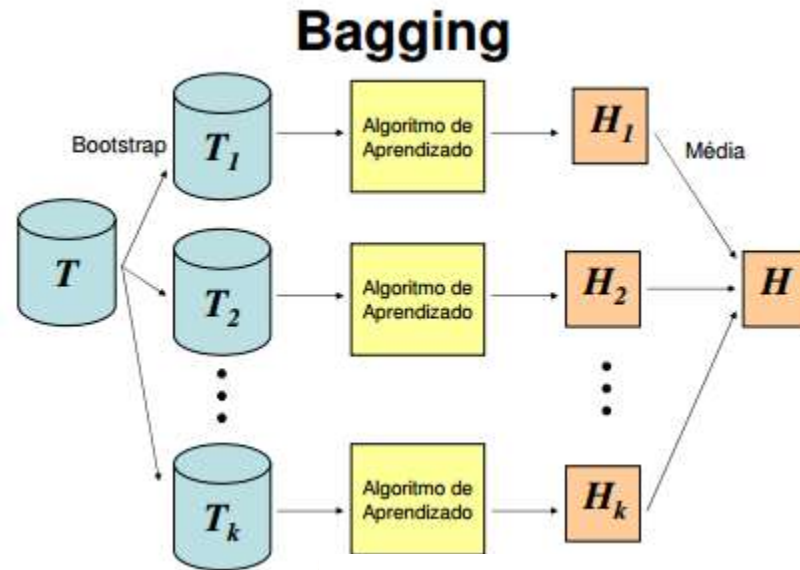


Voting

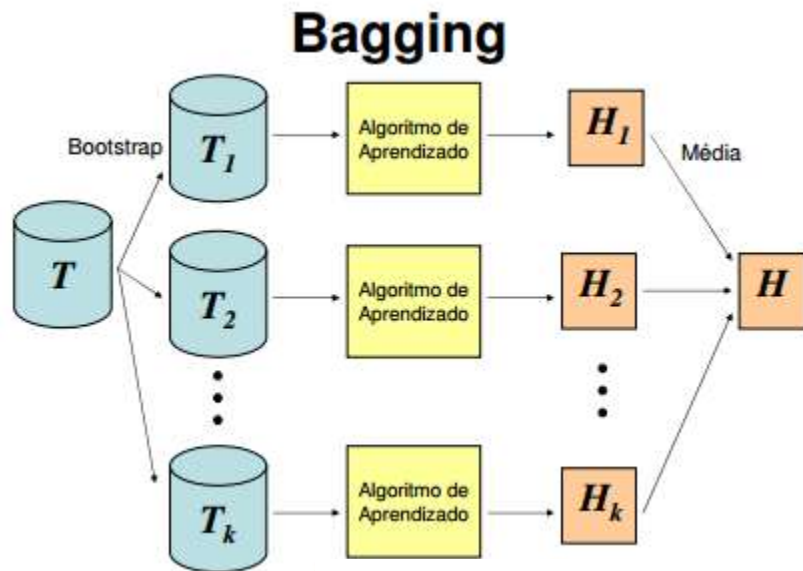
Para construção de múltiplos modelos (normalmente de tipos diferentes) e estatísticas simples (como a média) são usadas para combinar as previsões.



Bootstrap Aggregating (Bagging)



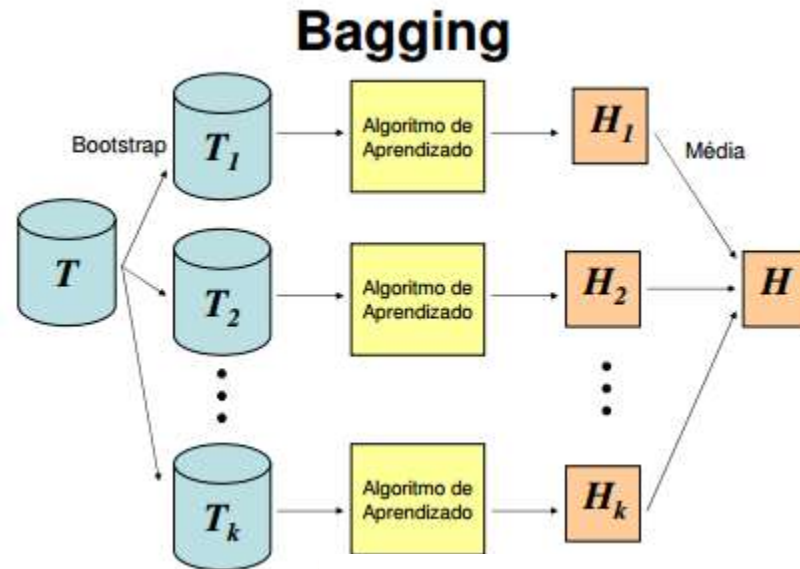
Bootstrap Aggregating (Bagging)



Bootstraps (amostras diferentes da base de dados que são usadas para aprender hipóteses diferentes)



Bootstrap Aggregating (Bagging)



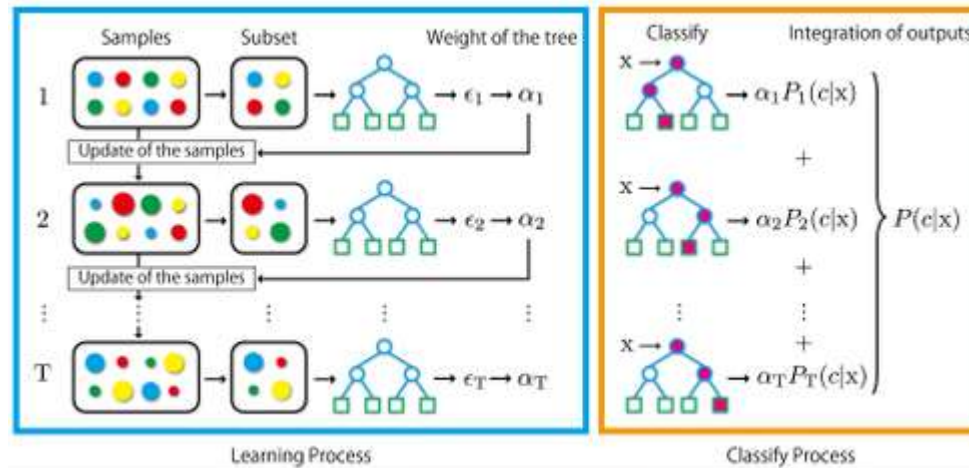
Boosting

- AdaBoost (**Adaptive Boosting**)
- Gradient Boosting
- XGBoost

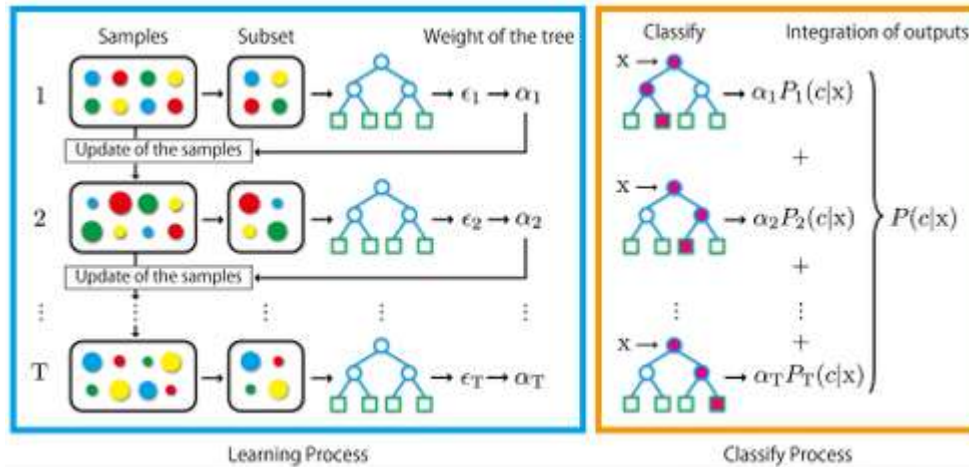
O termo "Boosting" refere-se a uma família de algoritmos que converte modelos fracos em um modelo forte



Boosting



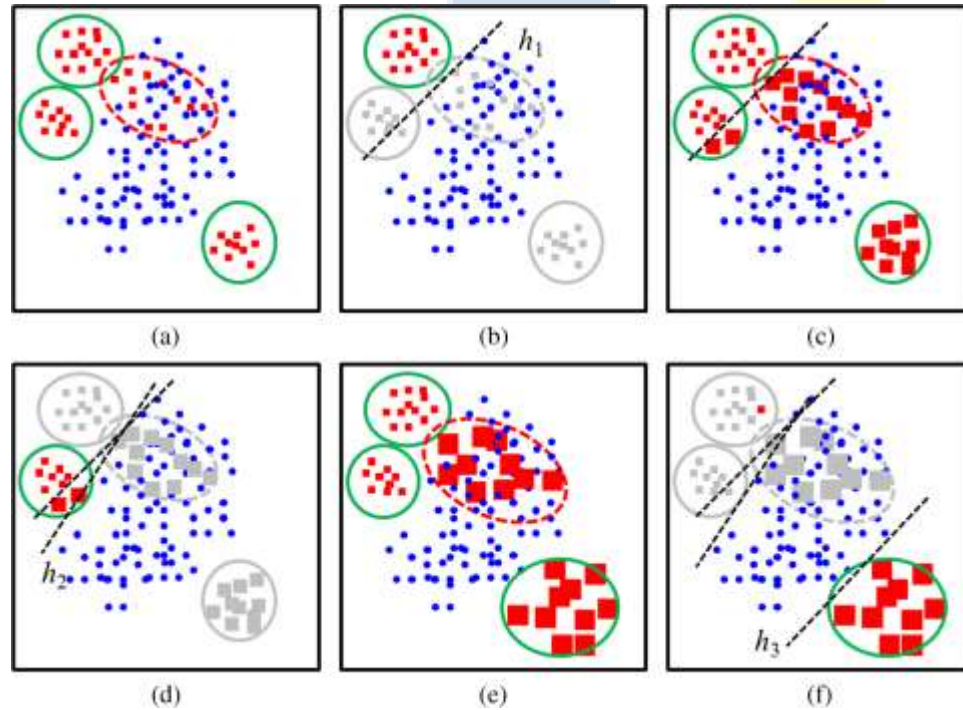
Boosting



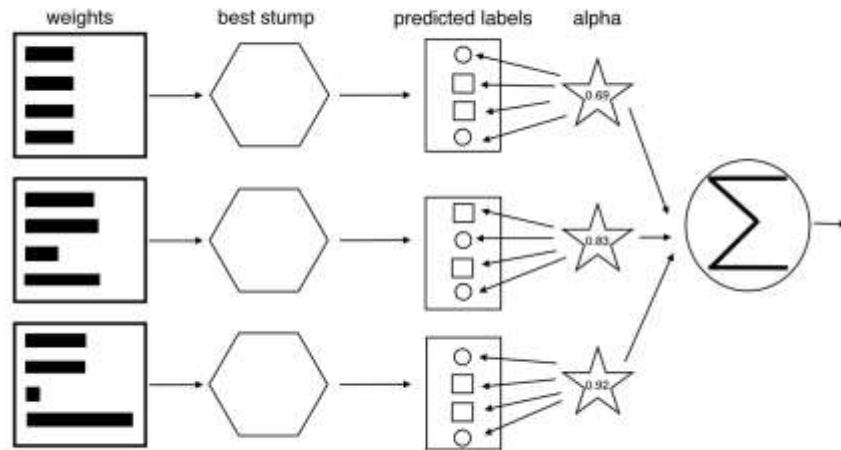
- AdaBoost.M1
- AdaBoost.M2
- AdaBoost.R
- Adaboost.R2
- AdaBoost.RT
- Boosting Correlation Improvement (BCI)



