

REDES NEURAIS

ALGORITMO DE BACKPROPAGATION E
AJUSTE DE HIPERPARÂMETROS

Docente: Dr. Thales Levi Azevedo Valente

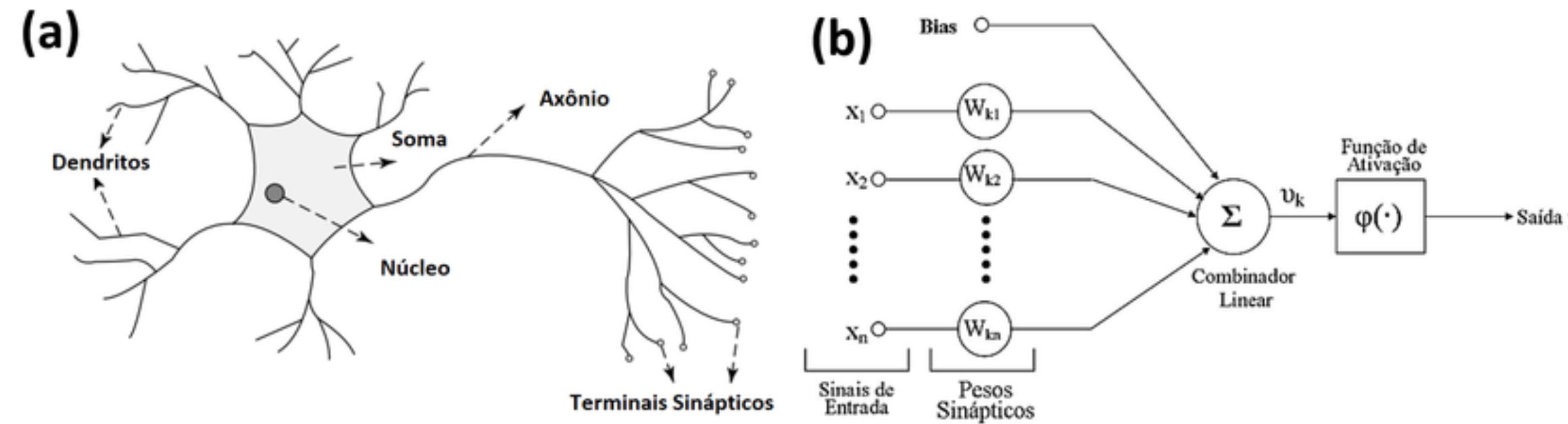
Discentes: Caio Reis, Katarina Ires e Melissa

SUMÁRIO

- 03 O que é uma rede neural
- 04 O que é o Backpropagation
- 05 Importância do Backpropagation
- 07 A estrutura de um MLP
- 10 Analogia com Backpropagation
- 11 Fases do Backpropagation
- 22 Visualização e exemplos no ipynb
- 23 Exemplo prático
- 24 Problemas no Backpropagation
- 33 O que são Hiperparâmetros?
- 35 Principais Hiperparâmetros
- 42 Técnicas para Ajuste de Hiperparâmetros
- 62 Impacto dos Hiperparâmetros
- 66 Técnicas para Evitar Overfitting
- 70 Por que Otimizar Hiperparâmetros é Importante?
- 71 Melhores Práticas para o ajuste de Hiperparâmetros

O QUE É UMA REDE NEURAL?

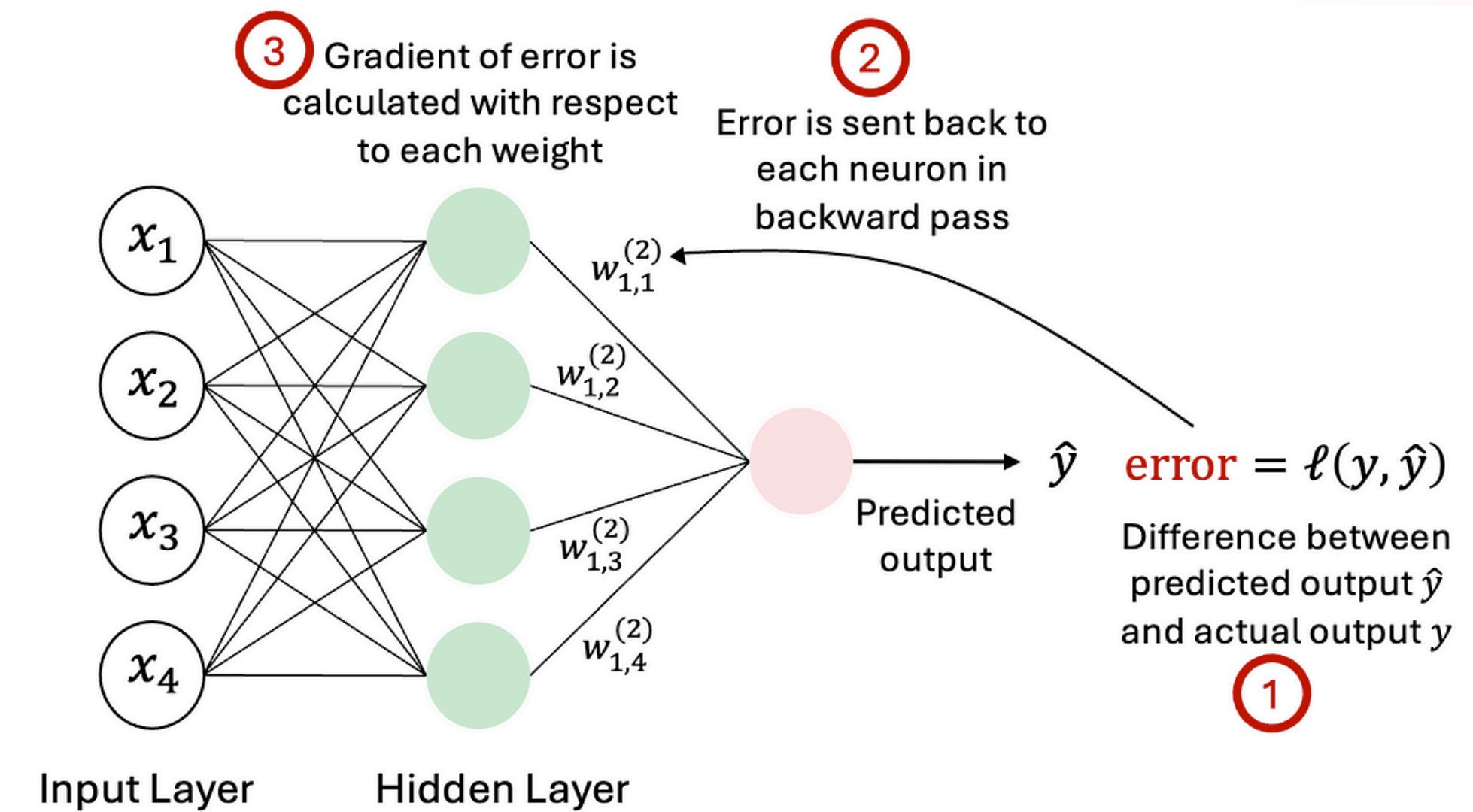
- Inspirada no cérebro humano
- Aprende com dados
- Usa conexões entre “neurônios” articiais
- Aplica-se em tarefas como: reconhecimento de voz, imagem, texto...



Fonte: Sarmento, Arianne & Dos Santos, Wellington. (2022)

O QUE É O BACKPROPAGATION?

O Backpropagation (propagação reversa) é um algoritmo fundamental no treinamento de redes neurais artificiais. Desenvolvido na década de 1970 e popularizado em 1986 por Rumelhart, Hinton e Williams, este algoritmo revolucionou o campo do aprendizado de máquina ao fornecer um método eficiente para treinar redes neurais multicamadas

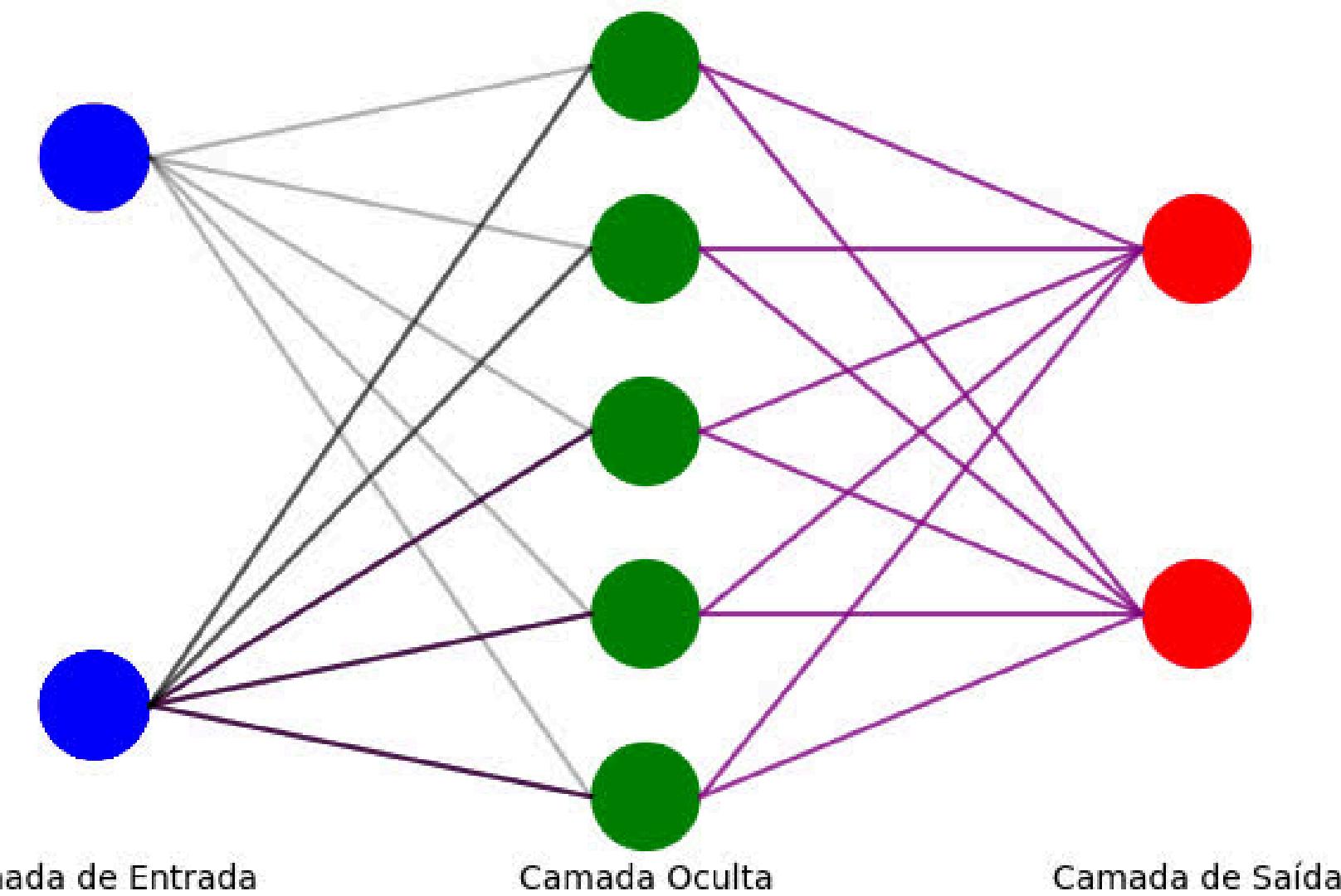


Fonte: medium

IMPORTÂNCIA DO BACKPROPAGATION

- Viabiliza o treinamento de redes multcamadas (rede com camadas ocultas)
- Aprendizado automático de características
- Base para Deep Learning
- Aplicabilidade universal

Propagação do Erro para Camadas Anteriores

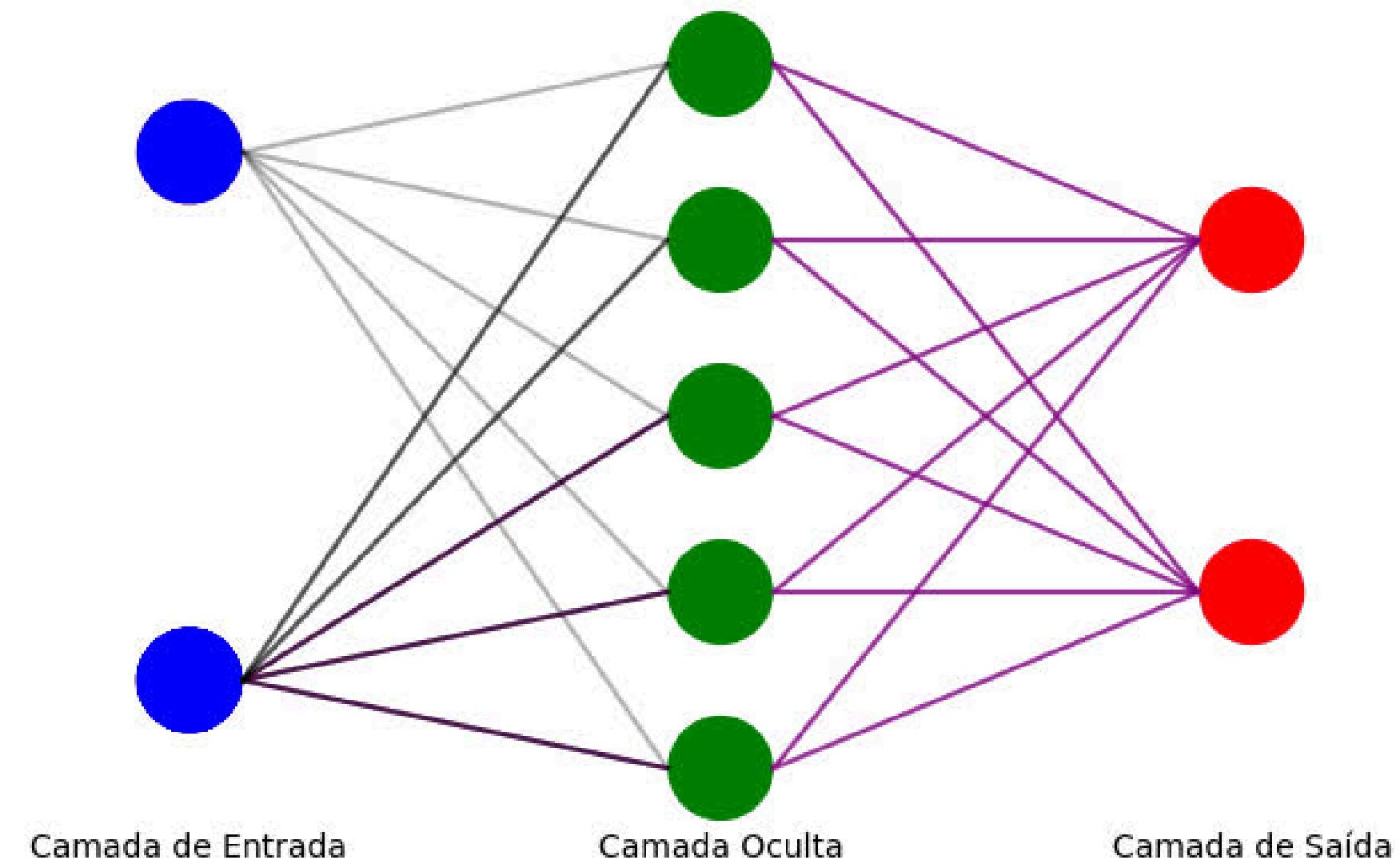


IMPORTÂNCIA DO BACKPROPAGATION

O que são camadas ocultas?

