Faculdade IMPACTA TECNOLOGIA

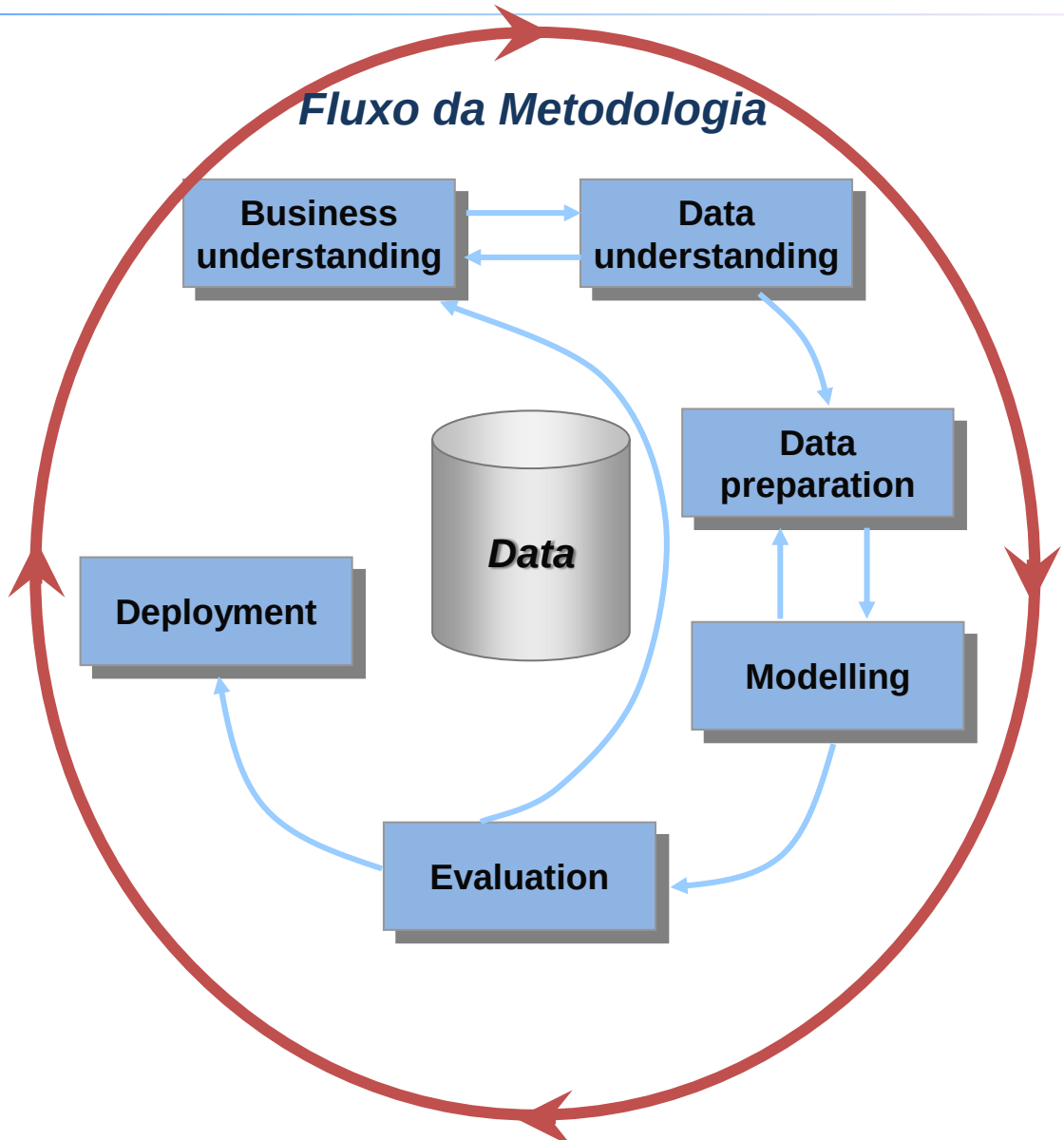
Graduação em BD e SI

Computação Cognitiva

RESUMO CRISP-DM
REVISÃO TÉCNICAS
DEEP LEARNING

Prof. Roberto Santos

CRISP-DM



CRISP-DM
CRoss Industry Standard Process for Data Mining

1 - Entendimento do Negócio

Procura identificar os objetivos e as necessidades na perspectiva de negócio, e assim transformar este conhecimento em uma tarefa de Data Mining. Na determinação dos objetivos do negócio, o primeiro passo é identificar as necessidades do cliente. O especialista em Data Mining também deverá identificar fatores importantes que poderão influenciar os resultados e os critérios de sucessos a serem avaliados.

2 - Entendimento dos Dados

Identificação da informação e dados relevantes para o estudo e uma primeira familiarização com o seu conteúdo, descrição, qualidade e utilidade. Nesta etapa objetiva-se também adquirir a informação com a qual se irá trabalhar, listando as suas fontes, o procedimento de leitura e os problemas preliminares detectados. A descrição dos dados deve descrever a forma como foram adquiridos, listando o seu formato, volume, significado e toda a informação relevante neste contexto.

Faculdade IMPACTA TECNOLOGIA

Entendimento do Negócio

- **Qual o contexto da aplicação?**
  - Qual área?
  - Qual empresa?
  - Qual setor?
- **Qual o problema de negócio a ser resolvido?**
- **Qual ou quais tarefas devemos utilizar?**
- **Quais os critérios de sucesso do projeto?**
  - Métricas claras de sucesso (para o negócio);
- **Descrição da solução.**

Faculdade IMPACTA TECNOLOGIA

Entendimento dos Dados

- **Estágio de familiarização com os dados do problema;**
- **Descrições dos dados disponíveis para modelagem;**
- **Identificação da qualidade;**
- **Primeiras indicações do que os dados podem oferecer;**
- **Quantidades de dados;**
- **Características dos dados;**
- **Primeiras percepções se os objetivos estabelecidos na fase anterior pode ser atingidos com os dados disponíveis.**

3 - Preparação dos dados

Conjunto de atividades destinadas a obter os dados finais, a partir do qual será criado e validado o modelo. A seleção dos dados e a escolha dos atributos são partes desta fase. Outra fase é a integração que representa a junção de dados provenientes de várias tabelas, para criar uma visão única (em geral apenas uma tabela), onde está toda a informação necessária para a análise.

4 - Modelagem

Nesta etapa são aplicadas as técnicas de Data Mining mais apropriadas dependendo das tarefas e objetivos levantados no Entendimento do Negócio. A criação de modelos representa a fase principal em um projeto de Data Mining, na qual técnicas de modelagem são aplicadas em um conjunto de dados.