Adaboost



Adaboost

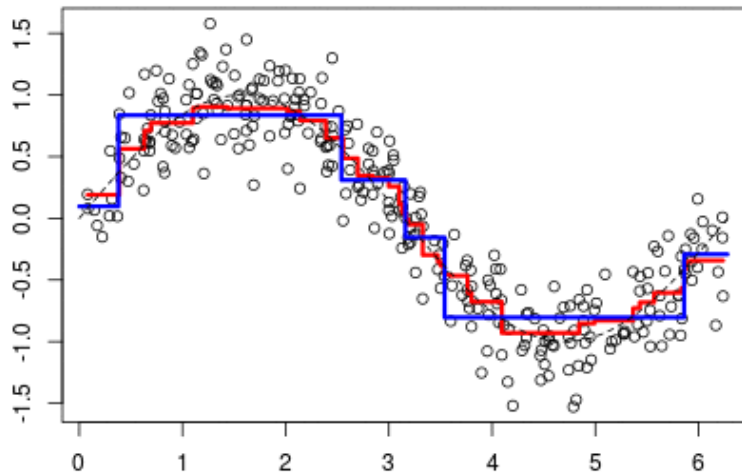


Entrada: $(x_1, y_1), \dots, (x_m, y_m)$

x = vetor de características
 $y = \{-1, +1\}$



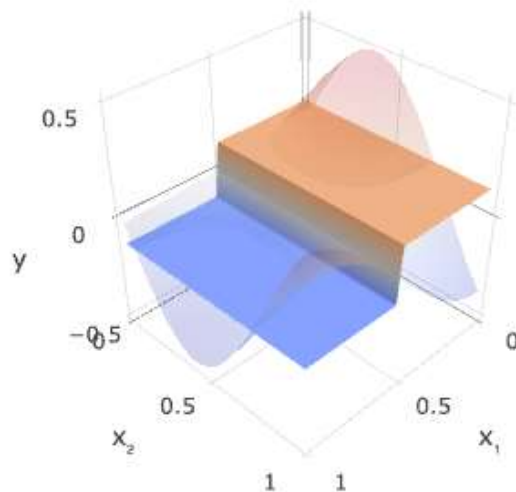
Boosting para Problemas de Regressão – AdaBoost.R

 $f(x)$
 $g(x)$

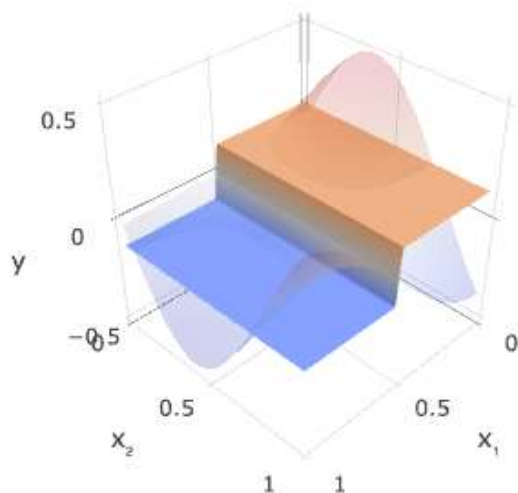
$$f(x_i) = g(x_i), \forall x_i \in X$$



Gradient Boosting = Gradient Descent + Boosting
GBM = Gradient Boosting Method



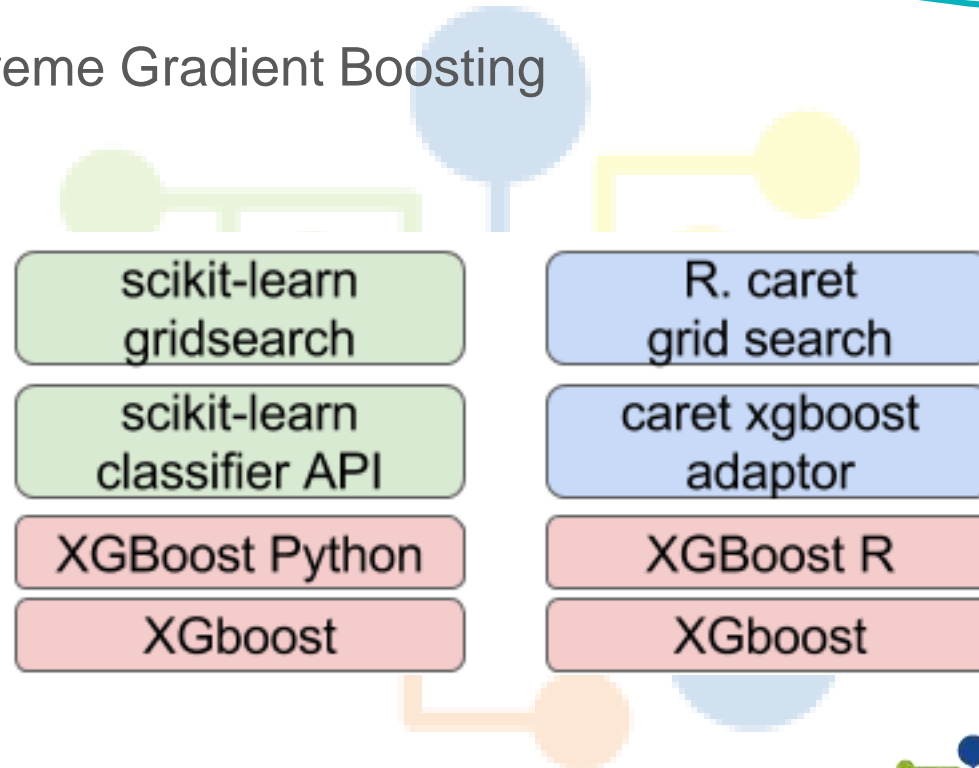
Gradient Boosting = Gradient Descent + Boosting
GBM = Gradient Boosting Method



Função de Perda:
 $y = ax + b + e$



XGBoost – eXtreme Gradient Boosting



Por que utilizar Métodos Ensemble?



Por que utilizar Métodos Ensemble?

Razões Estatísticas

Grandes volumes de dados

Pequenos volumes de dados



Por que utilizar Métodos Ensemble?