- Viabiliza o treinamento de redes multcamadas (rede com camadas ocultas)
- Aprendizado automático de características
- Base para Deep Learning
- Aplicabilidade universal

Propagação do Erro para Camadas Anteriores



A ESTRUTURA DE UM MLP

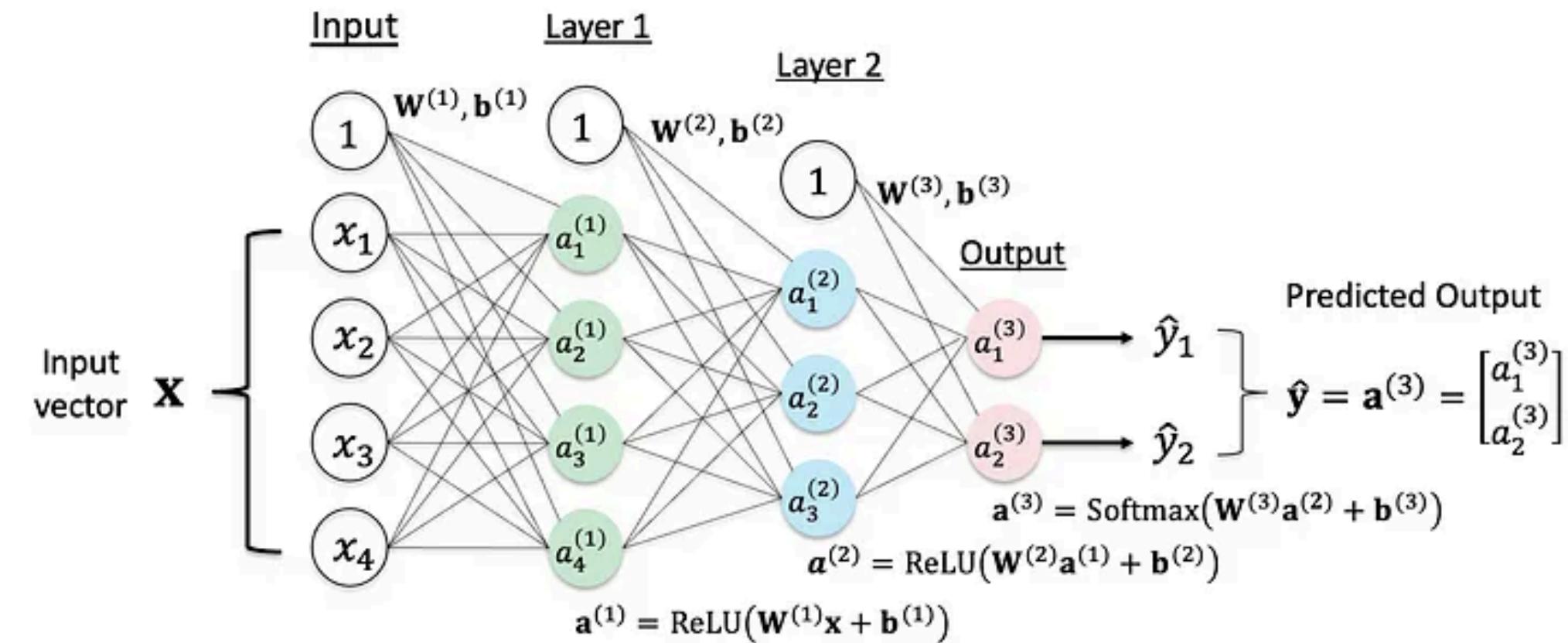
$$\mathbf{a}^{(l)} = g^{(l)}(\mathbf{W}^{(l)} \mathbf{a}^{(l-1)} + \mathbf{b}^{(l)})$$

$$\mathbf{a}^{(l)} = \begin{bmatrix} a_1^{(l)} \\ a_2^{(l)} \\ \vdots \\ a_{n_l}^{(l)} \end{bmatrix} = g^{(l)} \left(\begin{bmatrix} w_{1,1}^{(l)} & w_{1,2}^{(l)} & \dots & w_{1,n_l-1}^{(l)} \\ w_{2,1}^{(l)} & w_{2,2}^{(l)} & \dots & w_{2,n_l-1}^{(l)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n_l,1}^{(l)} & w_{n_l,2}^{(l)} & \dots & w_{n_l,n_l-1}^{(l)} \end{bmatrix} \begin{bmatrix} a_1^{(l-1)} \\ a_2^{(l-1)} \\ \vdots \\ a_{n_{l-1}}^{(l-1)} \end{bmatrix} + \begin{bmatrix} b_1^{(l)} \\ b_2^{(l)} \\ \vdots \\ b_{n_l}^{(l)} \end{bmatrix} \right)$$

onde:

- W(1) e b(1) são a matriz de peso e o vetor de polarização para a camada l
- a(1) é a saída de ativação de todos os neurônios na camada l
- g(1) é a função de ativação para a camada l

A ESTRUTURA DE UM MLP



Fonte: medium

$$\hat{\mathbf{y}} = f_{\theta}(\mathbf{x}) = \text{Softmax}(\mathbf{W}^{(3)} \text{ReLU}(\mathbf{W}^{(2)} \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) + \mathbf{b}^{(3)})$$

Fonte: medium

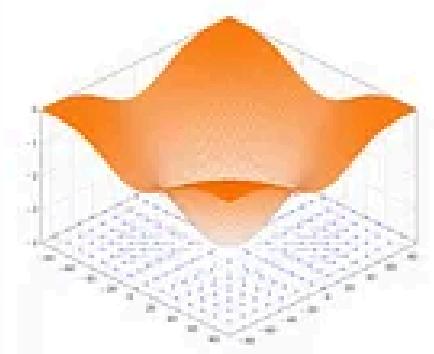
A ESTRUTURA DE UM MLP

- A necessidade de retropropagação:

Uma abordagem ingênua envolveria perturbar cada parâmetro individualmente e observar a variação na função de custo $\mathcal{L}(\theta)$. Esse método é computacionalmente proibitivo, especialmente para redes com milhões de parâmetros

- $\hat{y} = f_{\theta}(\mathbf{x}) = g^{(L)}(\mathbf{W}^{(L)} \dots g^{(2)}(\mathbf{W}^{(2)}g^{(1)}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) \dots + \mathbf{b}^{(L)})$
- $\theta = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}, \dots, \mathbf{W}^{(L)}, \mathbf{b}^{(L)}\}$ Parameters of the Neural Network

$$\mathbf{W}^{(l)} = \begin{bmatrix} w_{1,1}^{(l)} & w_{1,2}^{(l)} & \dots \\ w_{2,1}^{(l)} & w_{2,2}^{(l)} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad \mathbf{b}^{(l)} = \begin{bmatrix} b_1^{(l)} \\ b_2^{(l)} \\ \vdots \end{bmatrix}$$



$$\nabla_{\theta} \mathcal{L}(\theta_t) = \begin{bmatrix} \frac{\partial \mathcal{L}(\theta_t)}{\partial w_{1,1}^{(l)}} \\ \frac{\partial \mathcal{L}(\theta_t)}{\partial w_{1,2}^{(l)}} \\ \vdots \\ \frac{\partial \mathcal{L}(\theta_t)}{\partial b_i^{(l)}} \end{bmatrix}$$

Gradient Descent Algorithm

Initialization: start at θ_0

while ($\theta_{t+1} \neq \theta_t$)

{

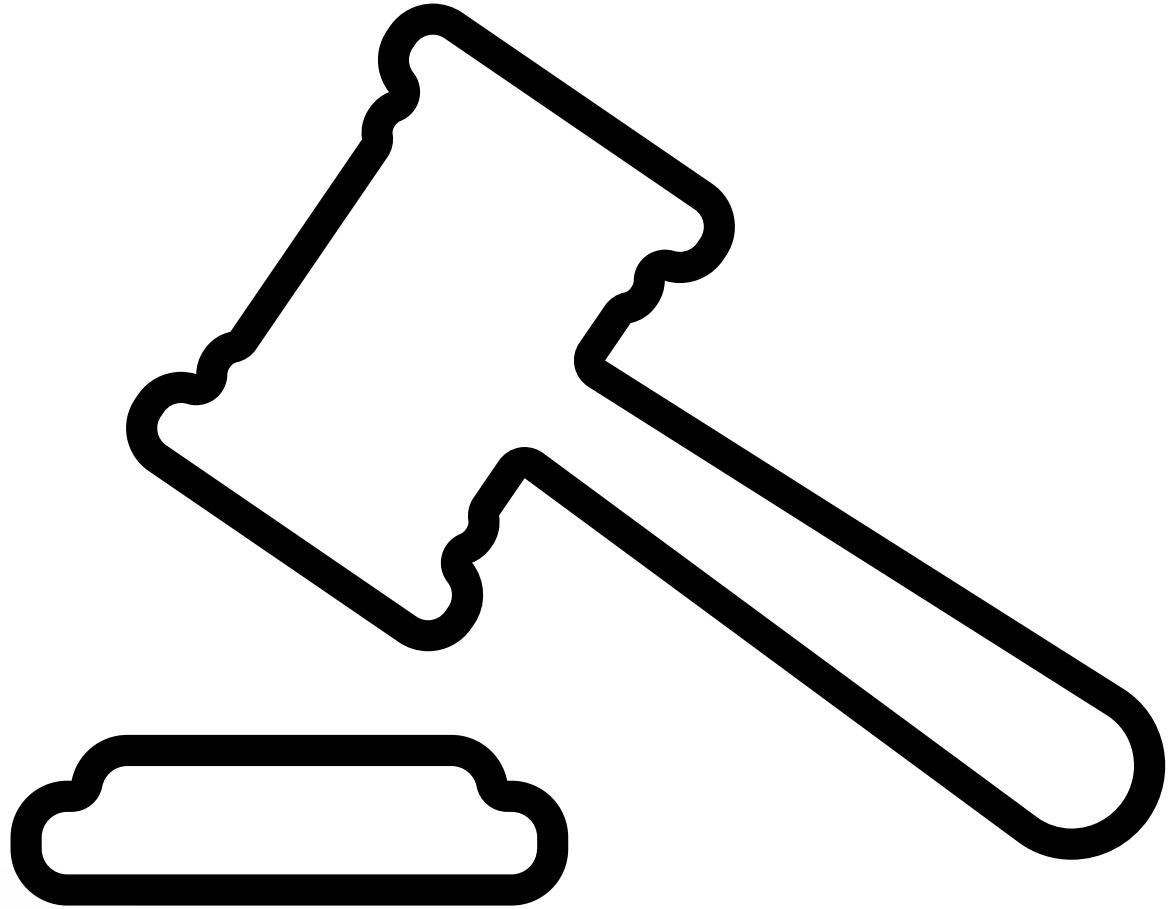
compute gradient of θ_t at t and update parameters

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta} \mathcal{L}(\theta_t)$$

To efficiently compute the gradient when dealing with a large number of parameters, we employ a technique known as backpropagation.

Fonte: medium

ANALOGIA COM O BACKPROPAGATION?



Atribuição de culpa



Aprender a cozinhar

FASES DO BACKPROPAGATION

Propagação do Erro para Camadas Anteriores

01

Feedforward (Propagação Direta)

02

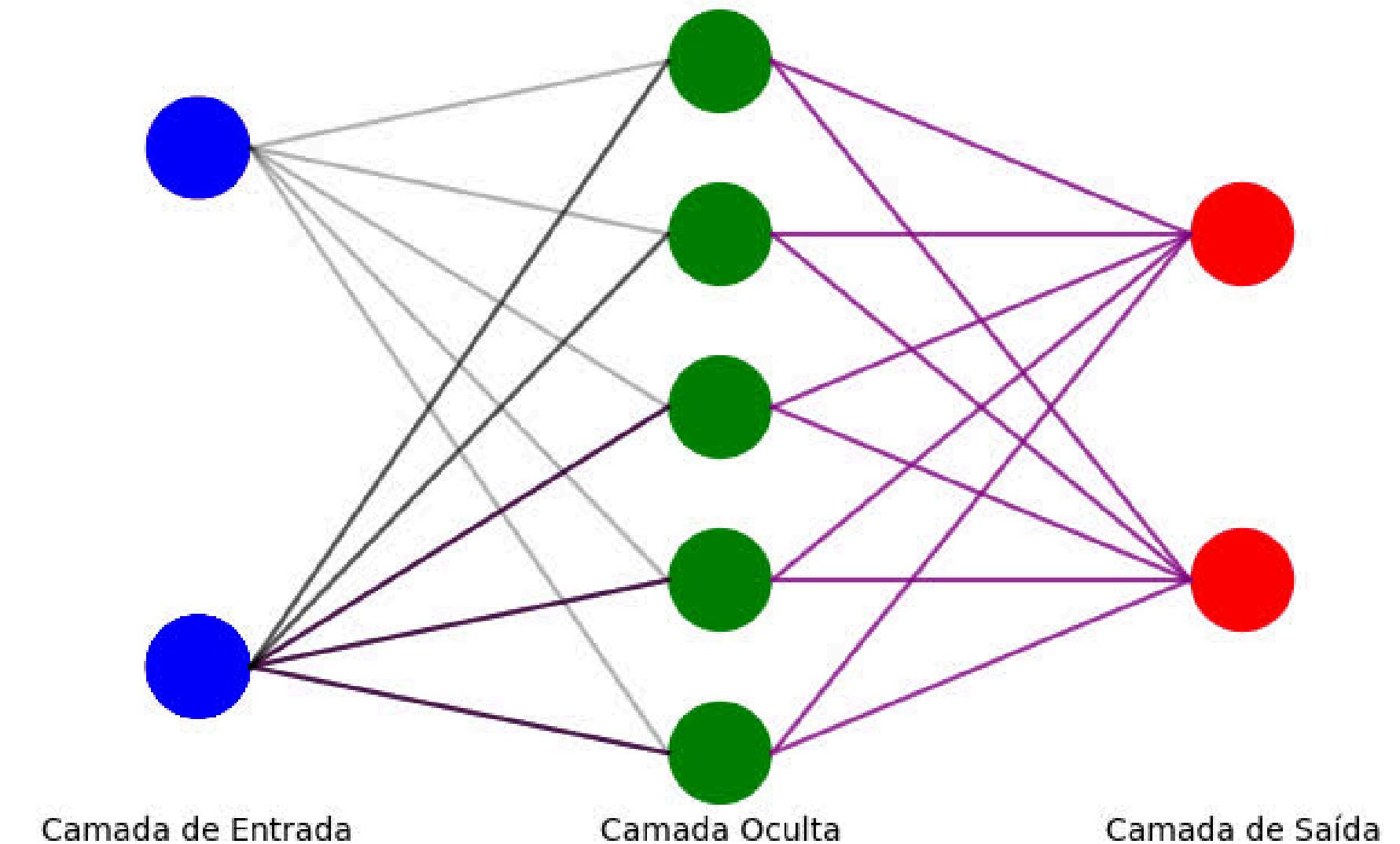
Cálculo do Erro

03

Backward Pass

04

Atualização dos pesos



Feedforward

Matemática do Feedforward

Para cada neurônio em uma camada, realizamos os seguintes cálculos:

$$z_j^l = \sum_{j'}^n w_{jj'}^l a_{j'}^{l-1} + b_j^l$$

Onde:

- z_j^l é a soma ponderada para o neurônio j na camada l
- $w_{jj'}^l$ é o peso da conexão entre o neurônio j' na camada $l - 1$ e o neurônio j na camada l
- $a_{j'}^{l-1}$ é a ativação do neurônio j' na camada $l - 1$
- b_j^l é o viés (bias) do neurônio j na camada l
- n é o número de neurônios na camada $l - 1$

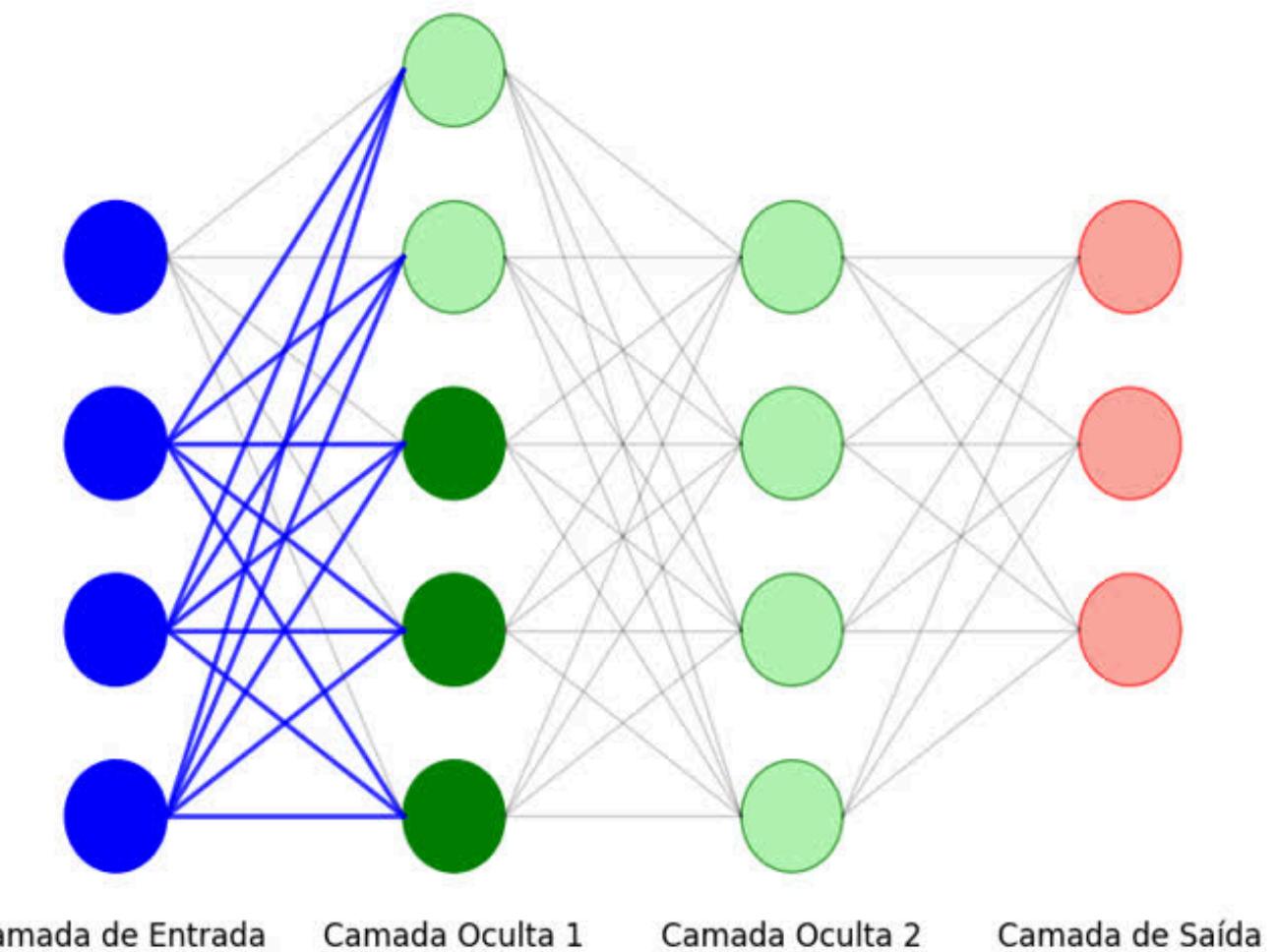
Aplicação da função de ativação:

$$a_j^l = f(z_j^l)$$

Onde:

- a_j^l é a ativação do neurônio j na camada l
- f é a função de ativação (como sigmoid, ReLU, etc.)

Fase de Feedforward: Propagando para Primeira Camada Oculta



Cálculo do Erro

Funções de Erro Comuns

Erro Quadrático Médio (MSE - Mean Squared Error)

O MSE é uma das funções de erro mais comuns para problemas de regressão:

$$E = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Onde:

- E é o erro total
- n é o número de exemplos
- y_i é o valor desejado (target) para o exemplo i
- \hat{y}_i é o valor previsto pela rede para o exemplo i

Suponha que:

- A saída da rede é 0.8
- O valor correto é 1.0
- A função de erro é

$$E = (1.0 - 0.8)^2 = 0.04$$

O backpropagation vai:

Calcular como esse erro depende de cada peso

Atualizar os pesos para reduzir esse erro na próxima rodada de treino

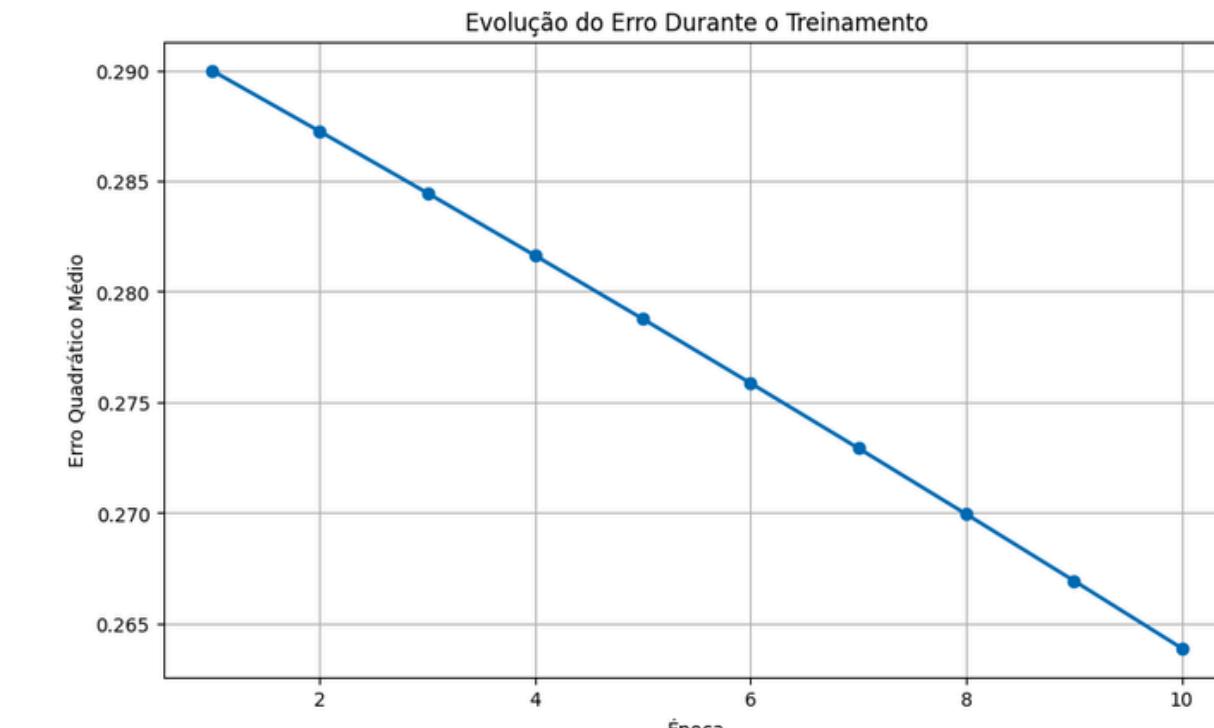
Entropia Cruzada (Cross-Entropy)

A entropia cruzada é frequentemente usada para problemas de classificação:

$$E = - \sum_{i=1}^n y_i \log(\hat{y}_i)$$

Onde:

- E é o erro total
- n é o número de classes
- y_i é o valor desejado (0 ou 1) para a classe i
- \hat{y}_i é a probabilidade prevista pela rede para a classe i

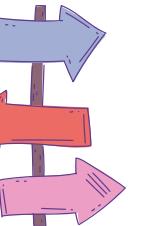


Importância do Cálculo do Erro

- Medida de desempenho



- Direção do aprendizado



- Base para o Backpropagation

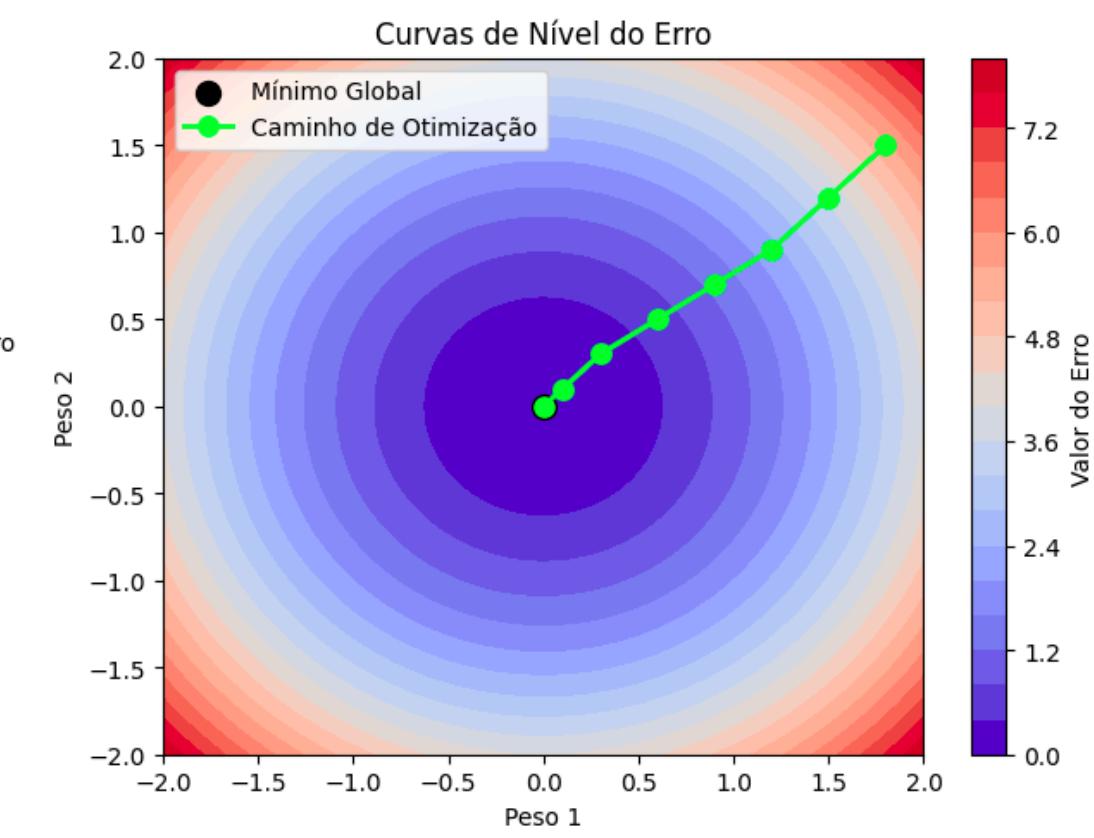
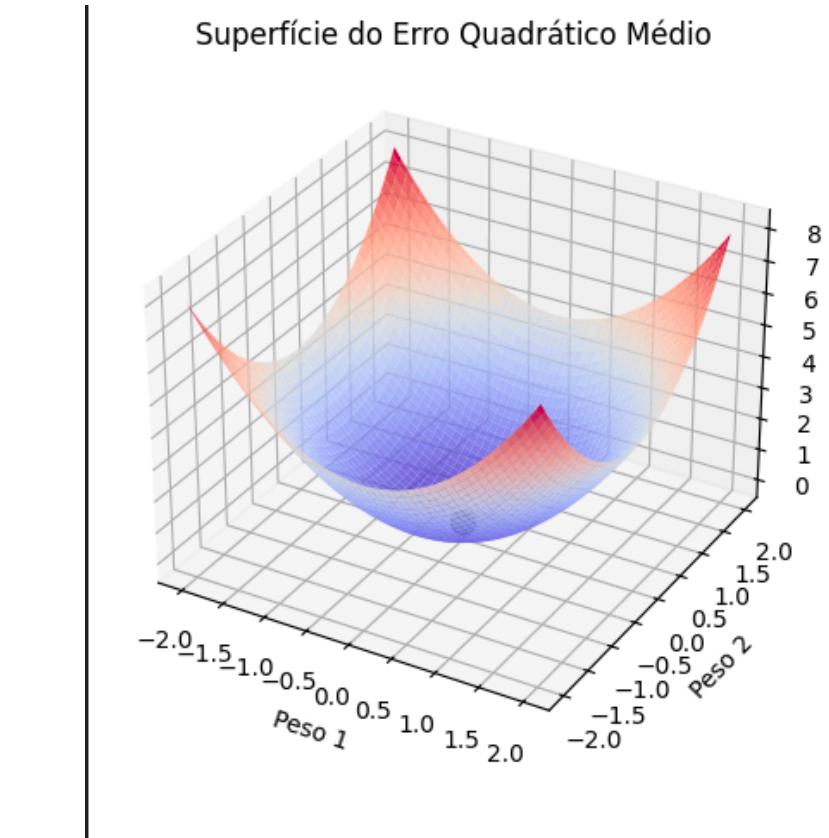
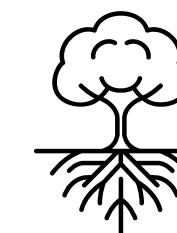


- Critério de parada



Importância do Cálculo do Erro

- Medida de desempenho
- Direção do aprendizado
- Base para o Backpropagation
- Critério de parada