Preparação dos Dados

Na prática é:

- Remover dados ou inutilidades se necessário;
- Decidir quais transformações devem ser feitas para tratar:
  - Preenchimento Errado;
  - Preenchimento Incompleto;
  - Falta de preenchimento (valores nulos);
  - Inconsistências;
  - Ausência de Semântica;
- Retirar amostra, se preciso;
- Efetuar as transformações necessárias para adequação dos dados às técnicas que serão utilizadas na modelagem;

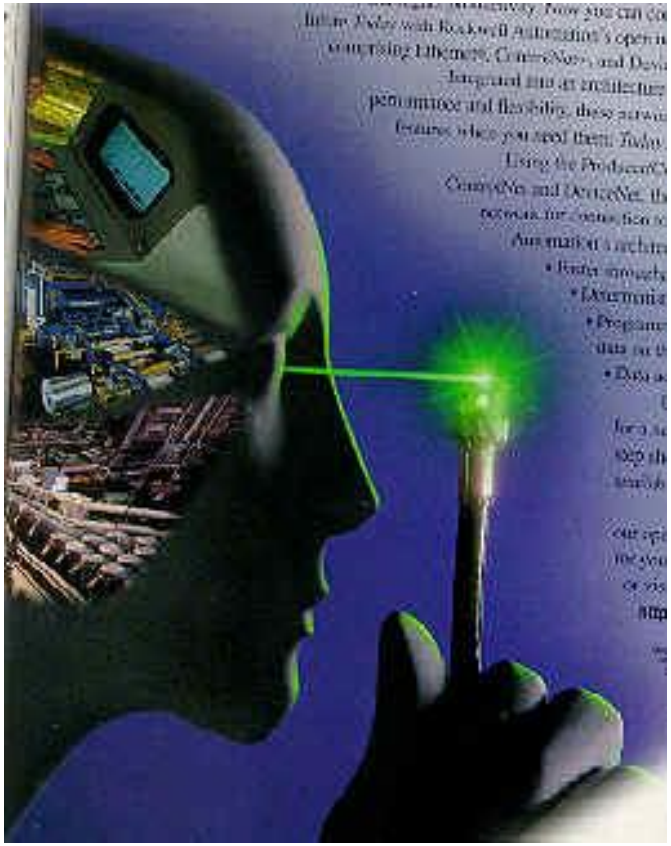
5 - Avaliação de Desempenho

Esta etapa consiste na avaliação do modelo, revendo os passos seguidos anteriormente e verificando se os objetivos de negócio foram alcançados.

6 - Distribuição (Implantação)

Conjunto de ações a fim de aplicar os resultados do projeto na organização. A distribuição dos resultados pode envolver a elaboração de relatórios, implantação de modelos desenvolvidos, desenvolvimento de software para aplicação dos resultados, entre outros. Dependendo do projeto, pode haver a necessidade de especificação de monitoramento e atualização periódica dos modelos.

Faculdade IMPACTA TECNOLOGIA



Técnicas de Data Mining

Faculdade IMPACTA TECNOLOGIA

10 Importantes Técnicas de Data Mining

Esta seria a minha ordem:

1. Logistic regression;
2. ANN - Artificial Neural Networks;
3. Classification and Regression Tree (C4.5, CART, Chaid, ID3);
4. Linear regression;
5. SVMs – Support Vector Machine;
6. k-means;
7. k-Nearest Neighbors (kNN);
8. Naive Bayes;
9. Bayesian Networks;
10. Genetic Algorithm.

Faculdade IMPACTA TECNOLOGIA

Algumas Técnicas de Data Mining

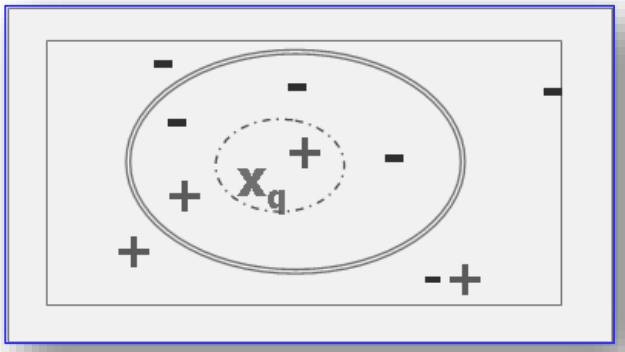
Métodos baseados em instâncias (KNN)

Algoritmo que considera todos os padrões (instâncias) como pontos no espaço *n*-dimensional R^n

Os vizinhos mais próximos de um padrão são definidos em termos da distância, por exemplo, Euclidiana.

Seja um padrão *x* arbitrário descrito pelo vetor de características $a_1(x)$, $a_2(x)$,...,$a_n(x)$, em que $a_r(x)$ representa o valor do r-ésimo atributo de *x*, então a distância euclidiana entre x_i e x_j.

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$



k = 1 classifica xq como **+**

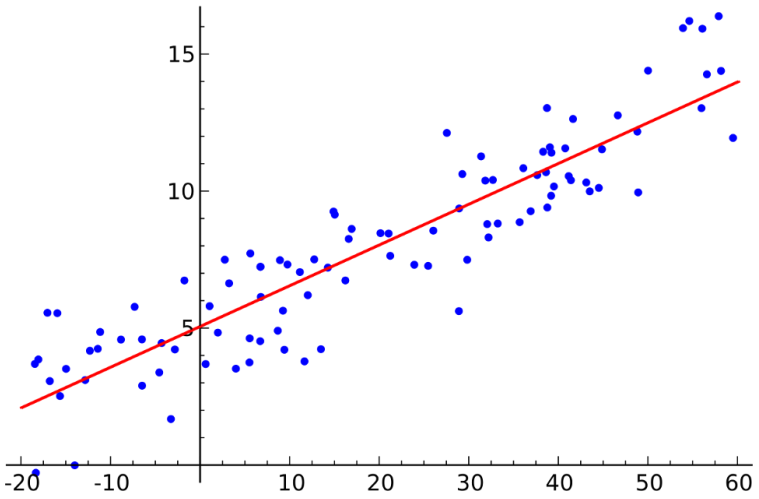
k = 5 classifica xq como **-**

Regressão Linear

Análise Multivariada

Obtenção de uma equação (modelo) que tenta explicar a variação da variável dependente (alvo) pelas variação das demais variáveis independentes (entradas).