Razões Estatísticas

Dividir e Conquistar

Grandes volumes de dados

Seleção de modelo

Pequenos volumes de dados

Diversidade





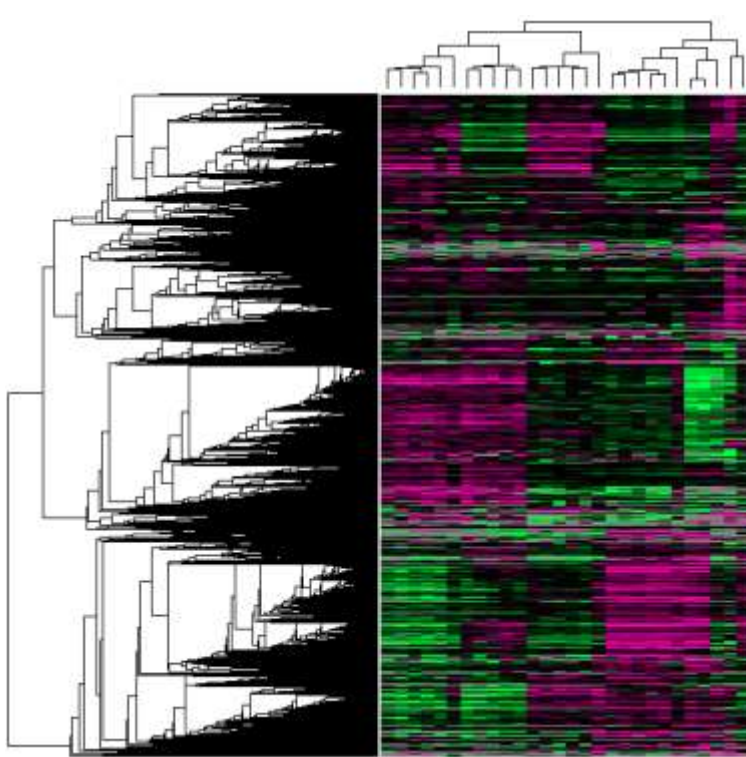
Data Science Academy

Redução de Dimensionalidade



Data Science Academy



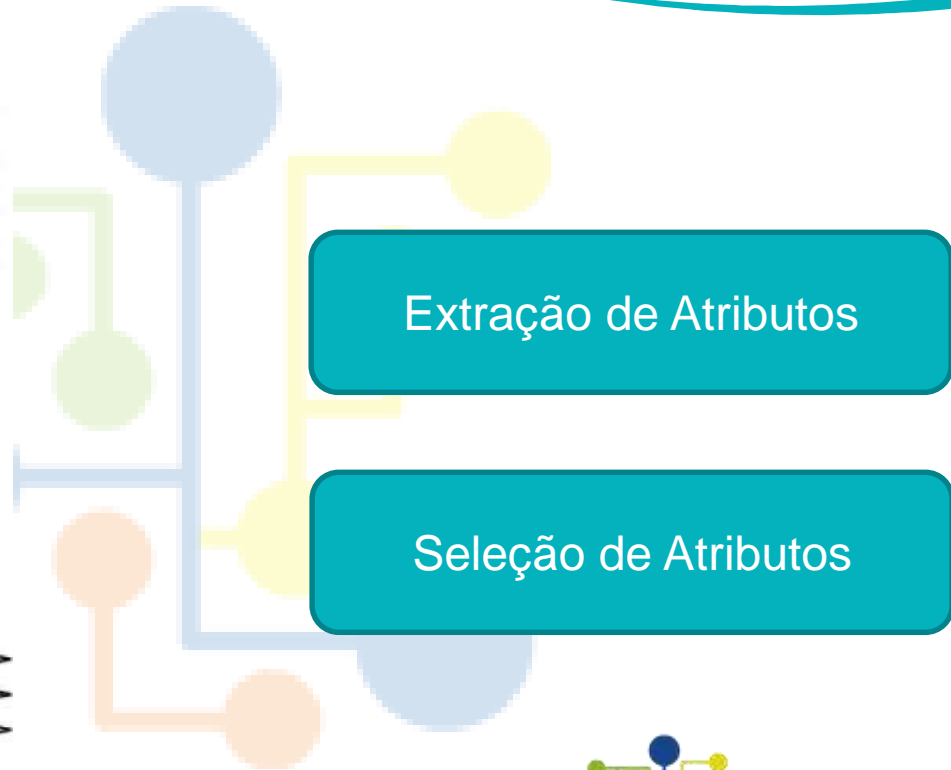
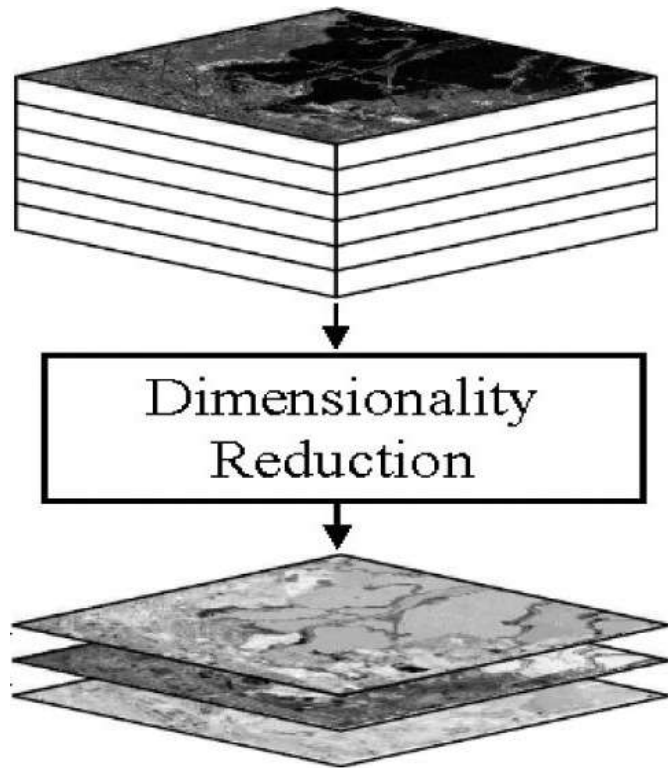


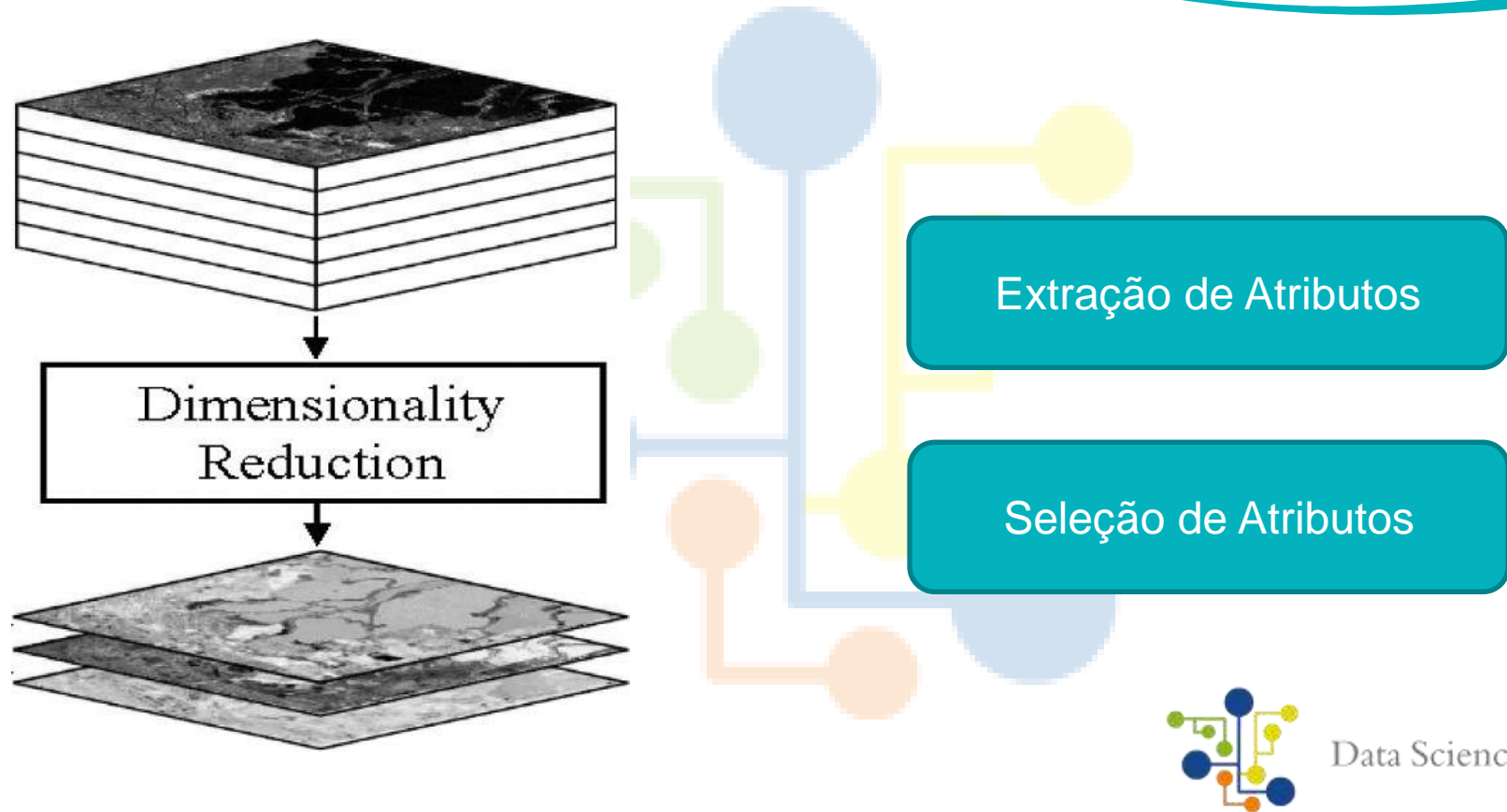
E esse aumento no volume de dados não é apenas em volume e de forma vertical, mas ocorre também na horizontal, com o aumento do número de dimensões ou atributos das bases de dados



O termo ***dimensionalidade*** é atribuído ao número de características de uma representação de padrões, ou seja, a dimensão do espaço de características









Extração de Atributos

Principal Component Analysis, Multidimensional Scaling e o FastMap.

Seleção de Atributos

Algoritmos de aprendizado de máquina, cálculo de dimensão fractal e wrapper.