Backpropagation

Princípio da Regra da Cadeia

Se temos uma função composta $E = f(g(h(w)))$, onde w é um peso da rede, a derivada de E em relação a w é dada por:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial f} \cdot \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial h} \cdot \frac{\partial h}{\partial w}$$

No contexto de uma rede neural, esta fórmula nos permite calcular como cada peso contribui para o erro total

Cálculo do Gradiente na Camada de Saída

Para os neurônios na camada de saída, o erro é calculado diretamente comparando a saída prevista com o valor desejado. O gradiente do erro em relação à saída do neurônio é:

$$\delta_j^L = \frac{\partial E}{\partial a_j^L} = \frac{\partial E}{\partial z_j^L} \cdot \frac{\partial z_j^L}{\partial a_j^L}$$

Onde:

- δ_j^L é o gradiente do erro para o neurônio j na camada de saída L
- a_j^L é a ativação (saída) do neurônio j na camada L
- z_j^L é a soma ponderada das entradas para o neurônio j na camada L

Para o erro quadrático médio (MSE) e a função de ativação sigmoid, temos:

$$\delta_j^L = (a_j^L - y_j) \cdot f'(z_j^L)$$

Onde:

- y_j é o valor desejado para o neurônio j
- $f'(z_j^L)$ é a derivada da função de ativação avaliada em z_j^L

Backpropagation

Cálculo do Gradiente nas Camadas Ocultas

O gradiente para um neurônio em uma camada oculta é:

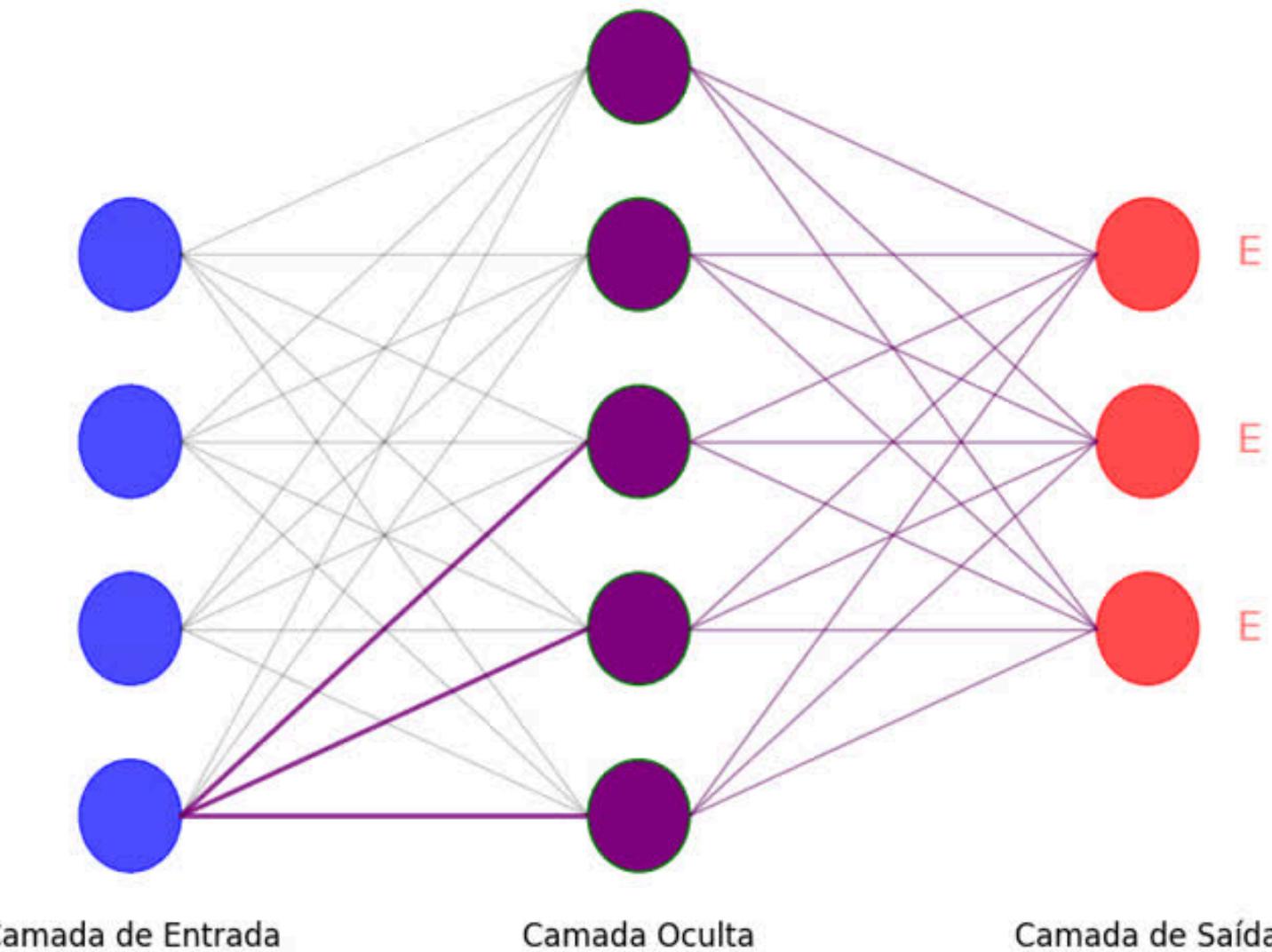
$$\delta_j^l = \left(\sum_{k=1}^{n_{l+1}} w_{kj}^{l+1} \cdot \delta_k^{l+1} \right) \cdot f'(z_j^l)$$

Onde:

- δ_j^l é o gradiente do erro para o neurônio j na camada l
- w_{kj}^{l+1} é o peso da conexão entre o neurônio j na camada l e o neurônio k na camada $l + 1$
- δ_k^{l+1} é o gradiente do erro para o neurônio k na camada $l + 1$
- $f'(z_j^l)$ é a derivada da função de ativação avaliada em z_j^l
- n_{l+1} é o número de neurônios na camada $l + 1$

Esta fórmula mostra como o erro é propagado de volta através da rede, camada por camada

Propagação do Erro: Camada Oculta → Camada de Entrada

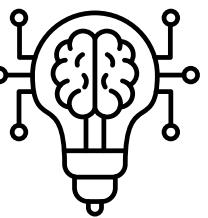


Importância do Backpropagation

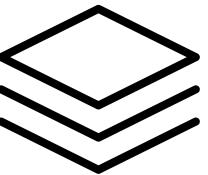
- Atribuição de responsabilidade



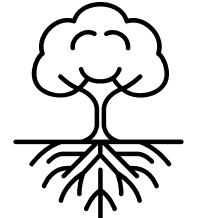
- Aprendizados eficientes



- Aprendizados em camadas profundas



- Base para algoritmos avançados



Atualização dos Pesos

Algoritmo do Gradiente Descendente

No contexto do treinamento de redes neurais, a função a ser minimizada é a função de erro, e os parâmetros são os pesos e vieses da rede. A regra de atualização básica do gradiente descendente é:

$$w'_{ji} = w_{ji} - \eta \frac{\partial E}{\partial w_{ji}}$$

Onde:

- w_{ji} é o peso da conexão entre o neurônio i na camada $l - 1$ e o neurônio j na camada l
- η (eta) é a taxa de aprendizado, um hiperparâmetro que controla o tamanho do passo na direção do gradiente
- $\frac{\partial E}{\partial w_{ji}}$ é o gradiente do erro em relação ao peso w_{ji}

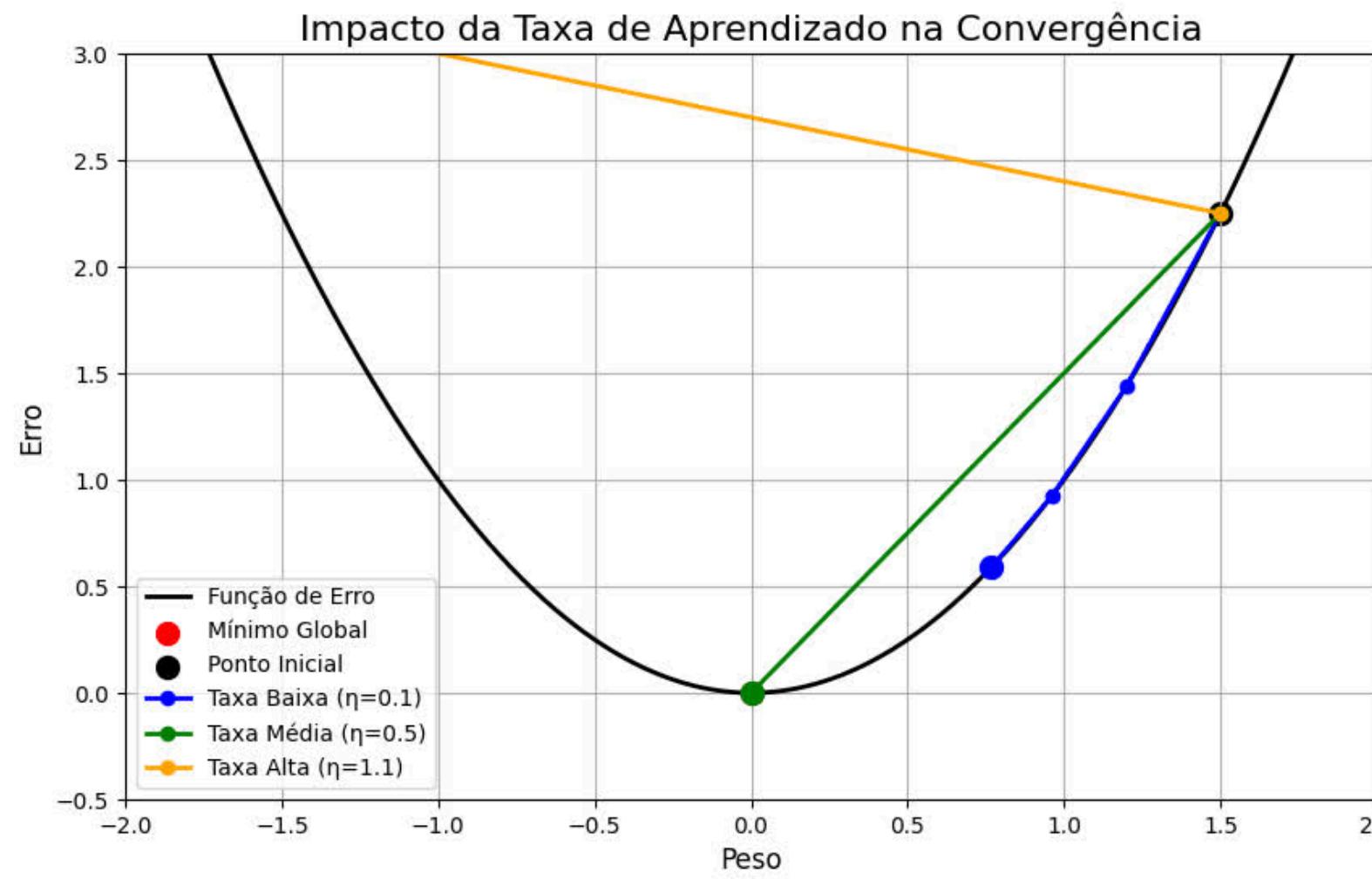
De forma similar, os vieses são atualizados usando:

$$b'_j = b_j - \eta \frac{\partial E}{\partial b_j}$$

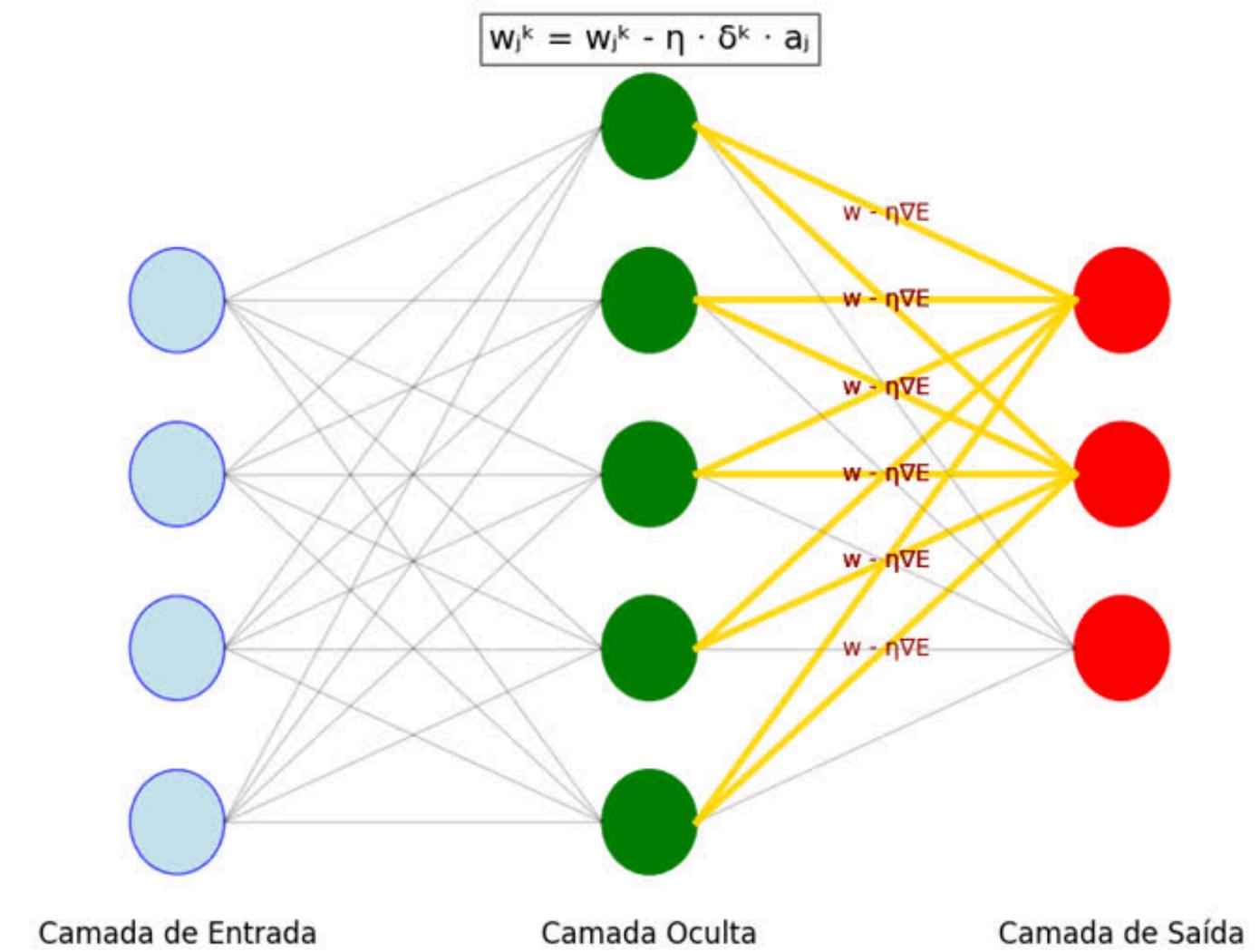
Atualização dos Pesos

Taxa de Aprendizado

- Taxa de aprendizado muito alta: Pode causar oscilações ou divergência, impedindo que o algoritmo encontre o mínimo da função de erro
- Taxa de aprendizado muito baixa: Pode resultar em um treinamento muito lento, exigindo muitas épocas para convergir