$$y_c = b_0 + b_1 x_1 + b_2 x_2 + \ldots b_n x_n + u_i$$

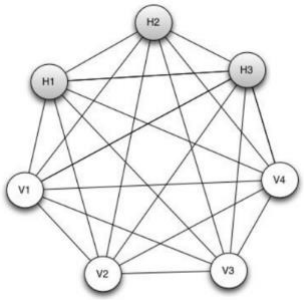


Faculdade IMPACTA TECNOLOGIA

Regressão linear

Exemplo

Modelos com Fórmulas Matemáticas



$$Y_i=\beta_0+\beta_1 X_{i1} + \beta_2 X_{i2} +\ldots+ \beta_p X_{ip} + \varepsilon_i$$

Regressões Lineares Múltiplas

Score de Propensão - Pagar Dívida

	Peso		*Variável*	
Y =	50	+		
	10	*	FLAG_IDADE_MAIOR_25	+
	-5	*	FLAG_IDADE_MENOR_26	+
	10	*	RENDA_MAIOR_5000	+
	-5	*	RENDA_MENOR_5001	+
	30	*	FLAG_SEM_RESTRICAO_SPC	+
	-40	*	FLAG_COM_RESTRICAO_SPC	

Exemplo: *Pessoa de 20 anos com renda > 5000 e sem restrições no SPC tem* ***Score = 85***

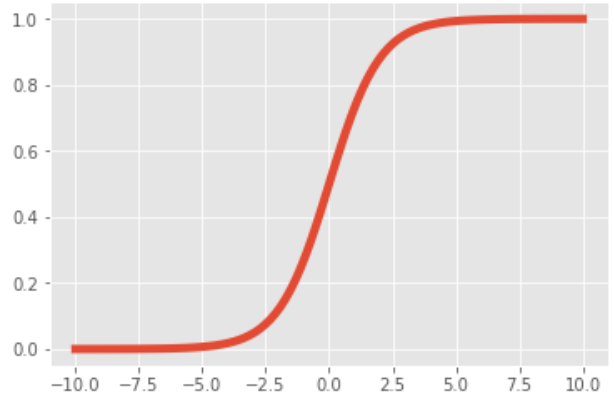
Regressão Logística

Alvo Binário

Semelhante a uma regressão linear, mas a sua variável resposta (alvo) é binária e a equação final é apresentada abaixo.

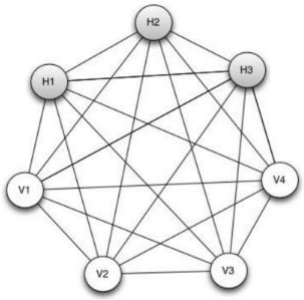
$$y_c = b_0 + b_1 x_1 + b_2 x_2 + \ldots b_n x_n$$

$$y = \frac{1}{1 + e^{-y_c}}$$ ou y = 1/(1 + EXP(-(y_c)))



Evolução: Regressão Logística
Alvo Binário

Modelos com Fórmulas Matemáticas

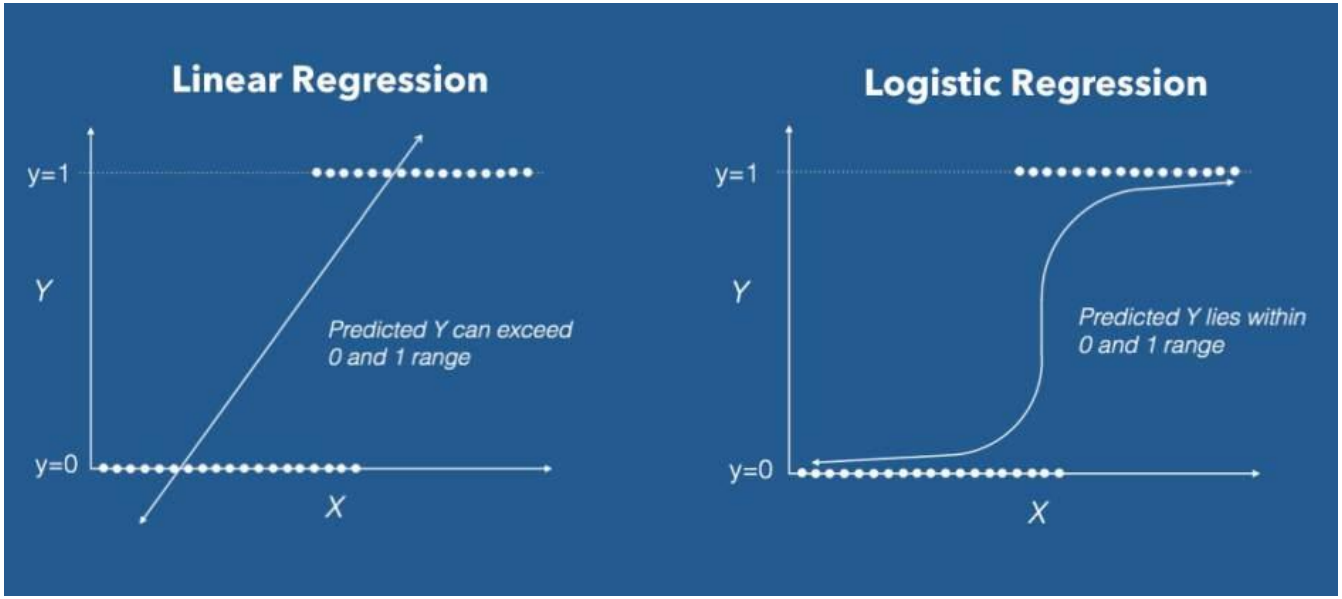


$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

Regressões Lineares Múltiplas

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 z_{1,i} + \cdots + \beta_k z_{k,i})}}$$

Regressões Logísticas

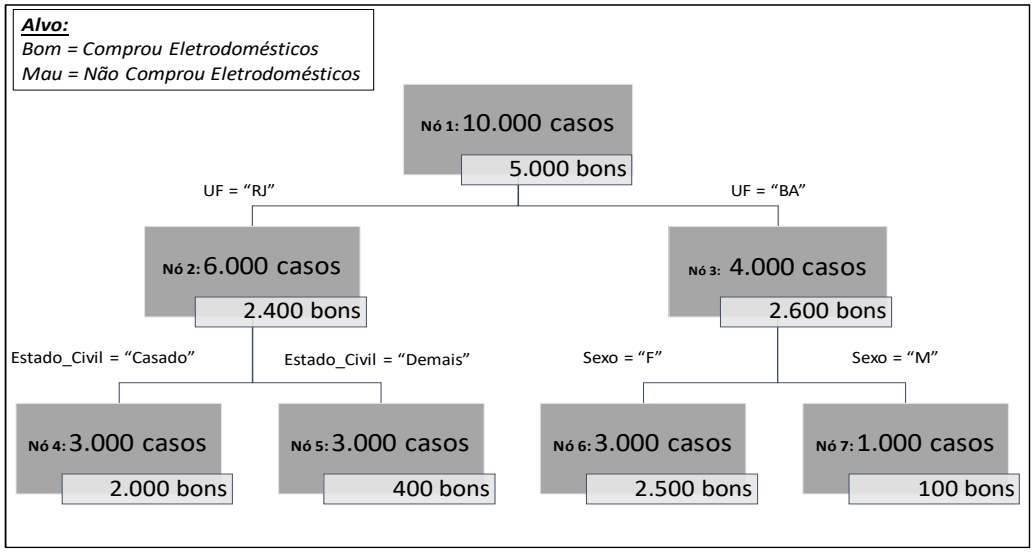


Faculdade IMPACTA TECNOLOGIA

Árvore de Decisão

Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore.