7 Técnicas para Redução da Dimensionalidade

Missing Values Ratio

Low Variance Filter

High Correlation Filter

**Random Forests /
Ensemble Trees**

**Forward Feature
Construction**

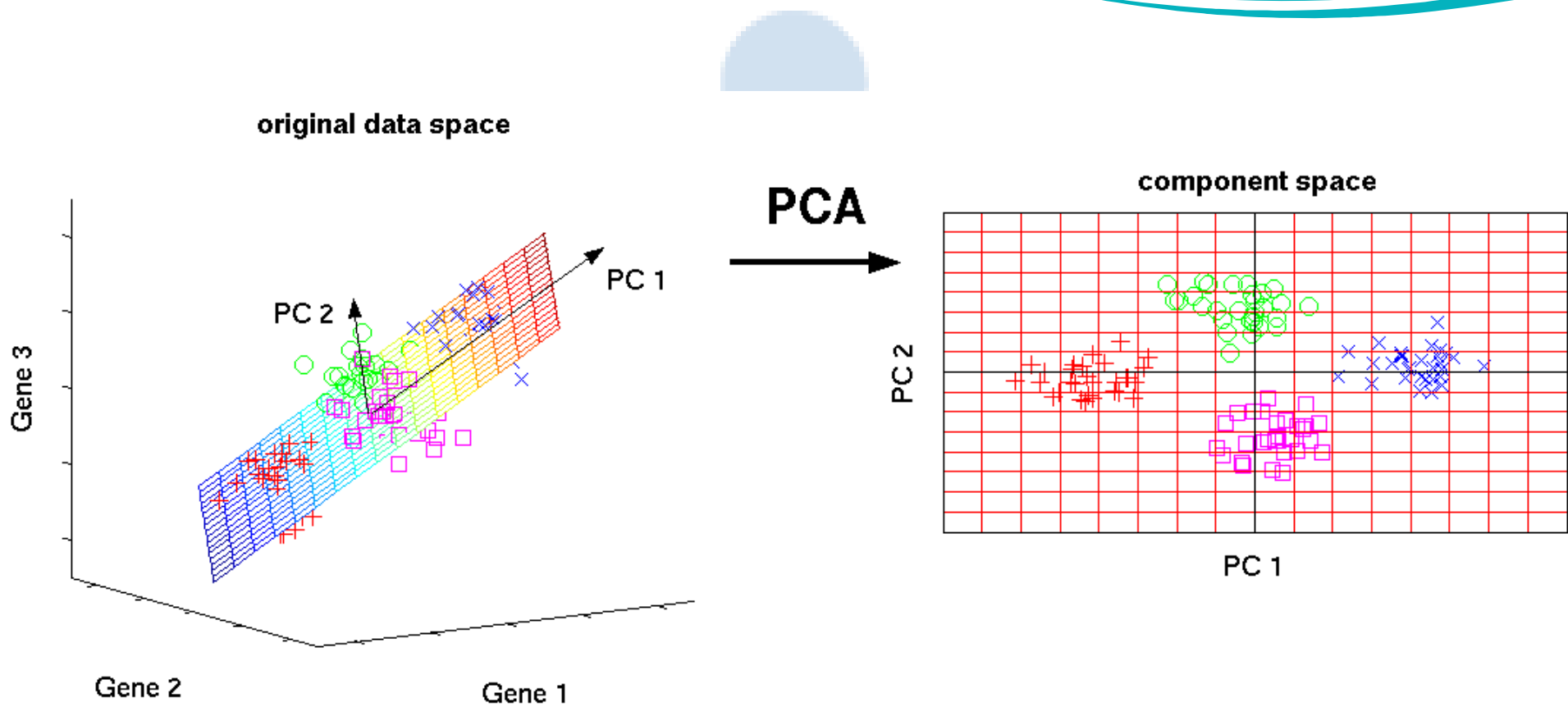
**Backward Feature
Elimination**

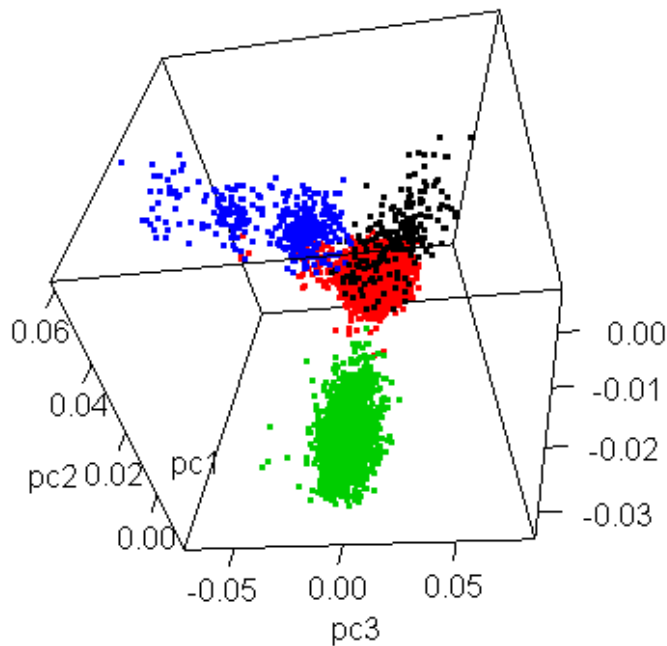
**Principal Component
Analysis (PCA)**



Principal Component Analysis (PCA)



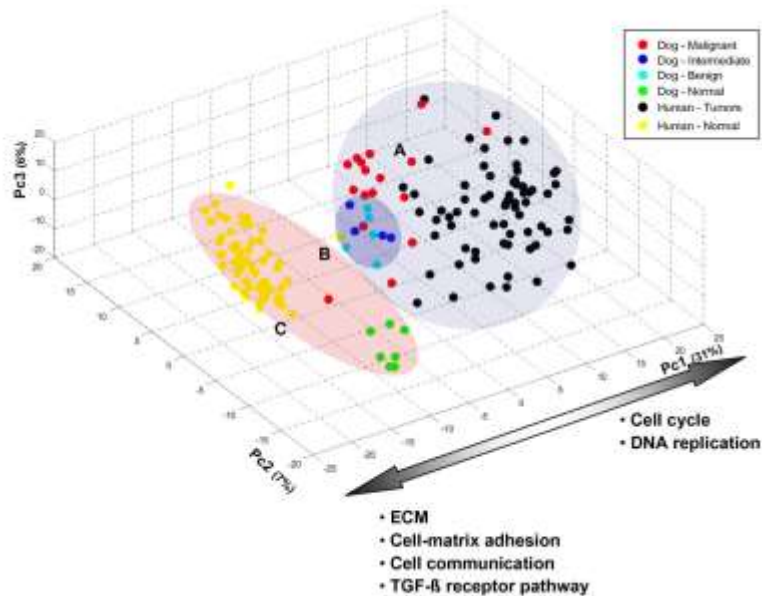




Cada componente resultante é uma combinação linear de n atributos.

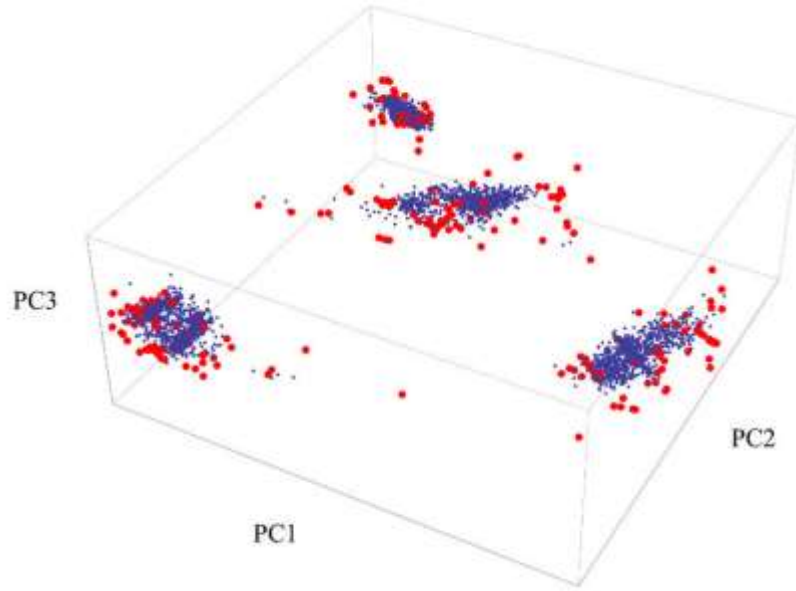
Cada componente principal é uma combinação de atributos presentes no dataset





O PCA precisa ser alimentado com dados normalizados. Utilizar o PCA em dados não normalizados pode gerar resultados inesperados.



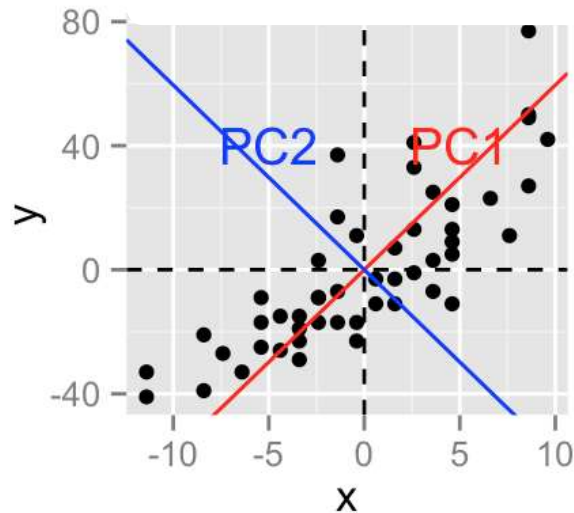


Esta técnica pode ser utilizada para geração de índices e agrupamento de indivíduos



A análise de componentes principais é associada à ideia de redução de massa de dados, com menor perda possível da informação.





Em termos gerais a PCA busca reduzir o número de dimensões de um dataset, projetando os dados em um novo plano.





Data Science Academy

Modelos Lógicos, Geométricos e Probabilísticos



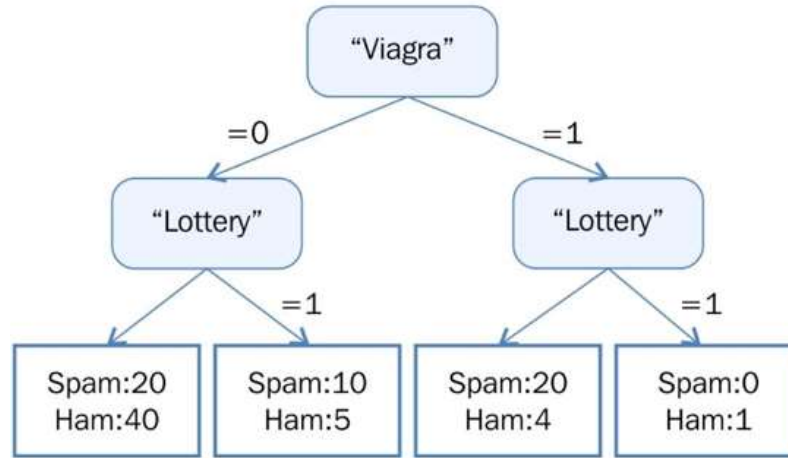
Data Science Academy

Lógicos

Geométricos

Probabilísticos

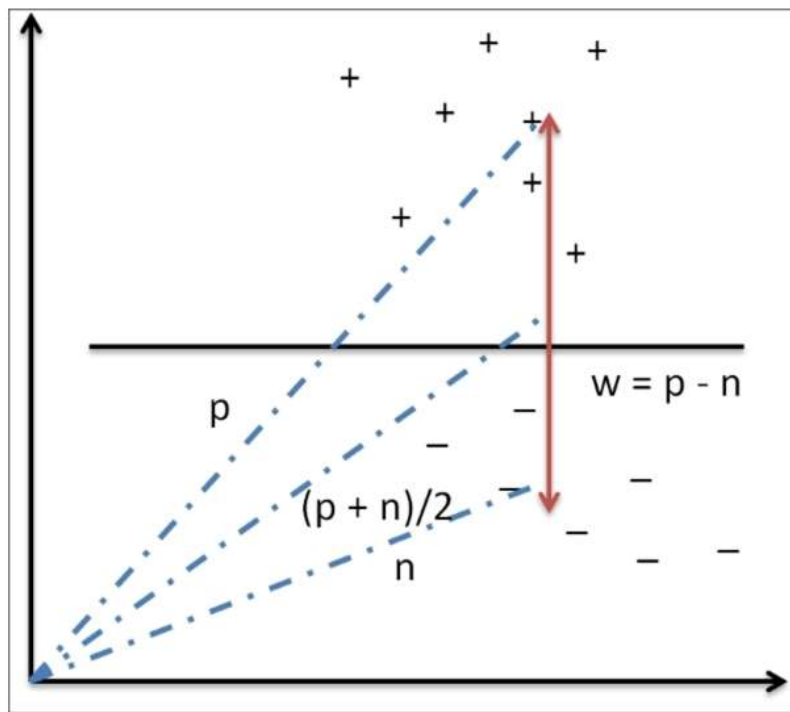




<div>"Lottery"</div> <div>=1</div>	Spam:10 Ham:5	Spam:0 Ham:1
	Spam:20 Ham:40	Spam:20 Ham:4
<div>=0</div>		<div>"Viagra"</div> <div>=1</div>

Lógicos





Geométricos



Viagra	Lottery	$P(Y = \text{Spam} (\text{Viagra}, \text{lottery}))$	$P(Y = \text{ham} (\text{Viagra}, \text{lottery}))$
0	0	0.31	0.69
0	1	0.65	0.35
1	0	0.80	0.20
1	1	0.40	0.60

Probabilísticos







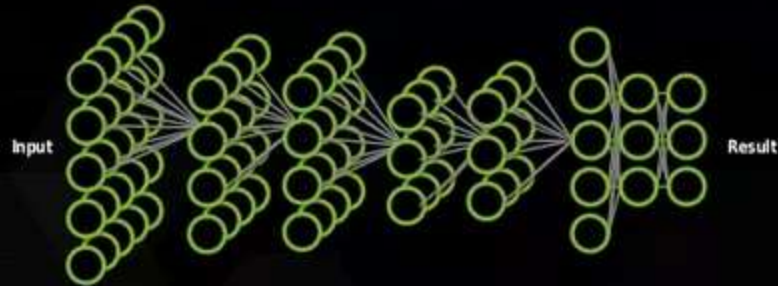
Data Science Academy

Deep Learning



Data Science Academy

WHAT MAKES DEEP LEARNING DEEP?