Atualização dos Pesos: Camada Oculta → Camada de Saída



Atualização dos Pesos

Variantes do Gradiente Descendente

Existem várias variantes do algoritmo do gradiente descendente que podem melhorar a convergência e o desempenho:

1. Gradiente Descendente Estocástico (SGD): Atualiza os pesos usando um único exemplo de treinamento por vez, em vez de todo o conjunto de dados
2. Gradiente Descendente com Momentum: Adiciona um termo de momentum que ajuda a acelerar a convergência e evitar mínimos locais

$$v = \gamma v - \eta \nabla E$$

$$w = w + v$$

Onde v é o vetor de velocidade e γ é o coeficiente de momentum

3. RMSProp: Adapta a taxa de aprendizado para cada parâmetro com base na magnitude dos gradientes recentes
4. Adam: Combina as ideias do momentum e do RMSProp para uma adaptação mais eficiente da taxa de aprendizado

VISUALIZAÇÃO E EXEMPLOS NO IPYNB



EXEMPLO PRÁTICO

Figure 1
Visualização do Algoritmo de Backpropagation em Redes Neurais
Mostrando todas as etapas do processo de aprendizado
Backpropagation - Época 23

Arquitetura da Rede Neural

Feedforward
Propagação para Frente (Feedforward)

1. Multiplicação de Ativação → **Ativação de Ativação** Camada 1

$$z^1 = W^1 a^0 + b^1 = [0.1653, 0.4551, 0.6317, 0.4961]$$

$$a^1 = \sigma(z^1) = [0.5412, 0.6118, 0.6529, 0.6215]$$

Camada 2

$$z^2 = W^2 a^1 + b^2 = [3.5097]$$

$$a^2 = \sigma(z^2) = [0.9710]$$

Função de Ativação Sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Cálculo do Erro (Binary Cross-Entropy)
Binary Cross-Entropy

$$E = - \sum (y \log(p) + (1 - y) \log(1 - p))$$

Saída da Rede (\hat{y}) Valor Esperado (y): Diferença ($\hat{y} - y$):

$$[0.9710] - [1.0000] \rightarrow [-0.0290]$$

Binary Cross-Entropy: **0.029468**

Backpropagation
Propagação Reversa (Backpropagation)

1. Cálculo das Deltas → **Cálculo dos Gradiientes** → **Propagação do Erro**

Cálculo dos Deltas

$$\delta^L = (\hat{y} - y) \odot \sigma'(z^L)$$

$$\delta^I = ((W^{I+1})^T \delta^{I+1}) \odot \sigma'(z^I)$$

Cálculo dos Gradiientes

$$\frac{\partial E}{\partial W^I} = \delta^I (a^{I-1})^T$$

$$\frac{\partial E}{\partial b^I} = \delta^I$$

Valores dos Deltas (Gradiente do Erro)

Camada 1: $[-0.0106, -0.0039, -0.0079, -0.0069]$
 Camada 2: $[-0.0290]$

$\sigma'(z) = \sigma(z) \odot (1 - \sigma(z))$

Atualização dos Pesos
Atualização dos Pesos e Biases

$$W^I = W^I - \eta \frac{\partial E}{\partial W^I}$$

$$b^I = b^I - \eta \frac{\partial E}{\partial b^I}$$

Taxa de aprendizado (η): 0.51

Exemplos de Atualização de Pesos

Camada	Peso Anterior	Gradiente	Peso Atualizado
Camada 1	-0.0746	-0.0077	-0.0707
Camada 2	1.4587	-0.0164	1.4671

Evolução da Perda (Loss)

Binary Cross-Entropy

Época

Arquitetura da Rede: [2, 4, 1] | Taxa de Aprendizado: 0.5

Redução do erro: 90.36%

PROBLEMAS NO BACKPROPAGATION E HIPERPARÂMETROS

01 Problemas no Backpropagation

02 O que são Hiperparâmetros?

03 Principais Hiperparâmetros

PROBLEMAS NO BACKPROPAGATION

- 01 Desvanecimento do Gradiente
- 02 Explosão do Gradiente
- 03 Overfitting (sobreajuste)
- 04 Sensibilidade a Hiperparâmetros



PROBLEMAS NO BACKPROPAGATION

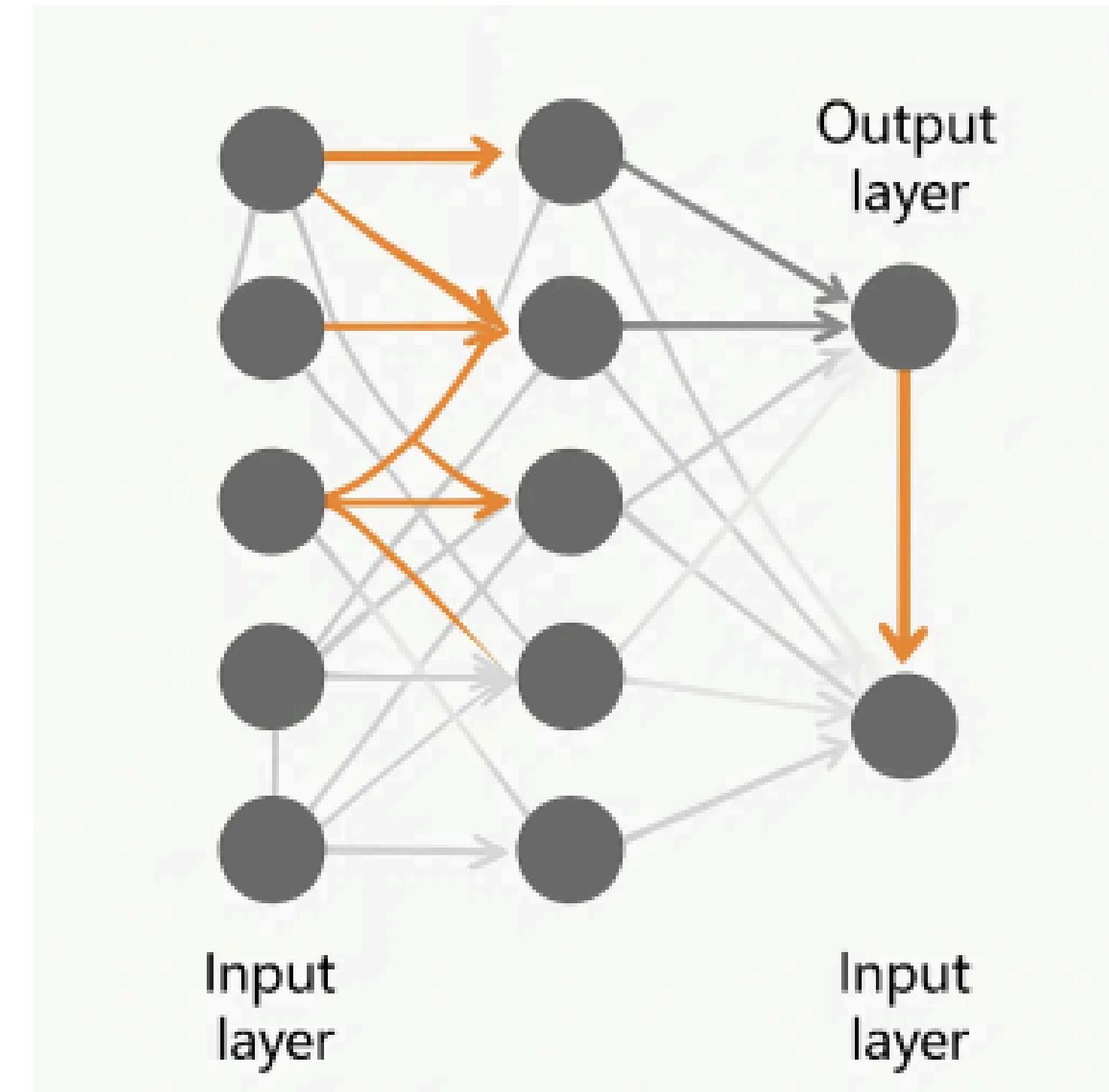
Desvanecimento do Gradiente

- **O que é?**

- Os gradientes se tornam extremamente pequenos à medida que são propagados de volta através das camadas, impedindo o aprendizado efetivo nas camadas iniciais.

- **Por que ocorre?**

- Funções Sigmoide/Hiperbólica: Gradientes máximos de 0.25 e 1.0;
- Redes Profundas: Multiplicação de gradientes pequenos;
- Saturação: Derivadas próximas de zero.



PROBLEMAS NO BACKPROPAGATION

Desvanecimento do Gradiente

- **Consequências:**

- Treinamento extremamente lento;
- Camadas iniciais não aprendem;
- Desempenho subótimo;
- Convergência prejudicada

- **Soluções:**

- ReLU e variantes;
- Batch Normalization;
- Skip Connections (ResNets);
- LSTMs/GRUs

PROBLEMAS NO BACKPROPAGATION

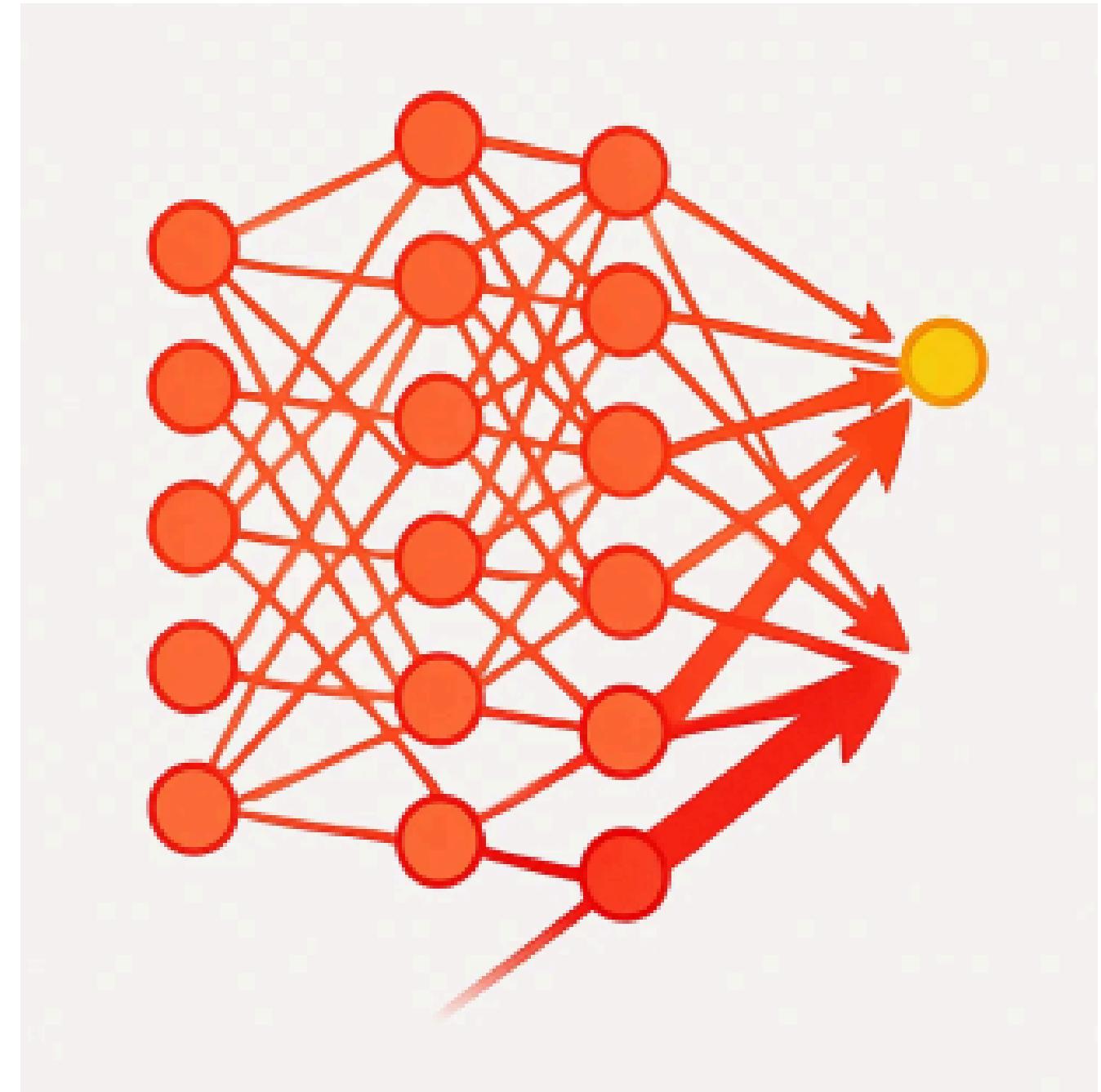
Explosão do Gradiente

- **O que é?**

- Os gradientes se tornam excessivamente grandes durante o Backpropagation, causando atualizações de peso instáveis e potencial divergência do treinamento.

- **Por que ocorre?**

- Pesos Iniciais Grandes: Multiplicação exponencial;
- Taxa de Aprendizado Alta: Amplifica gradientes;
- Redes Profundas: Acumulação de gradientes.



PROBLEMAS NO BACKPROPAGATION

Explosão do Gradiente

- **Consequências:**

- Flutuações erráticas na perda;
- Valores NaN ou Inf na perda;
- Pesos crescem descontroladamente;
- Instabilidade no treinamento.