Algoritmos podem elaborar uma árvore de decisão para prever as melhores escolhas de forma matemática.

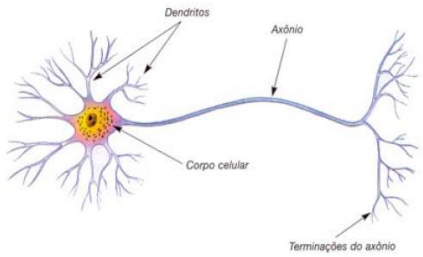


Redes Neurais Artificiais

Técnica inspirada no funcionamento do cérebro, onde neurônios artificiais, conectados em rede, são capazes de aprender e de generalizar.

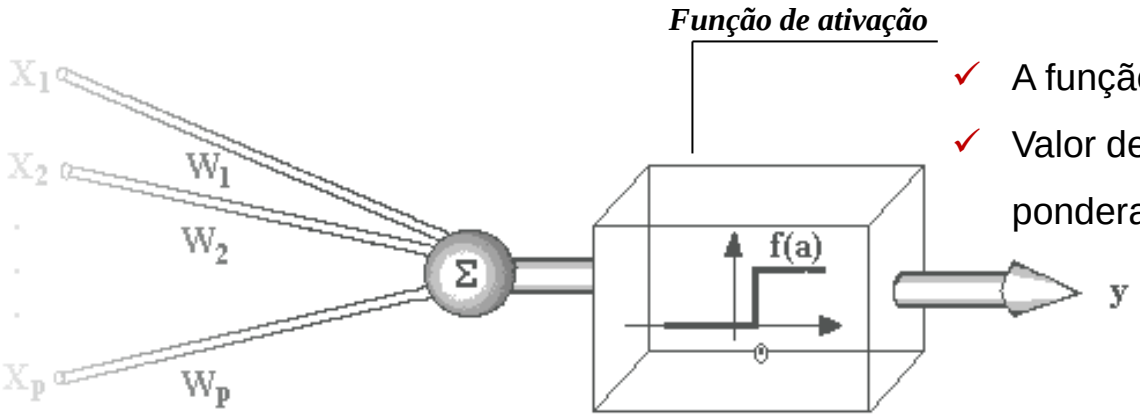
- **Supervisionado -** Utilizado quando se conhece a resposta do problema a ser solucionado (base com alvo).
  - **Ex.: MLP com Backpropagation;**

- **Não Supervisionado -** Utilizado quando **não** se conhece a resposta do problema a ser solucionado, normalmente para clusterização.
  - **Ex.: Mapa de Kohonen;**



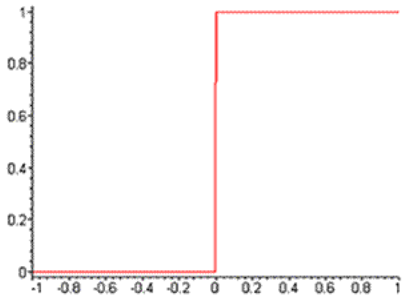
Redes Neurais Artificiais

Função de Ativação



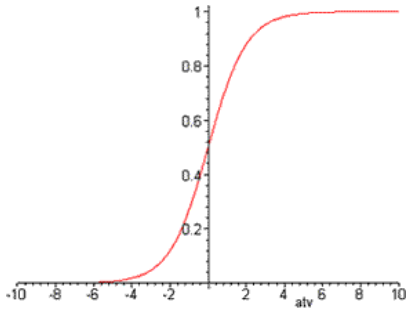
- ✓ A função de ativação *ativa* o neurônio
- ✓ Valor de ativação (soma das entradas ponderadas): $a = v = \sum x_i \cdot w_i$

- ➢ Função limiar (degrau):



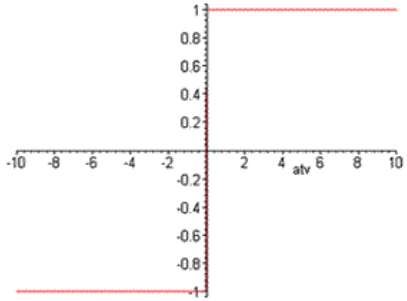
$$f(v) = \begin{cases} 1, \text{se } v \geq 0 \\ 0, \text{se } v < 0 \end{cases}$$

- ➢ Função sigmóide:



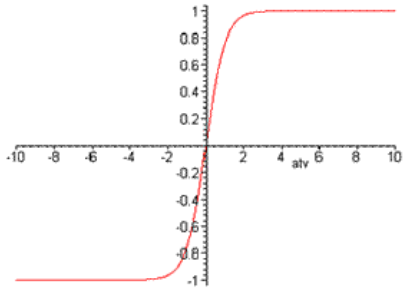
$$f(v) = \frac{1}{1 + e^{-av}}$$

- ➢ Função signum:



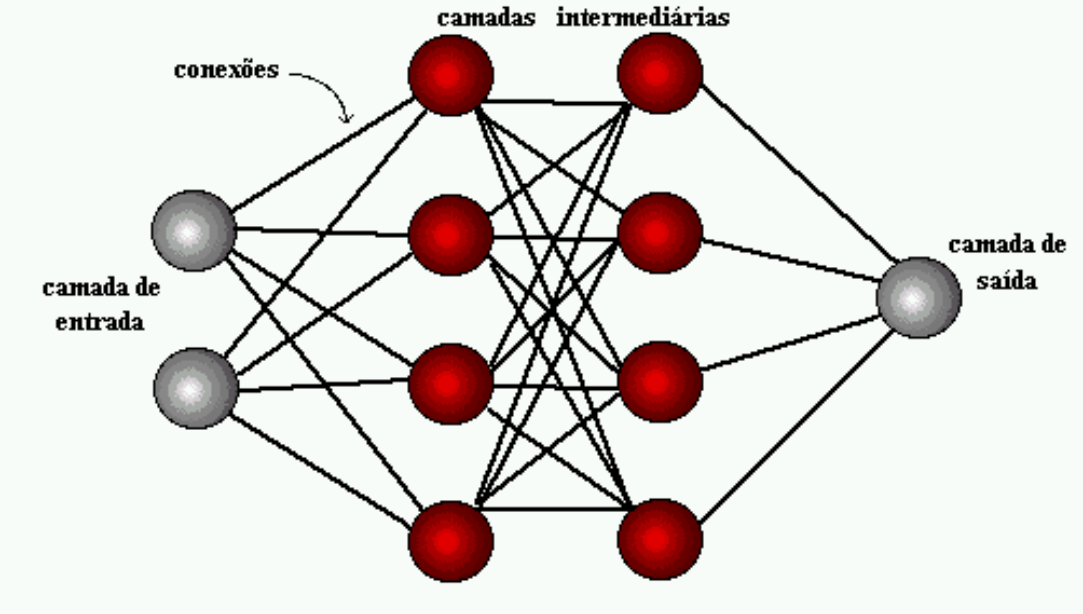
$$f(v) = b\frac{v}{|v|}, \text{para } v \neq 0$$

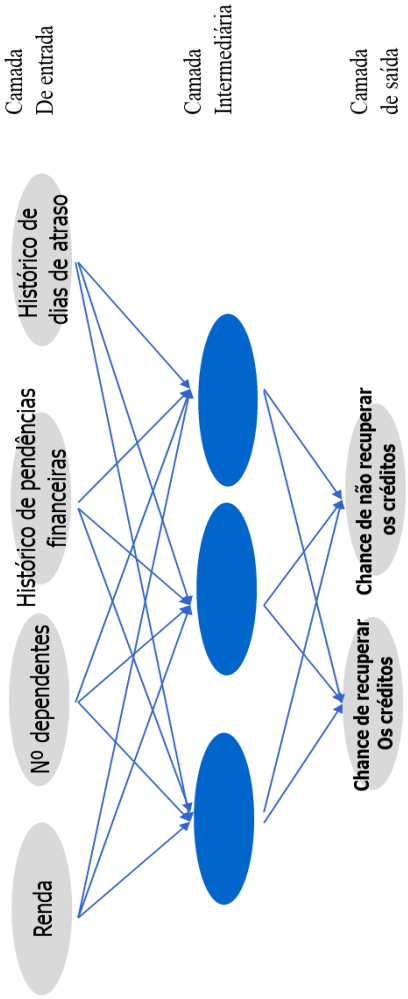
- ➢ Tangente hiperbólica:



$$f(v) = a\frac{e^{(bv)} - e^{(-bv)}}{e^{(bv)} + e^{(-bv)}}$$

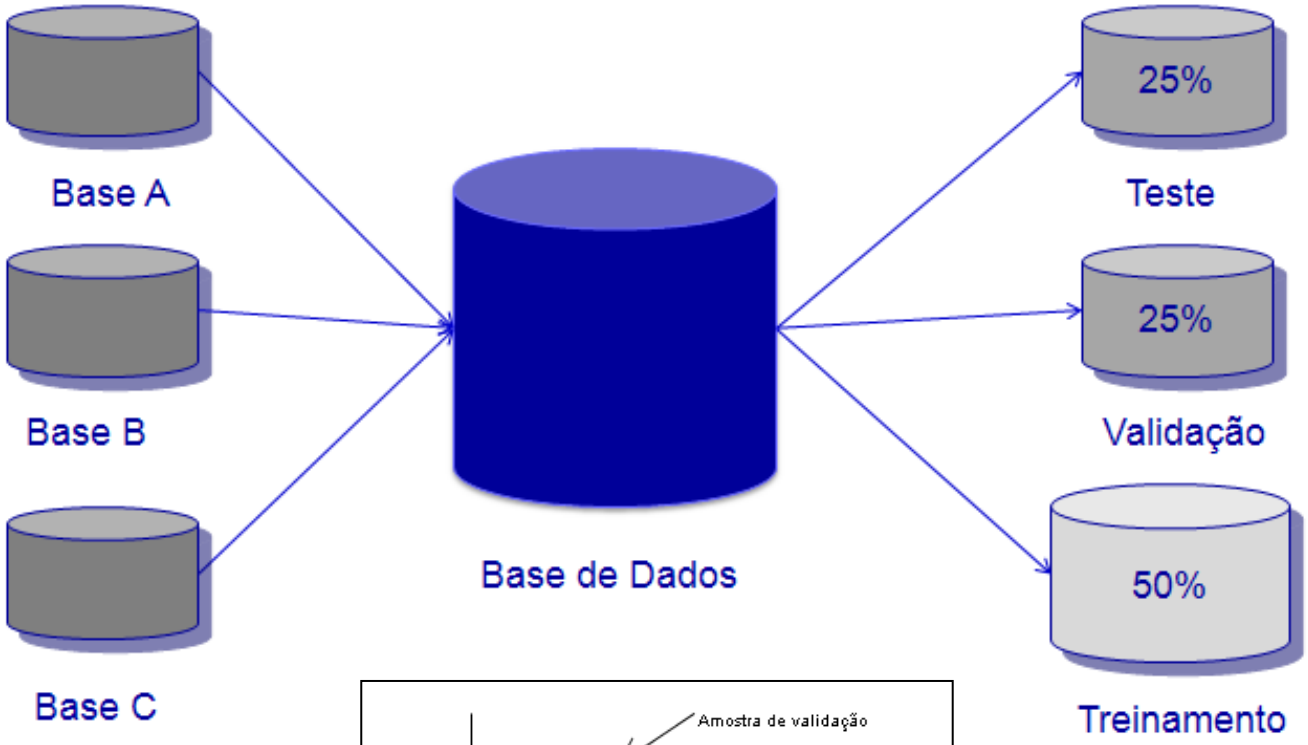
Redes Neurais Artificiais - MLP

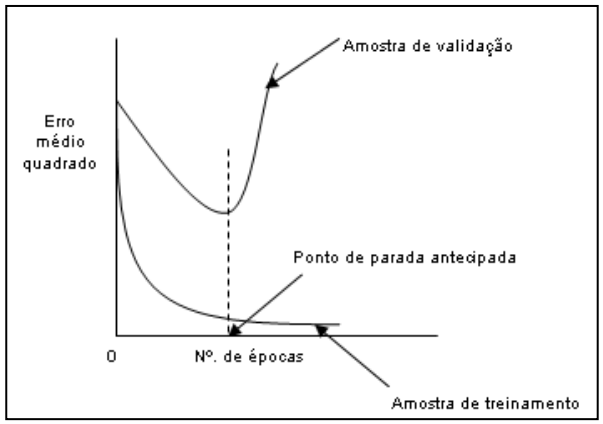




Redes Neurais Artificiais

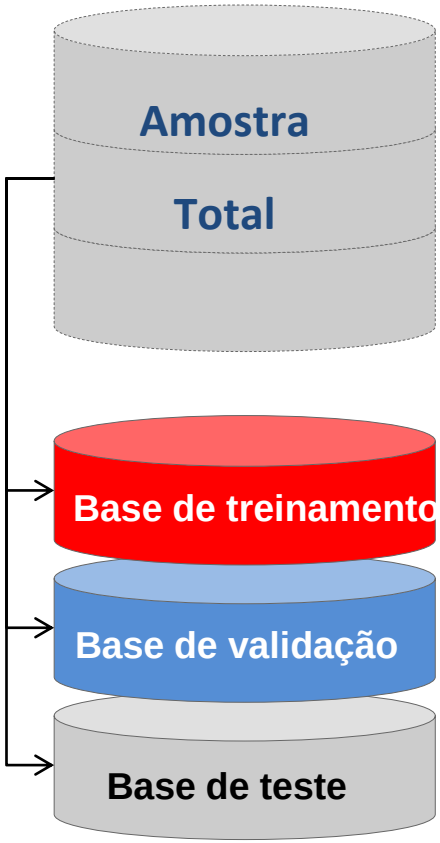
Definição dos Conjuntos



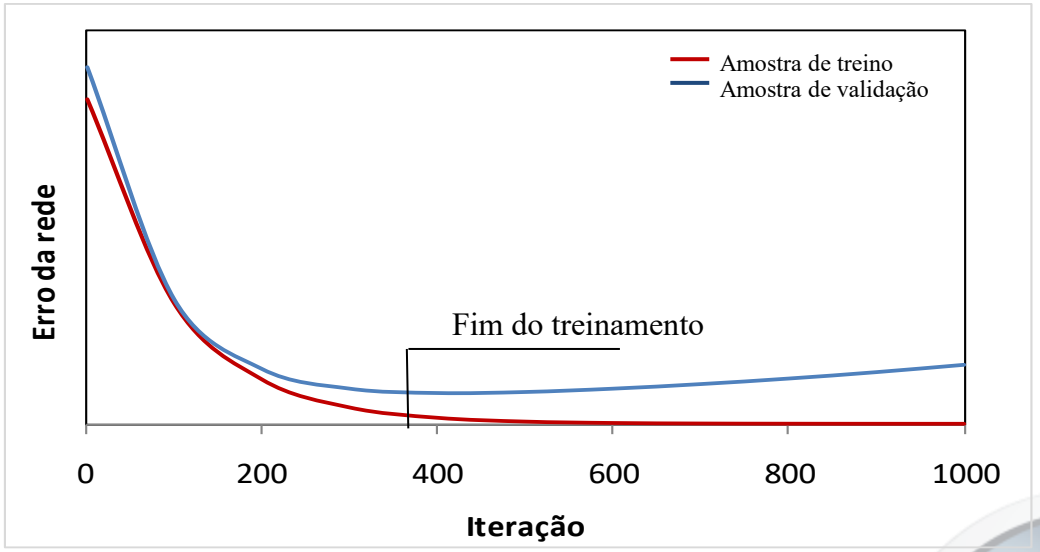


Redes Neurais Artificiais

Definição dos Conjuntos



- ✓ **Base de treinamento:** usada no treino para ajustar os pesos;
- ✓ **Base de validação:** usada como critério de parada de ajustes dos pesos;
- ✓ **Base de teste**: usada para verificar o desempenho da rede após todo o treinamento e se rede possui capacidade de generalização.