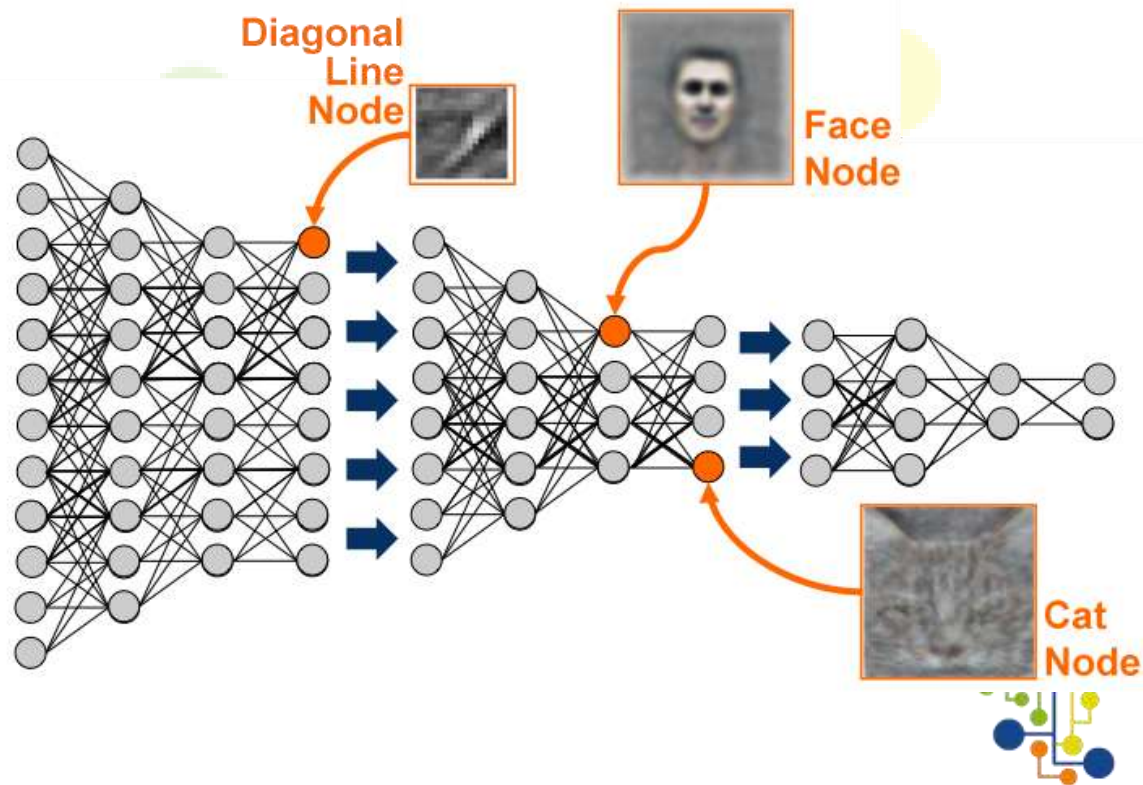
Today's Largest Networks

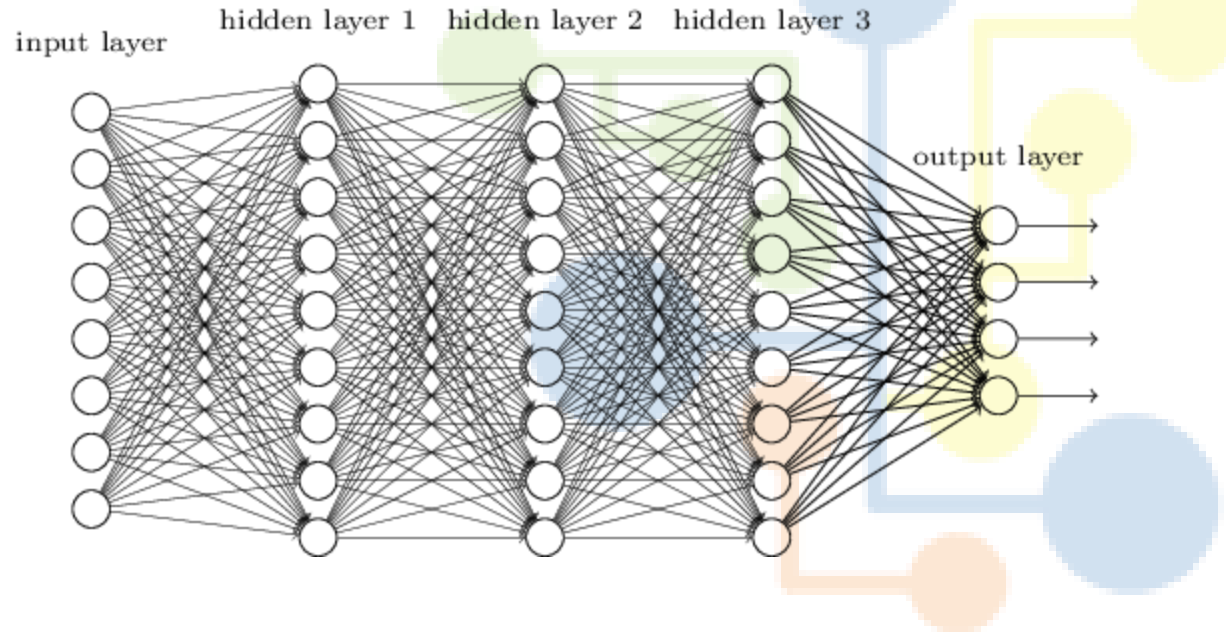
- ~10 layers
- 1B parameters
- 10M images
- ~30 Exaflops
- ~30 GPU days

Human brain has trillions of parameters - only 1,000 more.