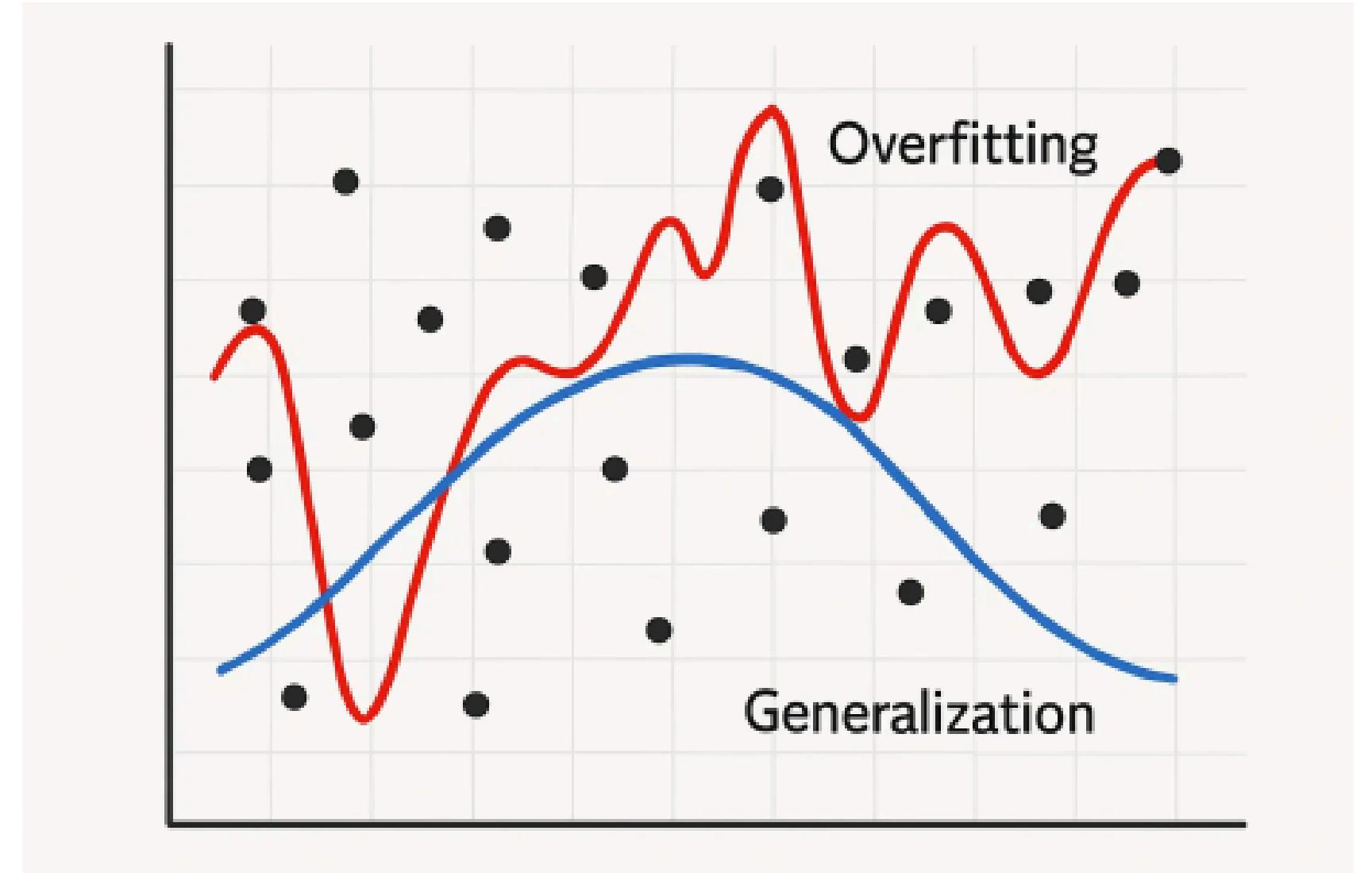
- **Soluções:**

- Gradient Clipping: Limita magnitude;
- Regularização L1/L2: Controla pesos;
- Taxa de Aprendizado Menor;
- Batch Normalization.

PROBLEMAS NO BACKPROPAGATION

Overfitting (sobreajuste)

- **O que é?**
 - O modelo aprende os dados de treinamento excessivamente, incluindo ruído e particularidades, perdendo a capacidade de generalizar para dados novos.
- **Por que ocorre?**
 - Treinamento Excessivo: Muitas épocas;
 - Modelos Complexos: Alta capacidade;
 - Dados Insuficientes: Memorização.



PROBLEMAS NO BACKPROPAGATION

Overfitting (sobreajuste)

- **Como Identificar?**

- Alta precisão no treino, baixa na validação;
- Perda de treino ↓ , perda de validação ↑ ;
- Diferença significativa de performance.

- **Soluções:**

- Regularização L1/L2: Penaliza complexidade;
- Dropout: Desativa neurônios aleatoriamente;
- Early Stopping: Para quando validação piora;
- Data Augmentation: Aumenta diversidade;
- Cross-Validation: Avalia generalização

PROBLEMAS NO BACKPROPAGATION

Sensibilidade a Hiperparâmetros

- **Por que ocorre?**

- Natureza Iterativa: Backpropagation ajusta pesos em pequenos passos controlados pelos hiperparâmetros;
- Superfície Complexa: Função de perda não-convexa com múltiplos mínimos locais;
- Interdependência: Hiperparâmetros interagem entre si de forma complexa.

- **Consequências:**

- Convergência Lenta/Divergência;
- Subajuste ou Sobreajuste;
- Desempenho Subótimo;
- Instabilidade no Treinamento.

- **Soluções:**

- Otimização de Hiperparâmetros: Grid Search, Random Search, Bayesian Optimization para encontrar combinações ideais de forma sistemática.

PROBLEMAS NO BACKPROPAGATION

Sensibilidade a Hiperparâmetros

Hiperparâmetros Críticos

- 01** Taxa de Aprendizado
O mais crítico - controla tamanho dos passos
- 02** Função de Ativação Impacta diretamente os gradientes
- 03** Otimizador
Estratégia de atualização dos pesos
- 04** Tamanho do Batch
Afeta estabilidade e convergência

O QUE SÃO HIPERPARÂMETROS?

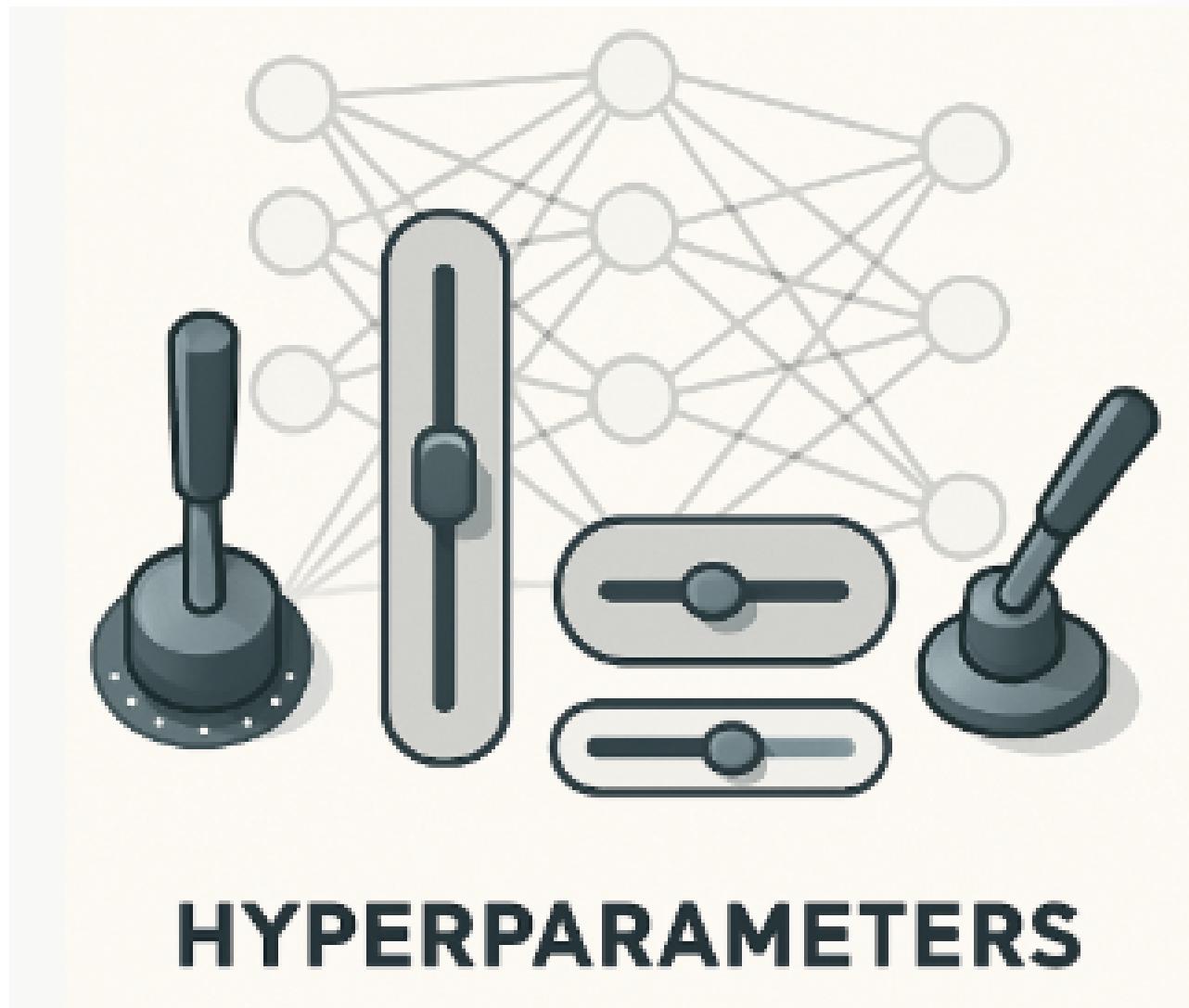
⚙ Definição:

- Variáveis de configuração externas que são definidas antes do treinamento e controlam como o modelo aprende;
- São as "ferramentas" que moldam o processo de aprendizado.

★ Por que são importantes?

- Desempenho: Impactam diretamente a precisão do modelo;
- Velocidade: Controlam rapidez do treinamento;
- Complexidade: Definem capacidade do modelo;
- Estabilidade: Influenciam robustez do treinamento.

Enquanto o Backpropagation é o motor que impulsiona o aprendizado, os hiperparâmetros são as alavancas de controle que guiam esse motor.



O QUE SÃO HIPERPARÂMETROS?

Parâmetros vs. Hiperparâmetros

PARÂMETROS DO MODELO

- Aprendidos a partir dos dados;
- Pesos e vieses dos neurônios;
- Ajustados automaticamente pelo;
- Backpropagation;
- Resultado do processo de treinamento.

HIPERPARÂMETROS

- Definidos pelo usuário antes do treino;
- Taxa de aprendizado, arquitetura • Controlam o processo de treinamento;
- Experimentação e ajuste manual.

PRINCIPAIS HIPERPARÂMETROS

Taxa de Aprendizado (η)

- Controla o tamanho do passo na atualização dos pesos a cada iteração.

↑ ALTA

- ✓ Convergência rápida;
- ✗ Oscilações;
- ✗ Pode divergir;
- ✗ "Salta" sobre mínimo.

↓ BAIXA

- ✓ Convergência estável;
- ✓ Precisão maior;
- ✗ Treinamento lento;
- ✗ Pode travar em mínimos locais.

- Estratégias de Escolha:
 - Experimentação (0.1, 0.01, 0.001);
 - Learning Rate Scheduling;
 - Otimizadores Adaptativos (Adam).

PRINCIPAIS HIPERPARÂMETROS

Número de Épocas

- Uma época = passagem completa de todo o dataset pela rede.

POCAS ÉPOCAS

- ✓ Treinamento rápido;
- ✓ Menor custo computacional;
- ✗ Subajuste (underfitting);
- ✗ Baixa precisão.

MUITAS ÉPOCAS

- ✓ Aprendizado refinado;
- ✓ Alta precisão no treino;
- ✗ Sobreajuste (overfitting);
- ✗ Maior tempo/custo.

- Técnica Principal:
 - Early Stopping:
 - Para o treinamento quando a performance na validação começa a piorar, evitando overfitting..

PRINCIPAIS HIPERPARÂMETROS

Tamanho do Batch

- Número de exemplos processados antes de atualizar os pesos.
 - Batch GD (Batch = Dataset): Gradiente preciso, mas lento para grandes datasets;
 - SGD (Batch = 1) Ruidoso, mas escapa de mínimos locais;
 - Mini-Batch ($1 < \text{Batch} < \text{Dataset}$) Equilíbrio ideal - mais comum na prática

BATCH GRANDE

- ✓ Gradientes estáveis;
- ✓ Melhor paralelização;
- ✗ Mais memória;
- ✗ Mínimos "planos".

BATCH PEQUENO

- Batch Pequeno;
- ✓ Melhor generalização;
- ✓ Menos memória;
- ✗ Gradientes ruidosos ;
- ✗ Convergência instável.

PRINCIPAIS HIPERPARÂMETROS

Arquitetura da Rede

- Número de camadas e neurônios por camada definem a capacidade do modelo.

REDE SIMPLES

- ✓ Menos overfitting;
- ✓ Treinamento rápido;
- ✗ Pode ser underfitting;
- ✗ Capacidade limitada.

REDE COMPLEXA

- ✓ Alta capacidade;
- ✓ Padrões complexos;
- ✗ Risco de overfitting;
- ✗ Mais dados necessários.

- Estratégias de Escolha:
 - Complexidade do Problema: Mais complexo = mais camadas;
 - Volume de Dados: Mais dados = pode usar redes maiores;
 - Experimentação: Começar simples e aumentar gradualmente;
 - Arquiteturas Predefinidas: ResNet, VGG, Transformer.

PRINCIPAIS HIPERPARÂMETROS

Função de Ativação

- Introduz não-linearidade na rede, permitindo aprender relações complexas.

⚠ SIGMOIDE E HIPERBÓLICA

- Historicamente populares;
 - Problema: Desvanecimento do gradiente;
 - Gradientes pequenos em valores extremos;
 - Sigmoid: [0,1], Tanh: [-1,1].
-
- Escolha Prática:
 - Camadas Ocultas: ReLU (padrão);
 - Saída Binária: Sigmoide;
 - Saída Multiclasse: Softmax.

RELU E VARIANTES

- ReLU: $f(x) = \max(0, x)$;
- Mitiga desvanecimento do gradiente;
- Problema: Neurônios "mortos";
- Variantes: Leaky ReLU, ELU, PReLU

PRINCIPAIS HIPERPARÂMETROS

Otimizador

- Algoritmo que atualiza os pesos com base nos gradientes calculados.
- SGD: Simples, mas pode ser lento e oscilar;
- Adam: Adaptativo, robusto, escolha padrão;
- RMSprop: Bom para gradientes esparsos;
- Adagrad: Adapta taxa por parâmetro.
- **Características do Adam:**
 - Combina Adagrad + RMSprop;
 - Médias móveis dos gradientes;
 - Menos sensível à taxa de aprendizado;
 - Converge mais rápido.

PRINCIPAIS HIPERPARÂMETROS

Regularização L1/L2

- Adiciona termo de penalidade à função de perda para desencorajar pesos grandes.
- **Regularização L1 (Lasso):**
 - Penalidade: $\sum |w_i|$;
 - Efeito: Alguns pesos $\rightarrow 0$;
 - Resultado: Seleção de características;
 - Modelo: Esparsos.
- **Regularização L2 (Ridge):**
 - Penalidade: $\sum w_i^2$;
 - Efeito: "Encolhe" pesos proporcionalmente;
 - Resultado: Pesos menores;
 - Mais comum em redes neurais.
- **Taxa de Regularização (λ):**
 - λ Alto: Forte penalidade, risco de underfitting;
 - λ Baixo: Penalidade suave, risco de overfitting.