Maior chance da memorização (super treinamento) → **+ iterações**

Generalização é prejudicada se o número de iterações for grande

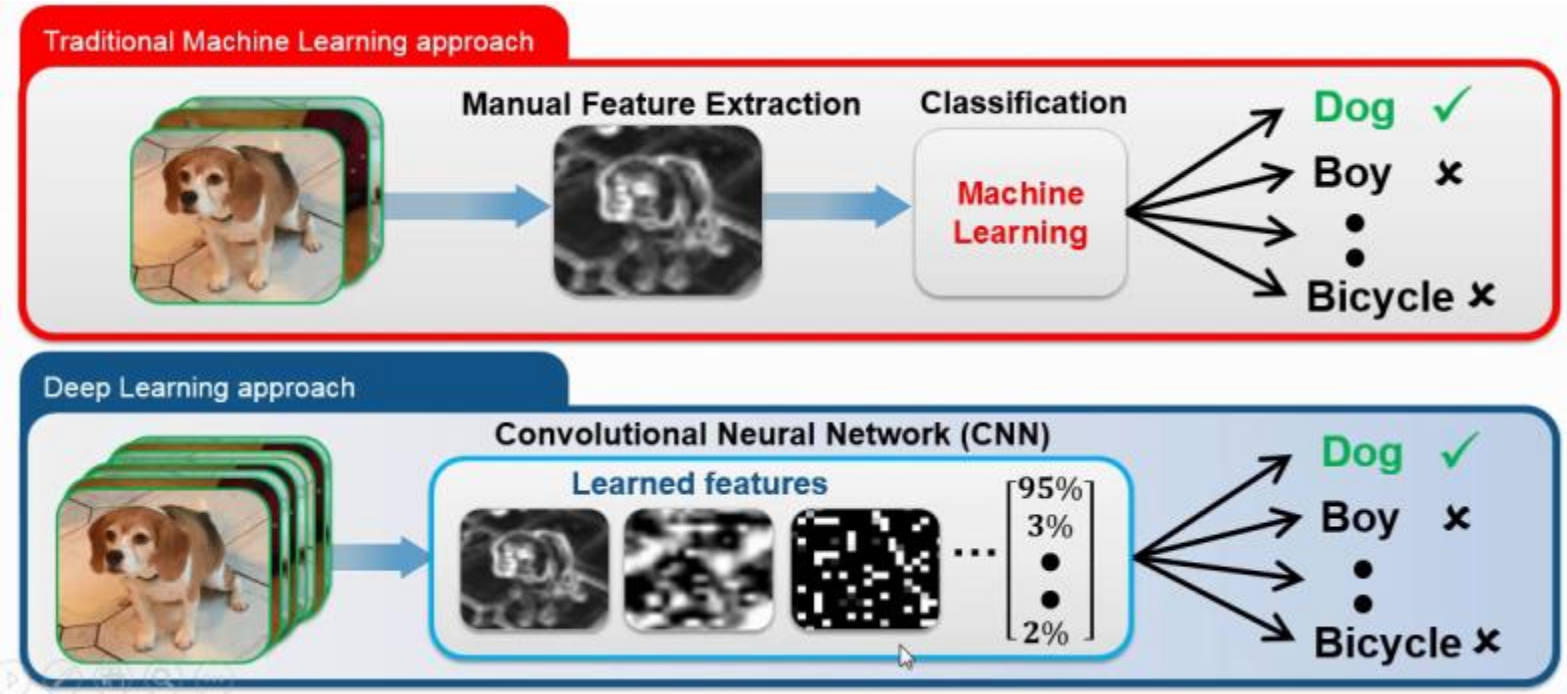
Memorização vs generalização

Faculdade IMPACTA TECNOLOGIA

Deep Learning

Aprendizado de Máquina

O Deep Learning pode aprender representações ou características úteis diretamente de imagens, texto e som.



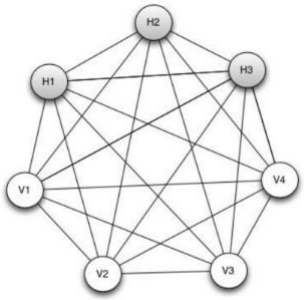
https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/deep_learning.html

Faculdade IMPACTA TECNOLOGIA

Aprendizado de Máquina

Exemplo de Regressão Linear

Modelos com Fórmulas Matemáticas



$$Y_i=\beta_0+\beta_1 X_{i1} + \beta_2 X_{i2} +\ldots+ \beta_p X_{ip} + \varepsilon_i$$

Regressões Lineares Múltiplas

Score de Propensão - Pagar Dívida

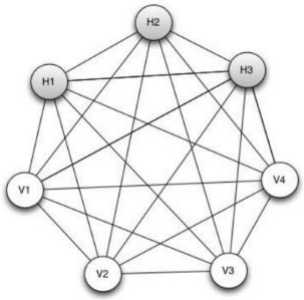
	Peso		*Variável*	
Y =	50	+		
	10	*	FLAG_IDADE_MAIOR_25	+
	-5	*	FLAG_IDADE_MENOR_26	+
	10	*	RENDA_MAIOR_5000	+
	-5	*	RENDA_MENOR_5001	+
	30	*	FLAG_SEM_RESTRICAO_SPC	+
	-40	*	FLAG_COM_RESTRICAO_SPC	

Exemplo: *Pessoa de 20 anos com renda > 5000 e sem restrições no SPC tem* ***Score = 85***

Aprendizado de Máquina

Evolução: Regressão Logística

Modelos com Fórmulas Matemáticas

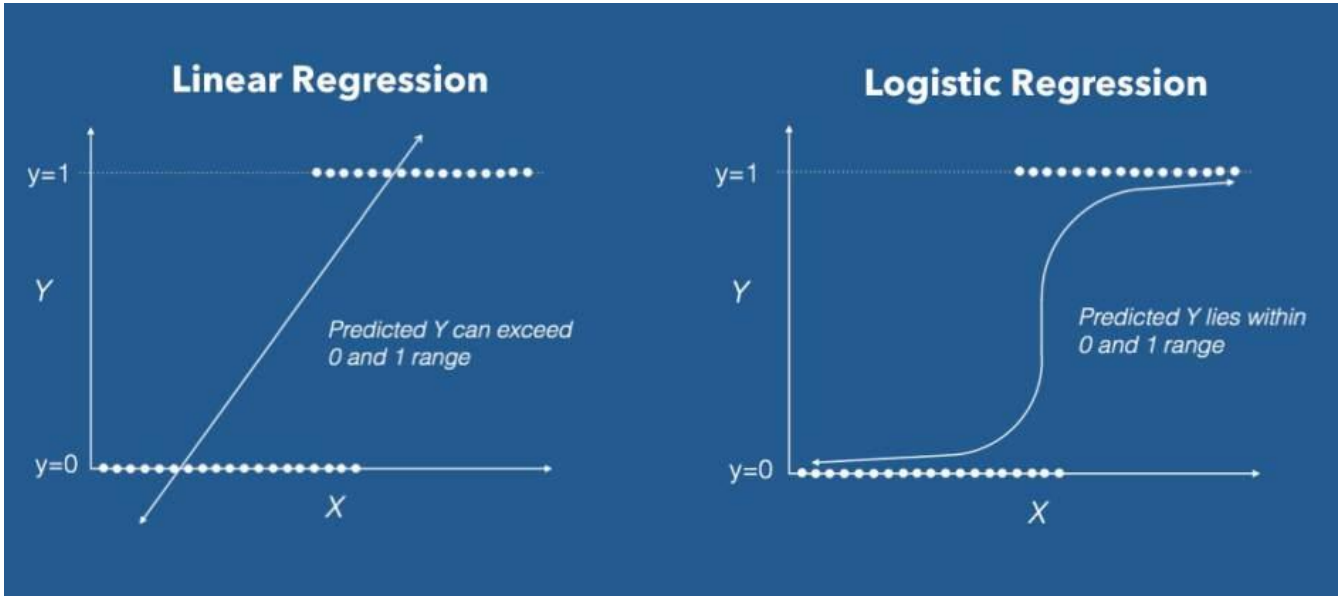


$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

Regressões Lineares Múltiplas

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 z_{1,i} + \cdots + \beta_k z_{k,i})}}$$