by NVIDIA

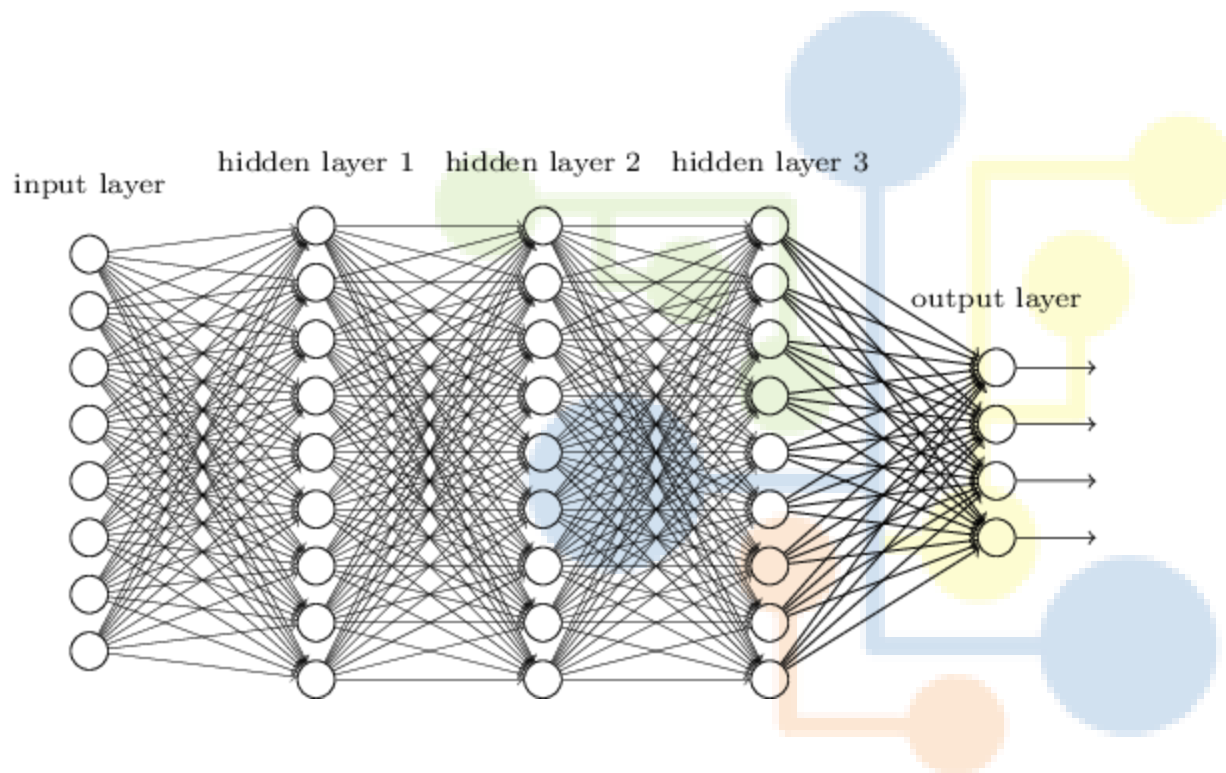






CNN
Convolutional Neural
Networks



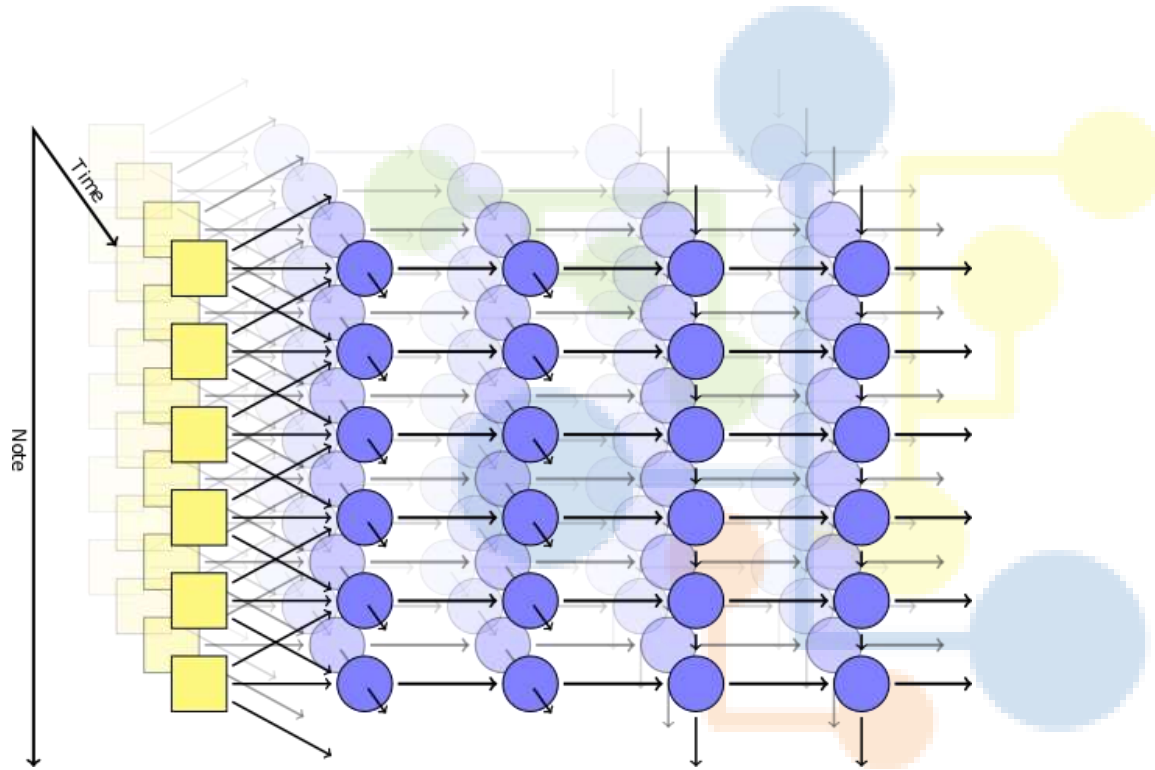


O nome "Rede Neural Convolucional" indica que a rede aplica uma operação matemática denominada convolução.

Convolução é um tipo especializado de operação linear.

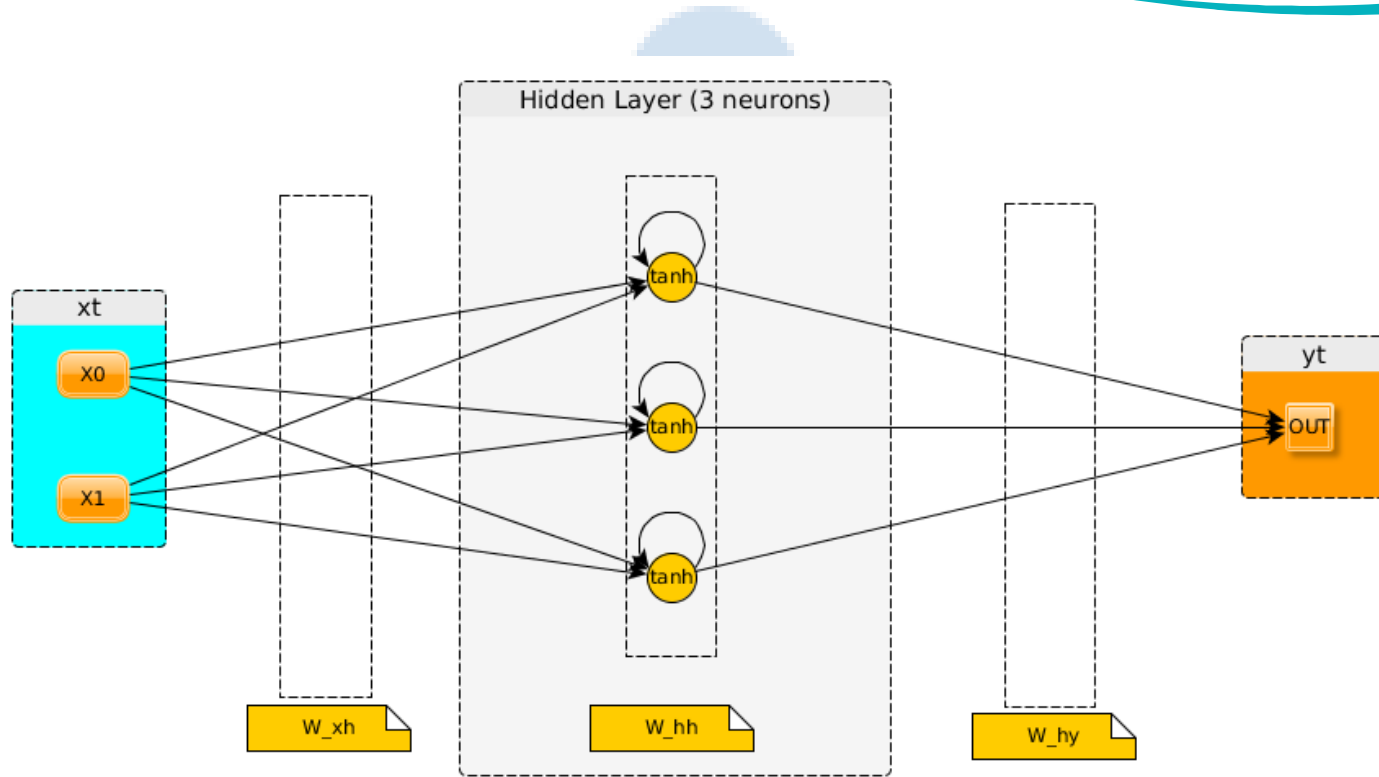
Redes convolucionais são redes neurais simples que utilizam convolução no lugar de multiplicação geral de matrizes em pelo menos uma das camadas.

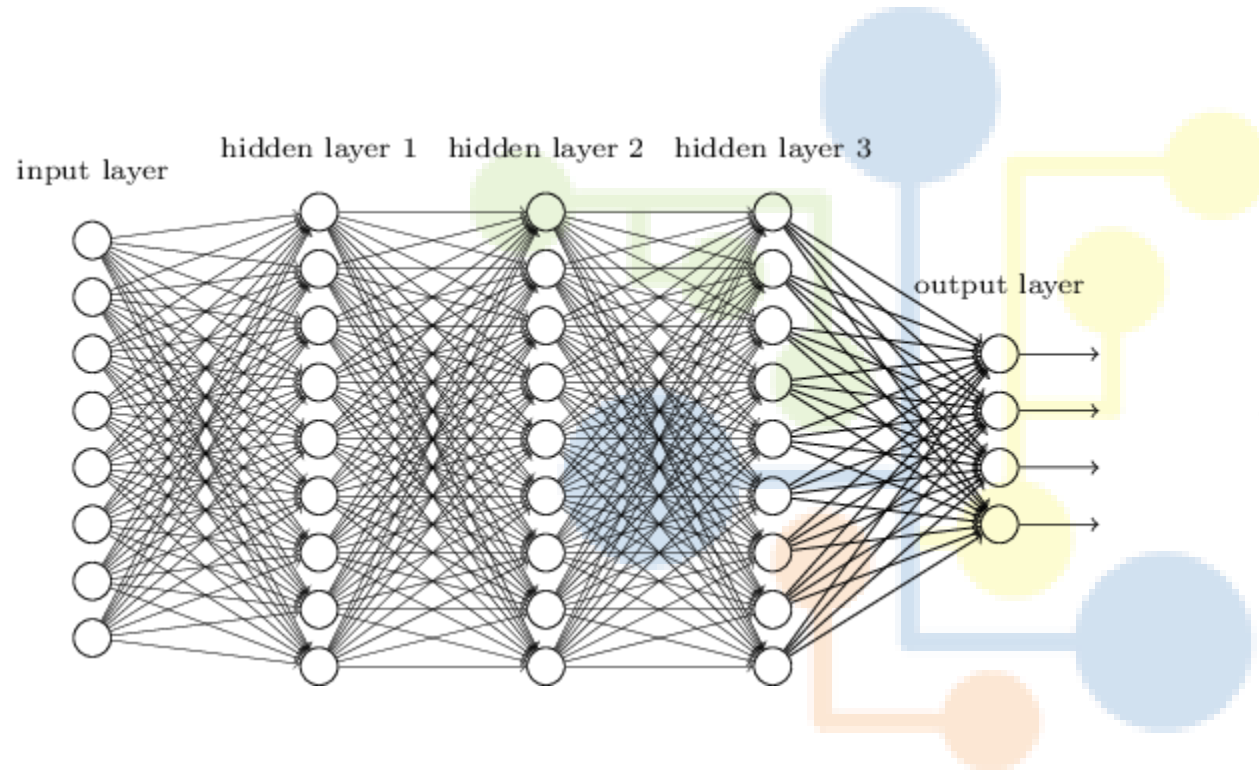




RNN Recurrent Neural Networks







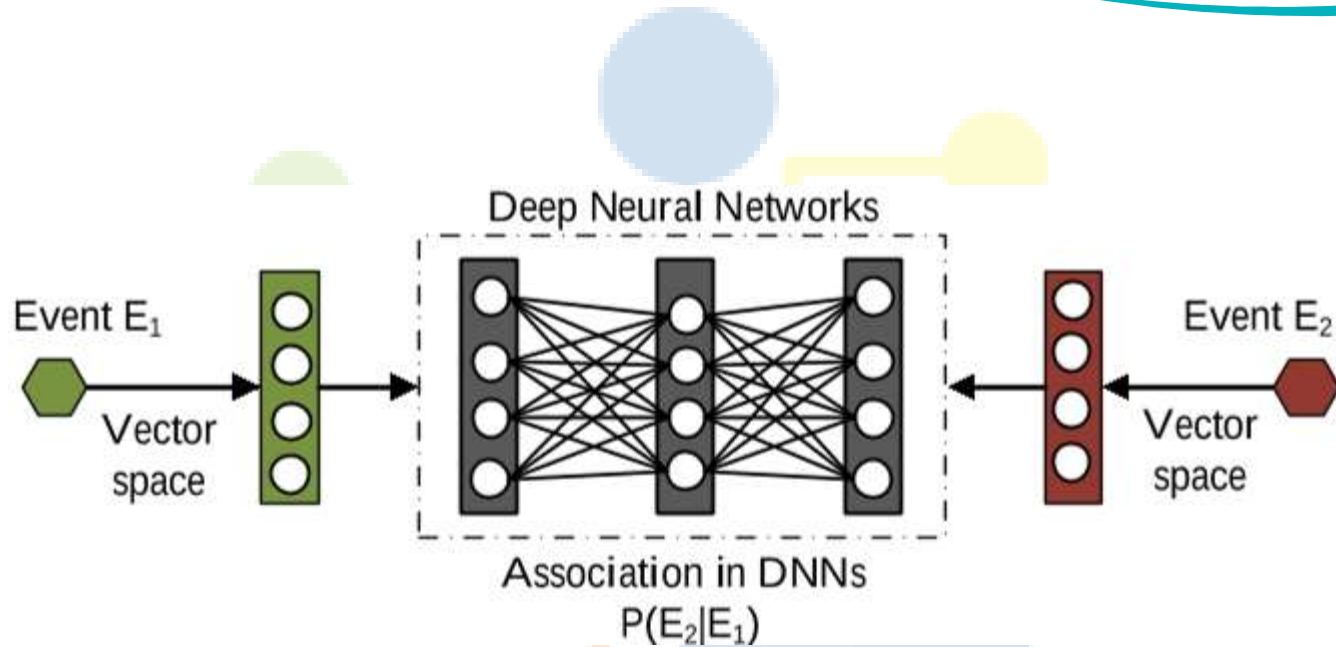
**Deep Learning é a
técnica de
aprendizado baseada
em Multi-Camadas
(Redes Neurais
Densas)**

