Machine Learning em Ciências Atuarias

Gabriel D'Assumpção de Carvalho

Ciências Atuarias - UFPE

24 de julho de 2025

Programação

Introdução

Inteligência Artificial (IA)

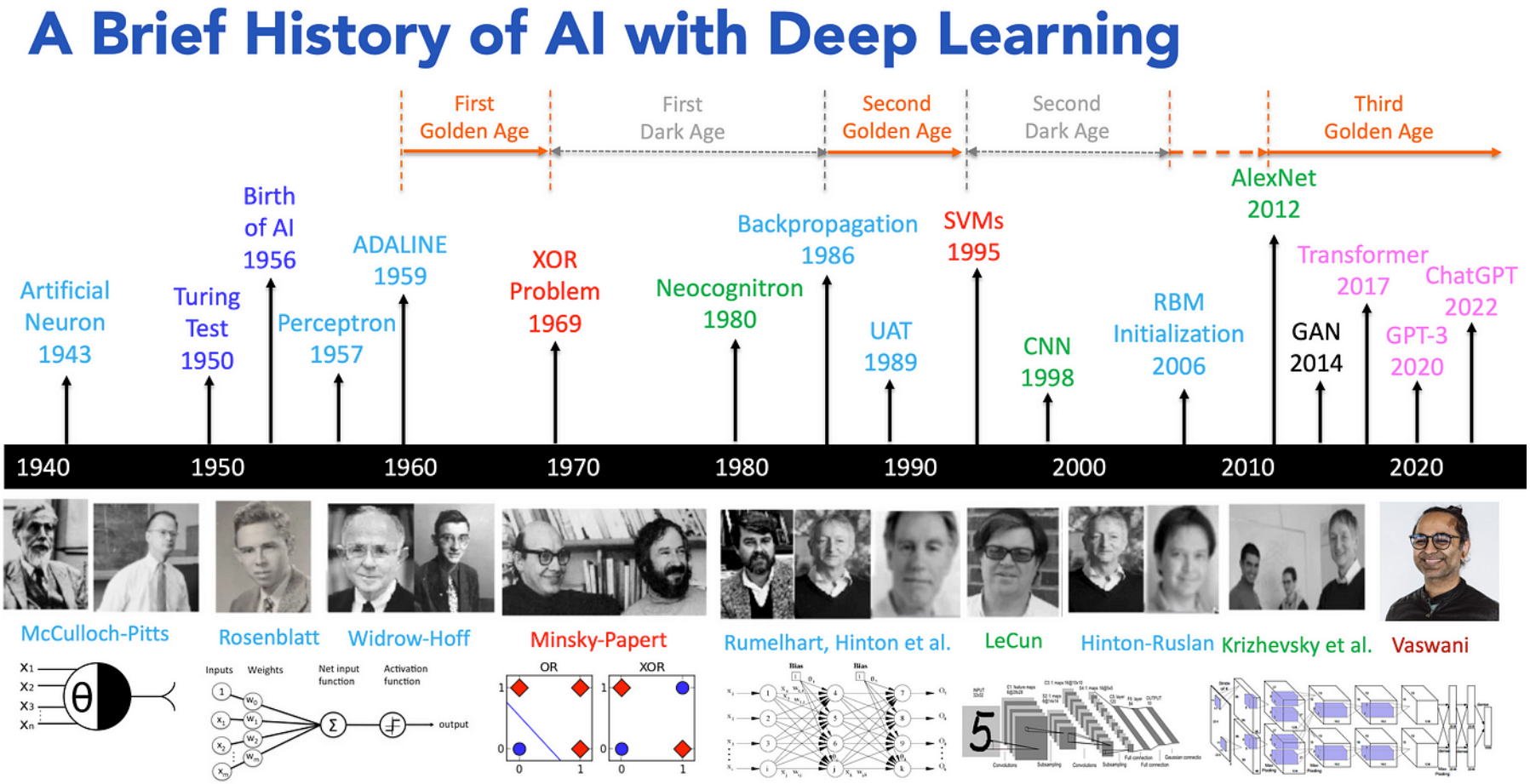

Figura: Linha Temporal da Inteligência Artificial

Áreas da IA

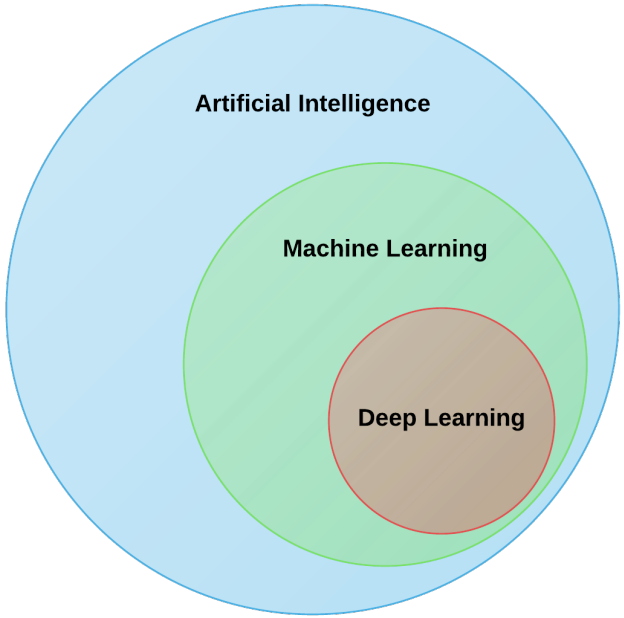

Figura: Conjuntos da IA: Machine Learning e Deep Learning

- **Machine Learning (ML):** Modelos estatísticos multivariados cujos parâmetros são estimados com base em dados observados.
- **Deep Learning (DL):** Subcampo de ML baseado em redes neurais artificiais. Os parâmetros são otimizados por meio do algoritmo de retropropagação (backpropagation).

Relação com Ciências Atuariais

As Ciências Atuariais utilizam ferramentas estatísticas e matemáticas para modelar e controlar riscos. Como o risco é um evento futuro e incerto, o atuário tem interesse em:

- Predição de eventos de fraude;
- Precificação de apólices de seguro;
- Modelagem de riscos operacionais, financeiros e atuariais.

Tipos de Aprendizado

- **Aprendizado Supervisionado:** Os dados são compostos por variáveis explicativas (X) e uma variável resposta (Y). O objetivo é estimar Y com base nas combinações de X.
  Exemplo: Prever o custo de um sinistro.
- **Aprendizado Não Supervisionado:** Os dados são compostos apenas por variáveis explicativas (X), e o objetivo é identificar padrões ou agrupar observações semelhantes.
  Exemplo: Segmentação de clientes de seguro.
- **Aprendizado por Reforço:** Um agente toma decisões em um ambiente para maximizar uma função de recompensa, aprendendo com tentativa e erro.
  Exemplo: Estratégias automatizadas de investimento.

Modelos Supervisionados

Modelos de Regressão

- **Regressão Linear:** Utilizado quando a variável resposta $Y \in \mathbb{R}$ apresenta relação linear com as variáveis explicativas X. O modelo é definido como:

$$Y = E[Y|X] = X^\top \beta + E[\varepsilon], \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

  Onde se assume que os erros são independentes e homocedásticos: $\varepsilon_i \perp \varepsilon_j, \ \forall i \neq j$.

- **Modelo Linear Generalizado (GLM):** Utilizado quando a relação entre Y e X não é linear ou quando Y pertence a domínios restritos (ex: $Y \in \mathbb{R}^+$ ou $Y \in [0, 1]$). A estrutura do modelo é:

$$f(\mathbb{E}[Y|X]) = X^\top \beta$$

  Onde $f(\cdot)$ é a função de ligação (link function), escolhida segundo a distribuição da variável resposta.

Modelos Lineares Generalizados (GLM)

Contexto: GLMs estendem a regressão linear para variáveis resposta com distribuições não-normais (Ex: Poisson, Gamma, Binomial).

Aplicações Atuariais:

- Precificação de apólices (ex: custo médio de sinistro).
- Modelagem da frequência de eventos (Poisson).
- Modelagem do custo com distribuição Gamma.

Referência Teórica: Artigo de aplicação em ciências atuariais com GLM.

Artigo: `PANN.pdf`

Support Vector Machine (SVM)

Objetivo: Separar observações em diferentes classes, mesmo quando linearmente inseparáveis no espaço original.

Exemplo Prático no R: Visualização da separação não-linear com Kernel e projeção em 3D.

- Dataset simulado com duas classes (uma circular dentro da outra).
- Aplicação de kernel polinomial implícito: $z = x^2 + y^2$.
- Visualização com `plotly::plot_ly()`.

`Código R no GitHub [Clique Aqui]`

K-NN (K-Nearest Neighbors)

Ideia Central: Classifique uma observação com base nas classes mais comuns entre seus K vizinhos mais próximos no espaço das variáveis explicativas.

Características:

- Método **não paramétrico** (não assume forma funcional).
- Altamente interpretável e intuitivo.
- Sensível à escala dos dados — uso comum de padronização.

Aplicação Atuarial:

- Detecção de outliers ou fraudes.
- Classificação de perfil de risco de clientes.

Modelos Não Supervisionados

K-means

O algoritmo K-means é uma técnica de agrupamento não supervisionado baseada em centroides. Seu objetivo é particionar os dados em k clusters distintos, minimizando a variabilidade interna de cada grupo. Cada cluster é representado por um centroide, que corresponde à média dos pontos atribuídos a ele.

- K é um hiperparâmetro definido previamente (número de clusters);
- Inicializa aleatoriamente k centroides no espaço dos dados;
- Atribui cada ponto ao cluster com o centroide mais próximo (geralmente usando distância Euclidiana);
- Recalcula os centroides como a média dos pontos atribuídos a cada cluster;
- Repete os passos de atribuição e atualização até convergência (ou atingir número máximo de iterações);
- Convergência: ocorre quando os centroides não se movem mais significativamente.