PRINCIPAIS HIPERPARÂMETROS

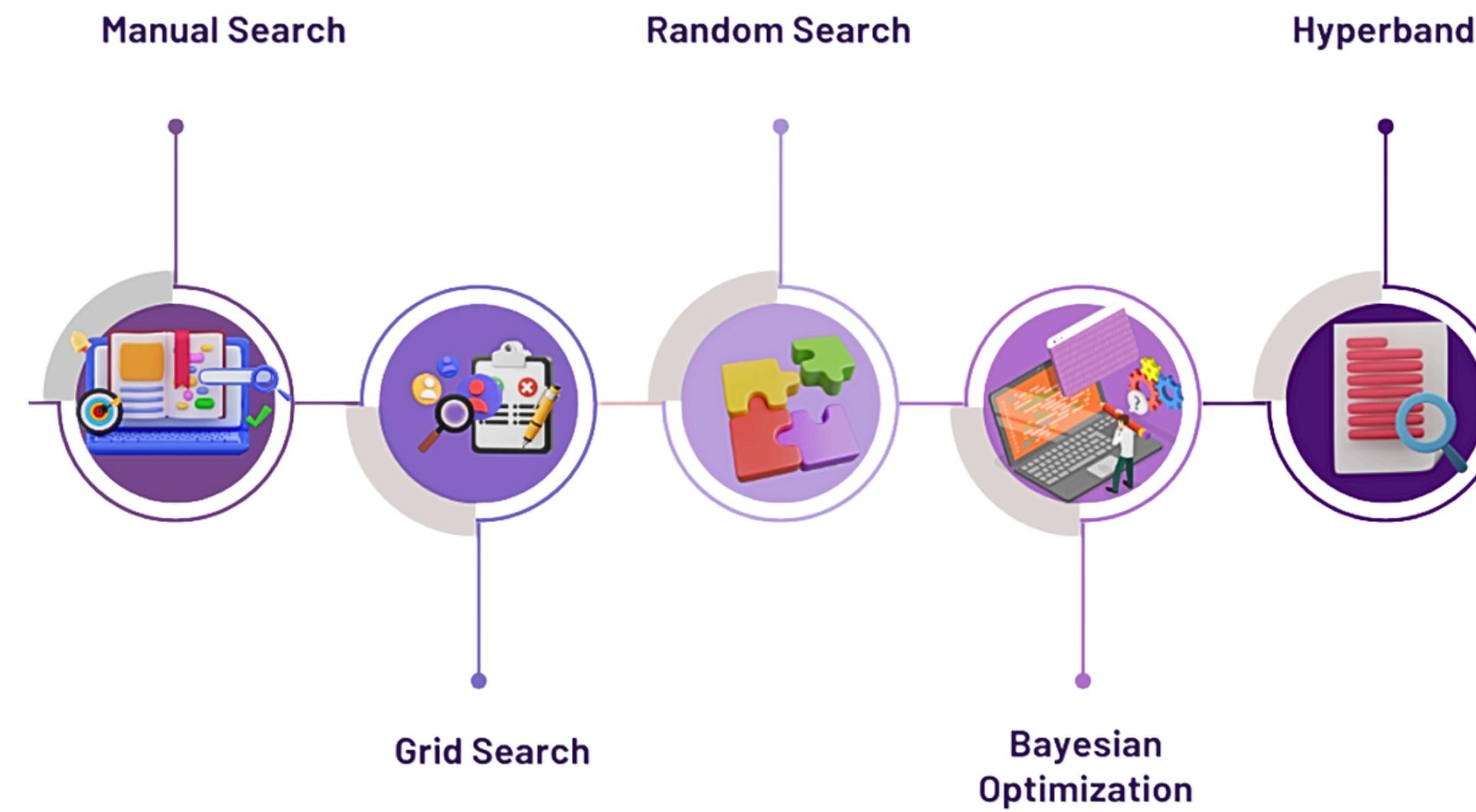
Dropout

- "Desliga" aleatoriamente um percentual de neurônios durante o treinamento.

⚙ Como Funciona?

- A cada iteração, neurônios diferentes são desativados;
 - Força a rede a não depender de neurônios específicos;
 - Apenas no treinamento - teste usa todos;
 - Pesos são escalados durante inferência.
-
- ↑ **Taxa Alta (ex: 0.7):**
 - ✓ Forte regularização;
 - ✗ Risco de underfitting.
 - ↓ **Taxa Baixa (ex: 0.2):**
 - ✓ Preserva capacidade;
 - ✗ Menos regularização.
-
- **Valores Típicos:**
 - Camadas Ocultas: 0.5;
 - Camada de Entrada: 0.2;
 - Experimentar: 0.2 - 0.5.

Técnica para Ajustes Hiperparâmetros



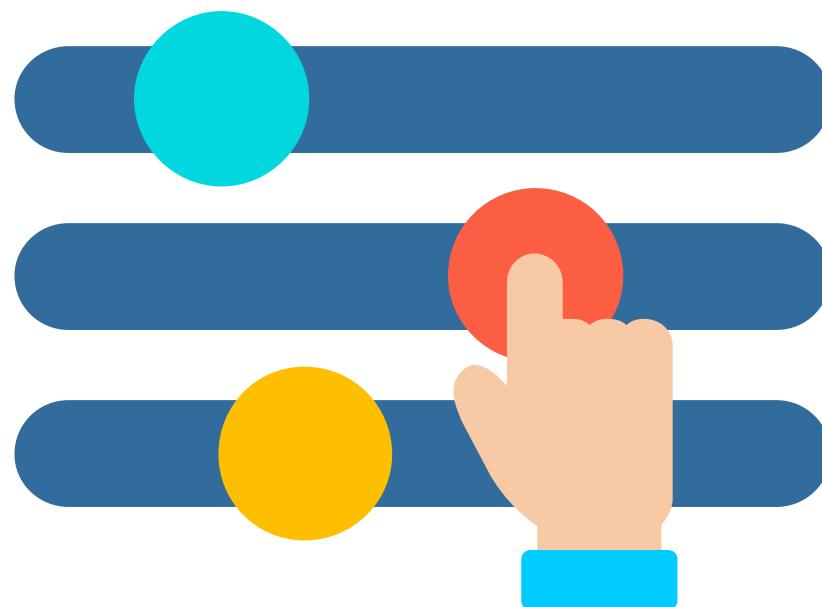
Fonte: AllAboutAi

Técnicas para Ajustes Hiperparâmetros

Manual Search

8.1.1 Definição

- Procedimento a qual utiliza sua experiência, intuição e análises de execuções anteriores para decidir, de **FORMA SUBJETIVA**, quais hiperparâmetros testar em seguida;
- Nesta abordagem, o "motor de otimização" é o próprio ser humano.



Técnicas para Ajustes Hiperparâmetros

Manual Search

8.1.2 Vantagens

- Desenvolvimento baseado na Intuição;
- Flexibilidade;
- Custo de Configuração Zero;
- Potencialmente Rápido (com experiência).

8.1.3 Desvantagens

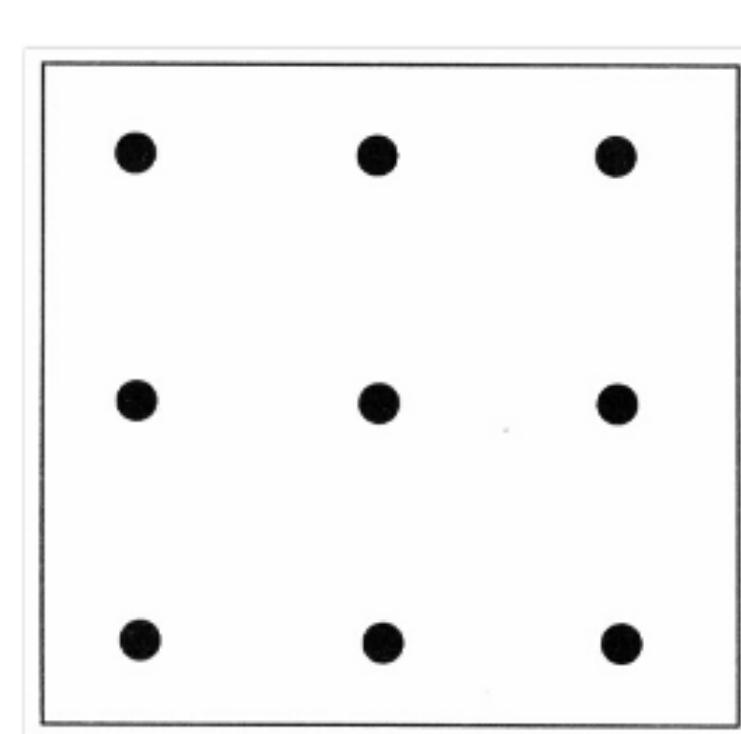
- Não é Cientificamente Reprodutível;
- Propenso a Vieses (Bias);
- Extremamente Lento e Tedioso;
- Não Escala.

Técnicas para Ajustes Hiperparâmetros

Grid Search

8.2.1 Definição

Avalia **TODAS AS COMBINAÇÕES** de hiperparâmetros a partir de uma grade de valores predefinida manualmente.



Fonte: DataCamp

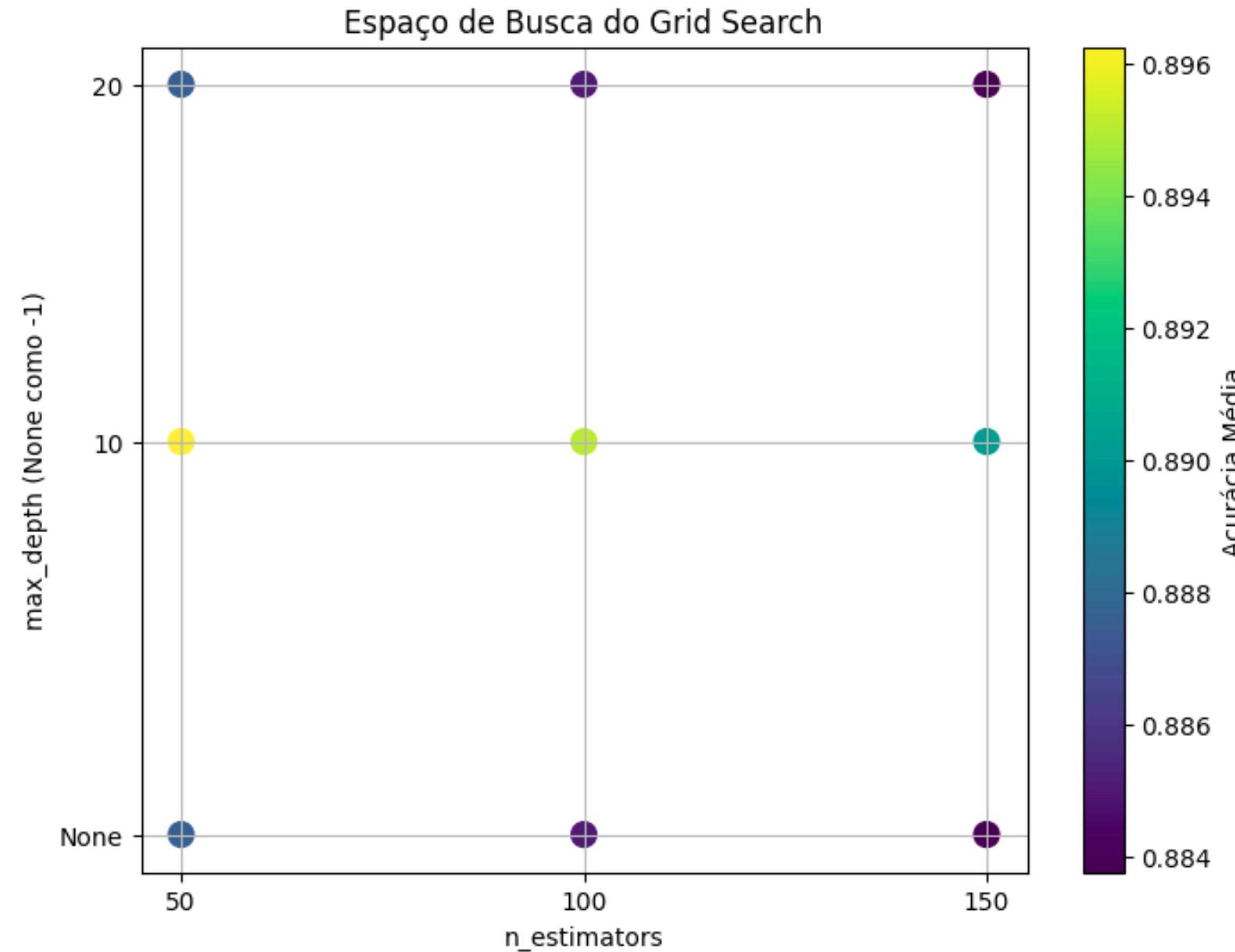
8.2.2 Fundamentação Matemática

- Considere dois hiperparâmetros, λ_1 com os valores $\{a, b\}$ e λ_2 com os valores $\{c, d, e\}$;
- O Grid Search analisa o produto cartesiano dos conjuntos de valores:
- Combinações = $\Lambda_1 \times \Lambda_2 = \{(a, c), (a, d), (a, e), (b, c), (b, d), (b, e)\}$.

Técnicas para Ajustes Hiperparâmetros

Grid Search

8.2.3 Representação Gráfica



Fonte: Autoria Própria

Técnicas para Ajustes Hiperparâmetros

Grid Search

8.2.4 Vantagens

- Simplicidade;
- Paralelização.

8.2.5 Desvantagens

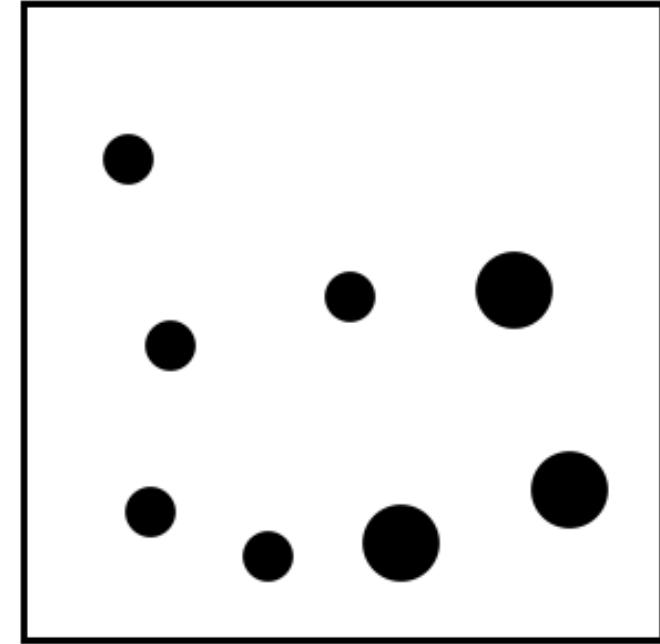
- Custo Computacional;
- Maldição da Dimensionalidade;
- Definição da Grade.

Técnicas para Ajustes Hiperparâmetros

Random Search

8.3.1 Definição

Avalia **UM NÚMERO FIXO** de combinações aleatoriamente a partir de distribuições estatísticas.