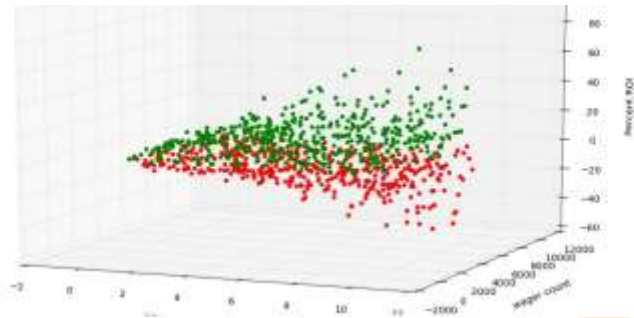




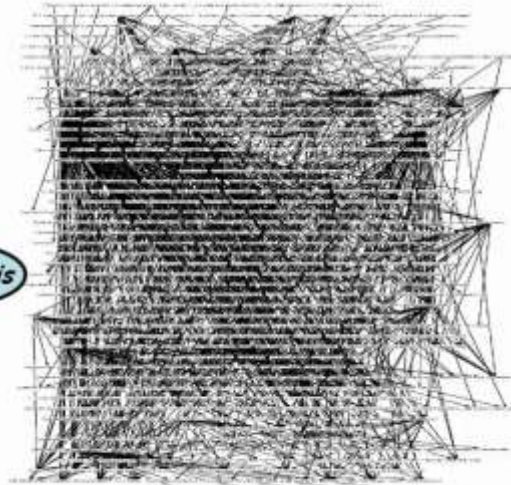
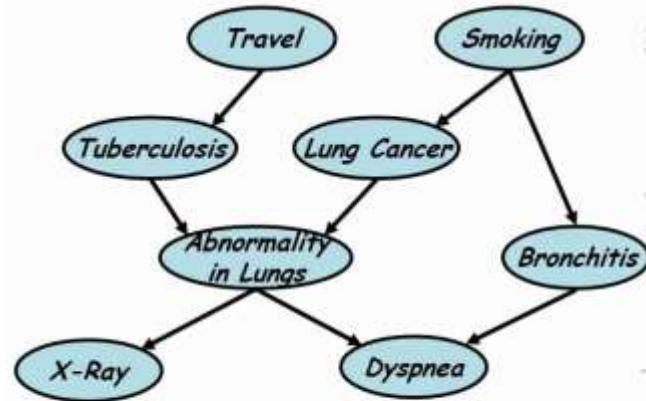
Reconhecimento de Voz e Processamento de Linguagem Natural





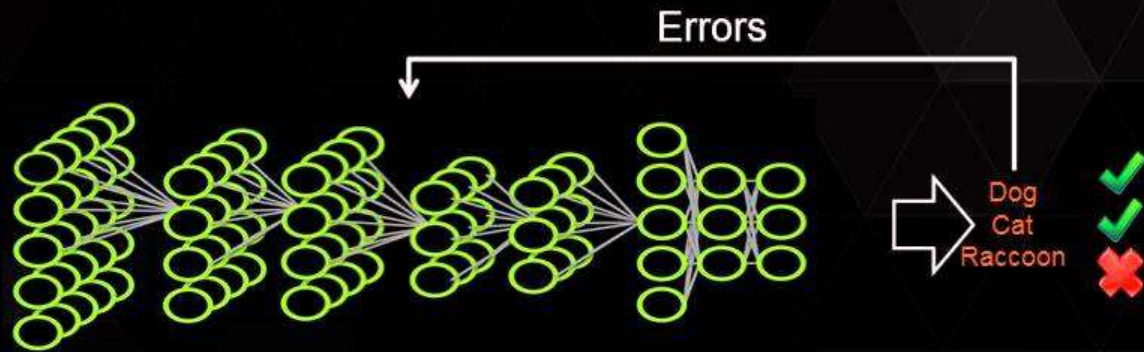


Muitos modelos probabilísticos são difíceis de treinar devido a dificuldade de realizar inferência

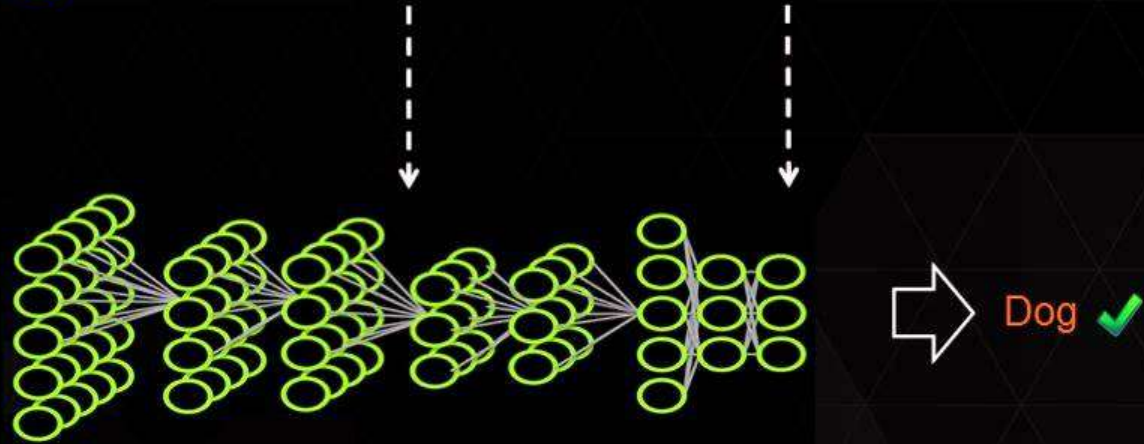


DEEP LEARNING APPROACH

Train:



Deploy:



Deep Learning está revolucionando a indústria aeroespacial





Data Science Academy

Simulação e Otimização



Data Science Academy





Podemos entender a simulação como um processo amplo que engloba não apenas a construção do modelo, mas todo o método experimental que se segue



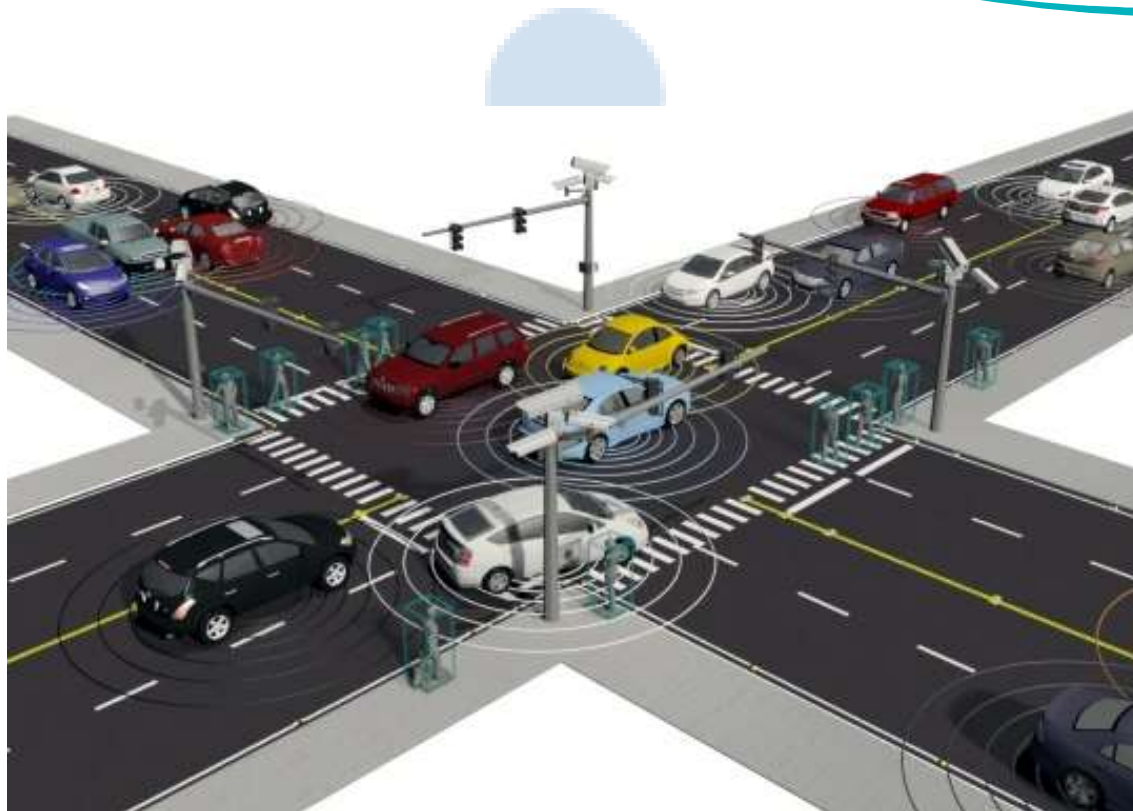
Simulação implica na modelagem de um processo ou sistema, de tal forma que o modelo imite as respostas do sistema real em uma sucessão de eventos que ocorrem ao longo do tempo





- Um dispositivo para compreensão de um problema;
- Um meio de comunicação para descrever a operação de um sistema;
- Uma ferramenta de análise para determinar elementos críticos e estimar medidas de desempenho;
- Uma ferramenta de projeto para avaliar problemas e propor soluções;
- Um sistema de planejamento de operações para trabalhos, tarefas e recursos;
- Um mecanismo de controle;
- Uma ferramenta de treinamento;
- Uma parte do sistema para fornecer informações on-line, projeções de situações e suporte à decisão.







Variáveis





Variáveis de
Estado





Entidade





Atributo





Recurso





Processes





Tempo de
Simulação





Filas





Eventos





A simulação é a técnica empregada durante a fase de construção do modelo preditivo





Pode ser necessário a simulação de uma grande variedade de alternativas e a criação de modelos preditivos pode gerar bons resultados



Machine Learning nos ajuda a simular um determinado evento através de dados e com isso, prever o comportamento futuro.