Regressões Logísticas

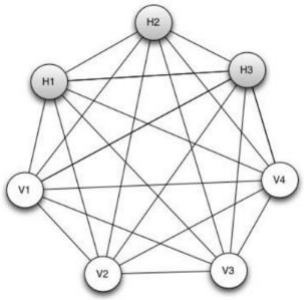


Faculdade IMPACTA TECNOLOGIA

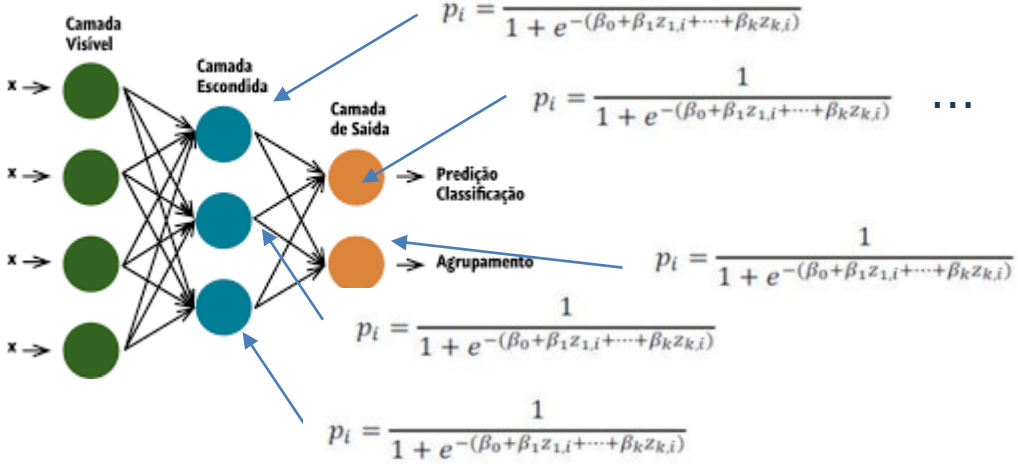
Aprendizado de Máquina

Deep Learning

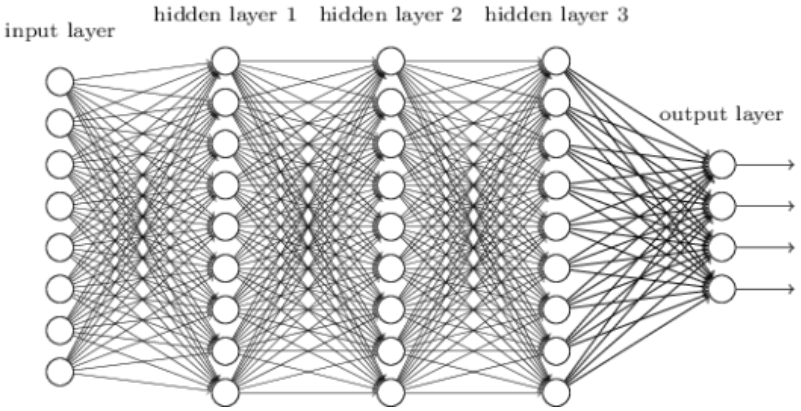
Modelos com Fórmulas Matemáticas

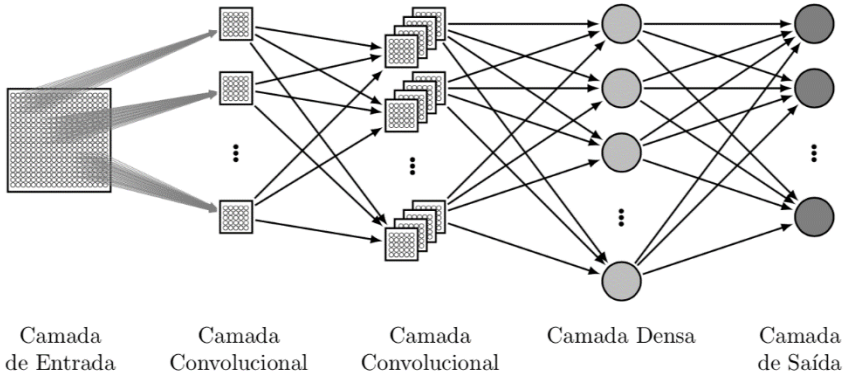


Redes Neurais Artificiais



Deep Learning

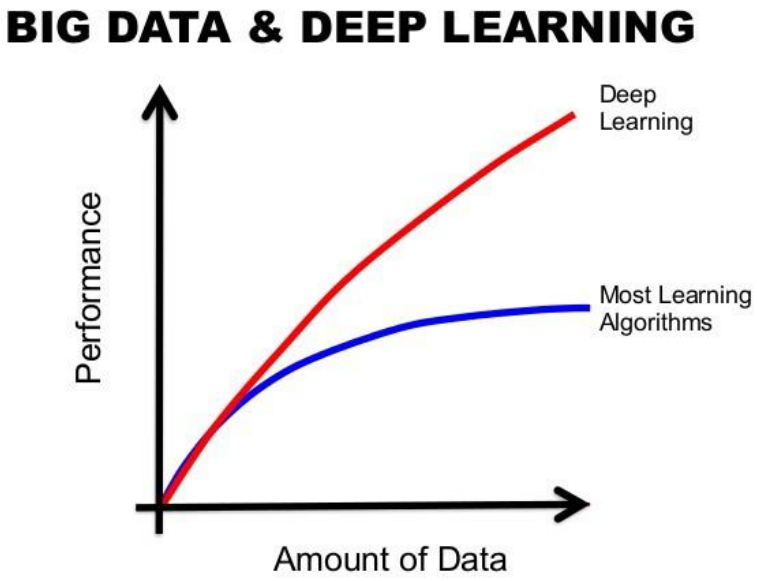




Aprendizado de Máquina

Deep Learning

Redes Neurais, em especial Deep Learning, são técnicas poderosas e capazes de representar conhecimentos complexos contidos nos dados.



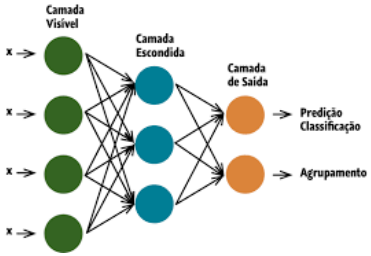
Faculdade IMPACTA TECNOLOGIA

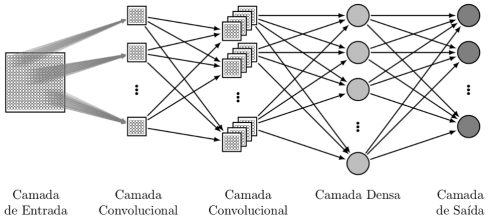
Aprendizado de Máquina

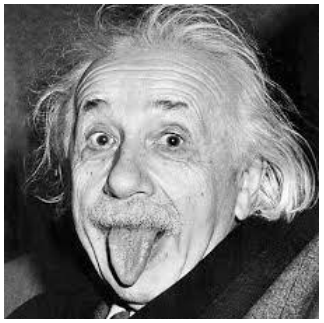
Deep Learning - A Metáfora

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 z_{1,i} + \cdots + \beta_k z_{k,i})}}$$







Aprendizado de Máquina

Deep Learning

- Ajusta elevada quantidade de parâmetros;
- Necessita de um grande número de exemplos (registros) para aprender eficientemente;
- Menor tempo de pré-processamento dos dados;
- Difícil compreensão/visualização humana (caixa preta);

Obrigado!

Prof. Roberto Santos
roberto.santos@faculdadeimpacta.com.br