Modelos Determinísticos
x
Modelos Estocásticos



Modelos Determinísticos



Modelos Estocásticos

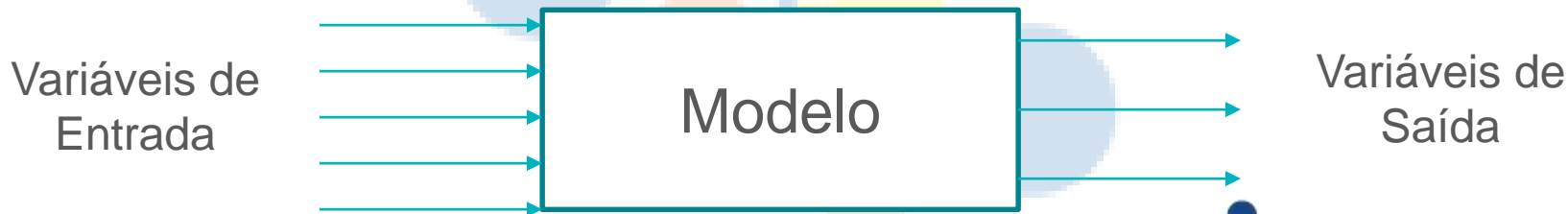


Quando uma variável de entrada de um sistema é aleatória, a variável de saída também será aleatória, no entanto, o sistema pode ter comportamento determinístico ou ser representado por um modelo determinístico





Modelo Computacional é um programa de computador cujas variáveis apresentam o mesmo comportamento dinâmico e estocástico do sistema real que representa



Velocidades
individuais
observadas em
uma avenida

Modelo

Comportamentos
Normais ou Anormais
e Tendências



Determinístico

Resultado do modelo é pré-determinado em função dos dados de entrada

Estocástico

Resultado do modelo não depende somente dos dados de entrada, mas também de outros fatores, normalmente aleatórios. Isso requer um modelo probabilístico.



Determinístico

Exemplo: Se uma pessoa tem mais de 16 anos, ela pode tirar carteira de motorista. Se tiver menos de 16, não pode.

Estocástico

Exemplo: Modelo para prever a reação de pessoas em um shopping, a uma situação de emergência. Um modelo probabilístico tenta descrever o comportamento "aleatório" das entidades



Modelos determinísticos e estocásticos podem ser combinados para resolver problemas que requerem muitas alternativas diferentes, tais como:

- Busca na Web e Extração de Informação
- Desenvolvimento de Novos Medicamentos
- Prever o Comportamento do Mercado Financeiro
- Compreender o Comportamento de Clientes
- Criação de Robôs

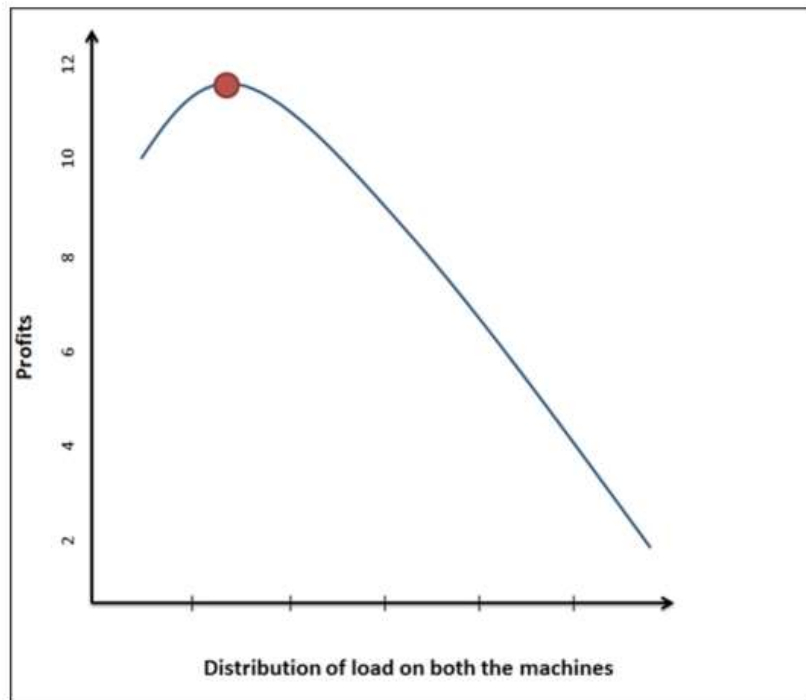




Otimização







Otimização



Aprendizado = Representação + Avaliação + Otimização



Aprendizado =

Representação

+

Avaliação

+

Otimização



Ao contrário da simulação onde existe incerteza associada com os dados de entrada, na otimização nós temos não somente acesso aos dados, mas também as informações de dependências e relacionamentos entre os atributos dos dados.



Generalização → Principal Objetivo na Construção do Modelo Preditivo



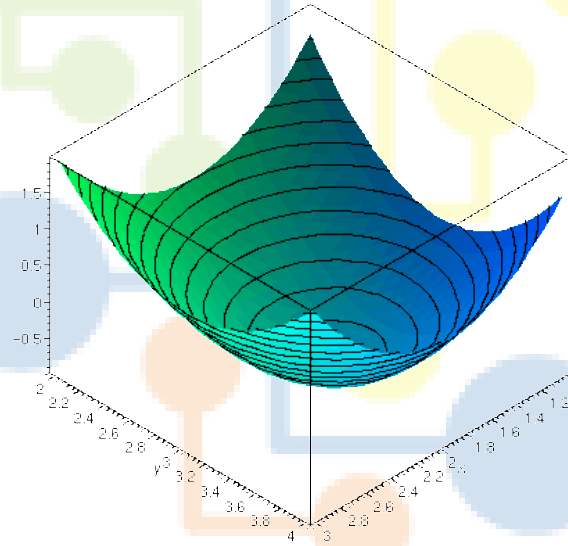
Seleção de Modelo



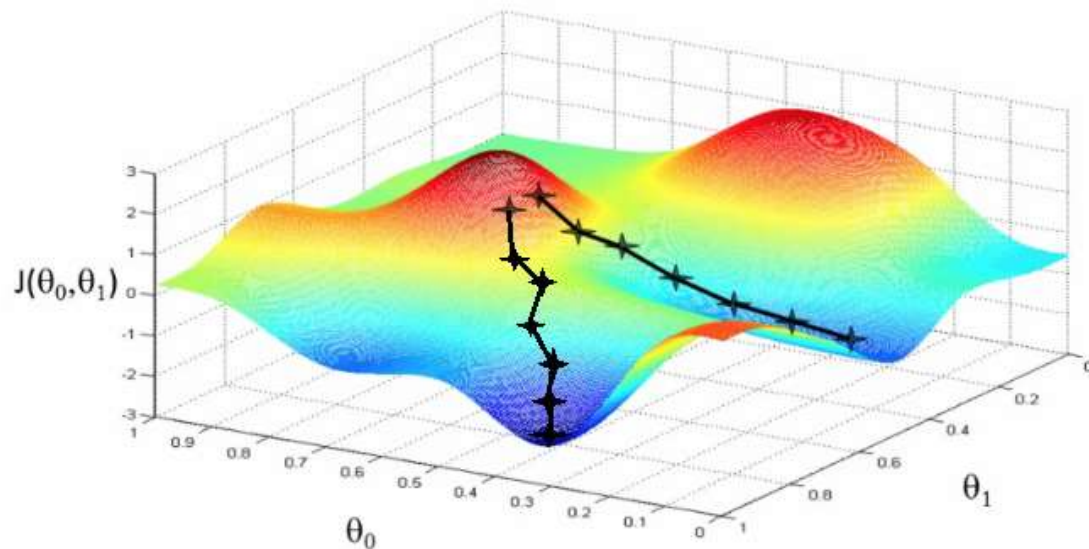
Definir Espaço de Parâmetro	Definir Configurações de Validação Cruzada
Definir Métrica	Treinar, Avaliar e Comparar



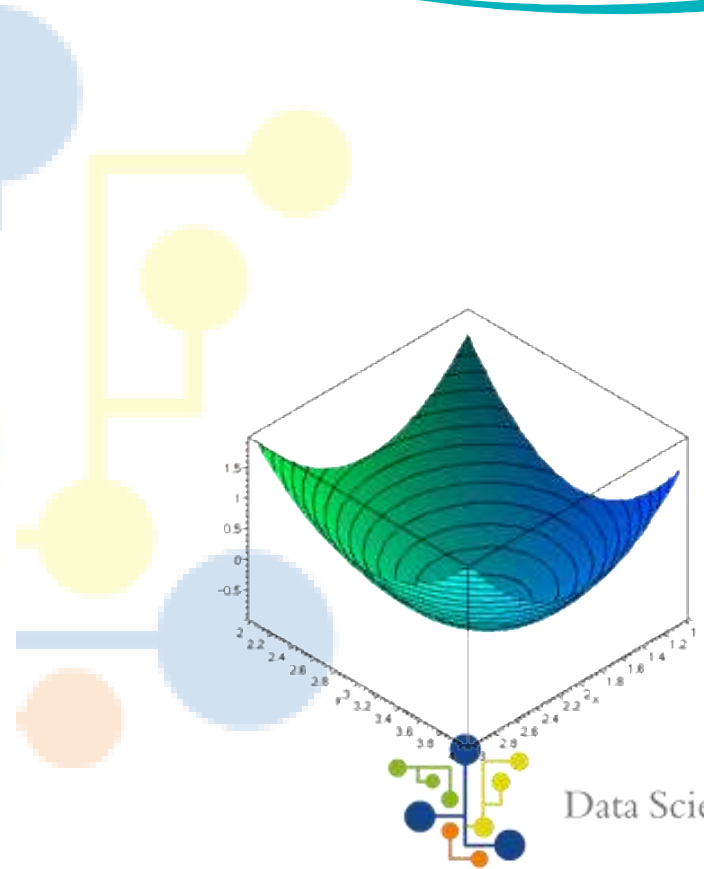
Gradiente Descendente (Gradient Descent)



Gradiente Descendente (Gradient Descent)



Gradiente Descendente (Gradient Descent)



Gradiente Descendente (Gradient Descent)



coefficient = 0.0
cost = $f(\text{coefficient})$
 $\text{delta} = \text{derivative}(\text{cost})$
coefficient = coefficient - ($\alpha * \text{delta}$)



Gradiente Descendente (Gradient Descent)

Batch Gradient Descent

Stochastic Gradient Descent





Data Science Academy
Obrigado