Fonte: DataCamp

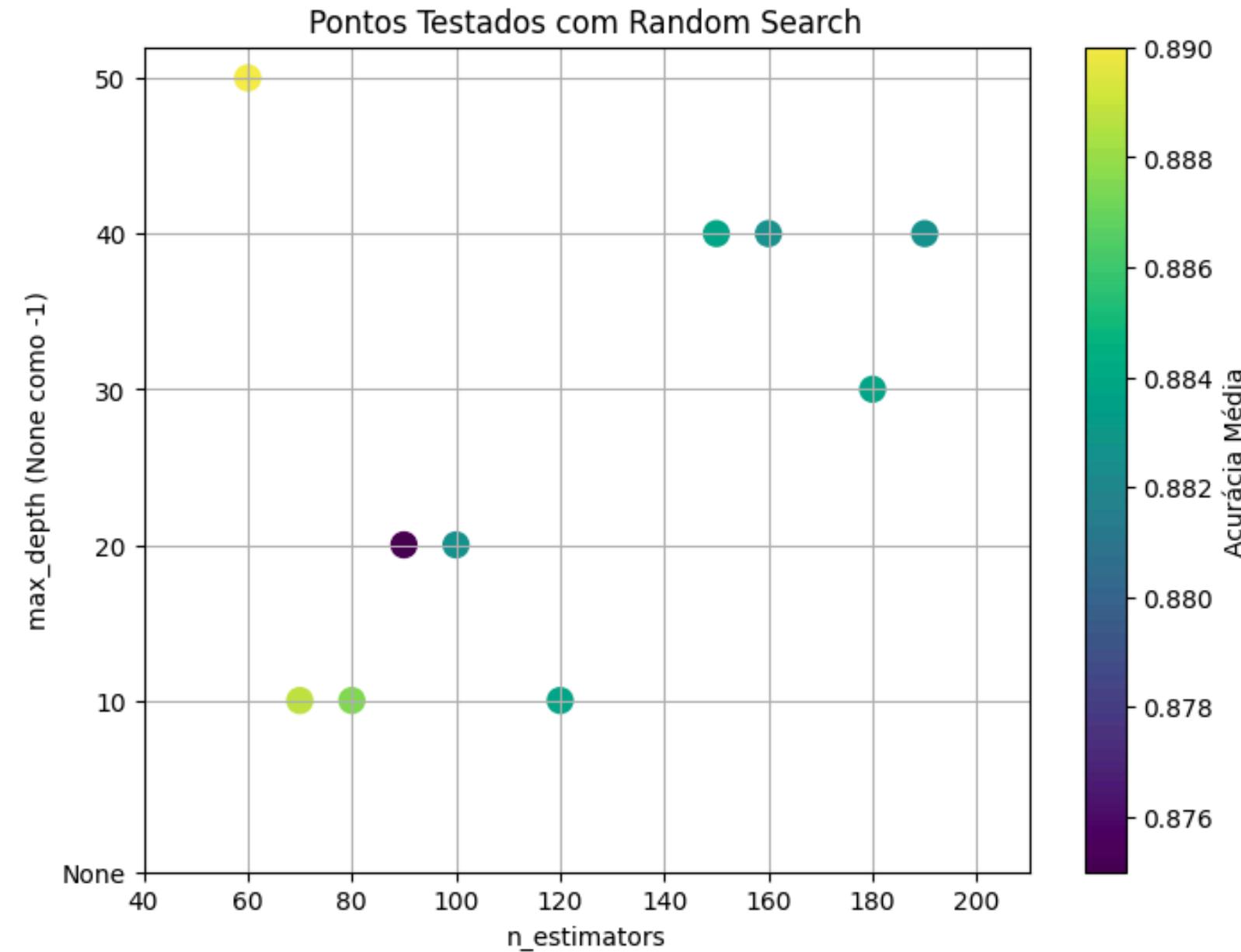
8.3.2 Fundamentação Matemática

- Para cada avaliação, uma configuração λ é extraída de uma distribuição de probabilidade que abrange o espaço de hiperparâmetros Λ ;
- $\lambda(i) \sim p(\lambda | \Lambda)$.

Técnicas para Ajustes Hiperparâmetros

Random Search

8.3.3 Representação Gráfica



Fonte: Autoria Própria

Técnicas para Ajustes Hiperparâmetros

Random Search

8.3.4 Vantagens

- Flexibilidade e Eficiência;
- Controle de Custo;
- Fácil de Paralelizar;
- Garantias Teóricas.

8.3.5 Desvantagens

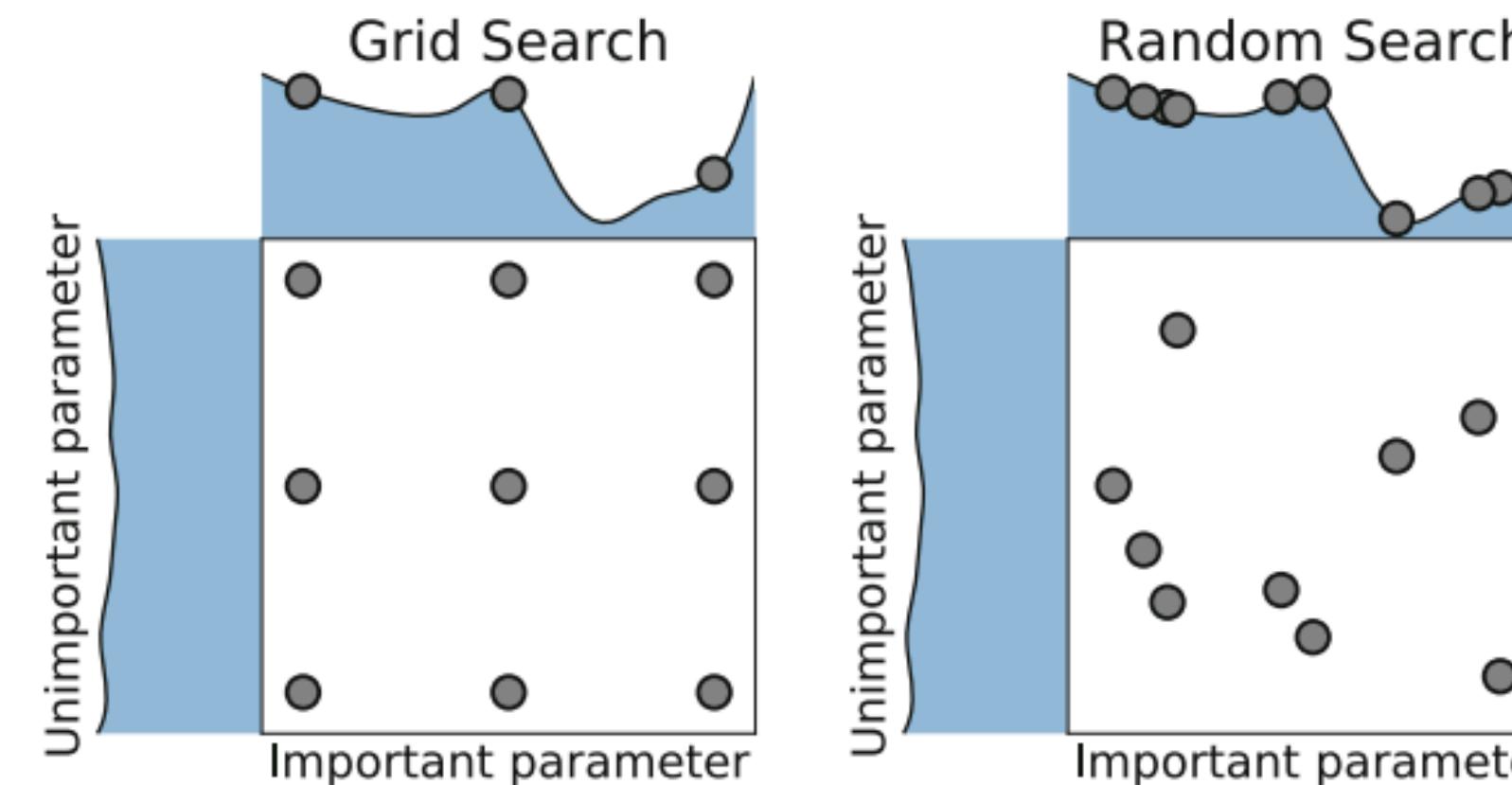
- Não aprende com o processo;
- Ainda pode ser caro.

Técnicas para Ajustes Hiperparâmetros

Random Search

8.3.6 Comparação com o Grid Search

Fig. 1.1 Comparison of grid search and random search for minimizing a function with one important and one unimportant parameter. This figure is based on the illustration in Fig. 1 of Bergstra and Bengio [13]



Fonte: HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (Ed.). Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer, 2019.

Técnicas para Ajustes Hiperparâmetros

Bayesian Optimization

8.4.1 Definição

Técnica de otimização sequencial e baseada em modelo. Ela "aprende" sobre a função objetivo (desempenho vs. hiperparâmetros) a cada avaliação e usa esse aprendizado para escolher a próxima combinação a ser testada.

8.4.2 Componentes Chave

- **Modelo Substituto (Surrogate Model):** Uma aproximação probabilística da função objetivo.
Ex.: Processos Gaussianos (GP);
- **Função de Aquisição (Acquisition Function):** Define qual ponto deve ser avaliado em seguida, equilibrando exploração (exploration) e exploração (exploitation).
Ex: Expected Improvement (EI).

Técnicas para Ajustes Hiperparâmetros

Bayesian Optimization

8.4.3 Fundamentação Matemática (Simplificada)

- **Processo Gaussiano (GP):** Modela a função $f(\lambda)$ como uma distribuição sobre funções:

$$f(\lambda) \sim GP(m(\lambda), k(\lambda, \lambda'))$$

$m(\lambda)$: função média;

$k(\lambda, \lambda')$: kernel de covariância (semelhança).

- **Expected Improvement (EI):** Determina o ganho esperado ao selecionar um novo ponto λ , em comparação com o melhor valor observado F_{\min} .

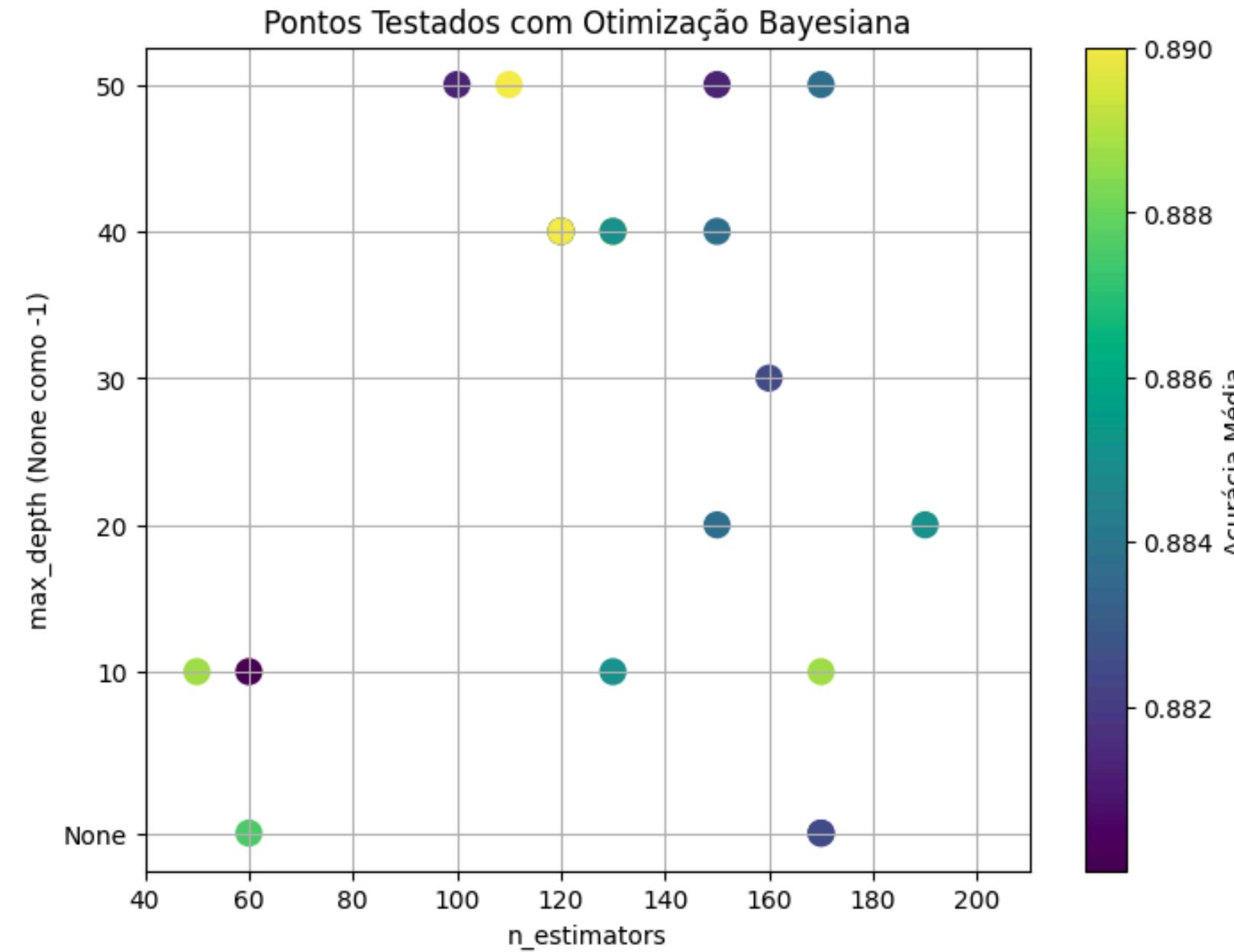
$$EI(\lambda) = E[\max(f_{\min} - y, 0)]$$

O objetivo é maximizar $EI(\lambda)$ para decidir qual configuração deve ser avaliada a seguir.

Técnicas para Ajustes Hiperparâmetros

Bayesian Optimization

8.4.4 Representação Gráfica



Fonte: Autoria Própria

Técnicas para Ajustes Hiperparâmetros

Bayesian Optimization

8.4.5 Vantagens

- Eficiência de Dados;
- Busca Inteligente;
- Desempenho Superior.

8.4.6 Desvantagens

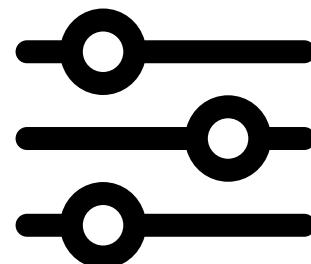
- Complexidade de Implementação;
- Custo de Overhead;
- Dificuldade de Paralelização;
- Risco de Mínimos Locais.

Técnicas para Ajustes Hiperparâmetros

Hyperband

8.5.1 Definição

Uma abordagem de alta velocidade baseada no conceito de Successive Halving (Metade Sucessiva). Ele aloca um orçamento (ex: épocas, subconjunto de dados) a um grande número de configurações e elimina iterativamente as de pior desempenho, focando os recursos nas mais promissoras.



Técnicas para Ajustes Hiperparâmetros

Hyperband

8.5.2 Como Funciona (Algoritmo)

P1 - Começa com n configurações aleatórias;

P2 - Treina todas por um recurso mínimo r (ex: 1 época);

P3 - Elimina a metade com pior desempenho;

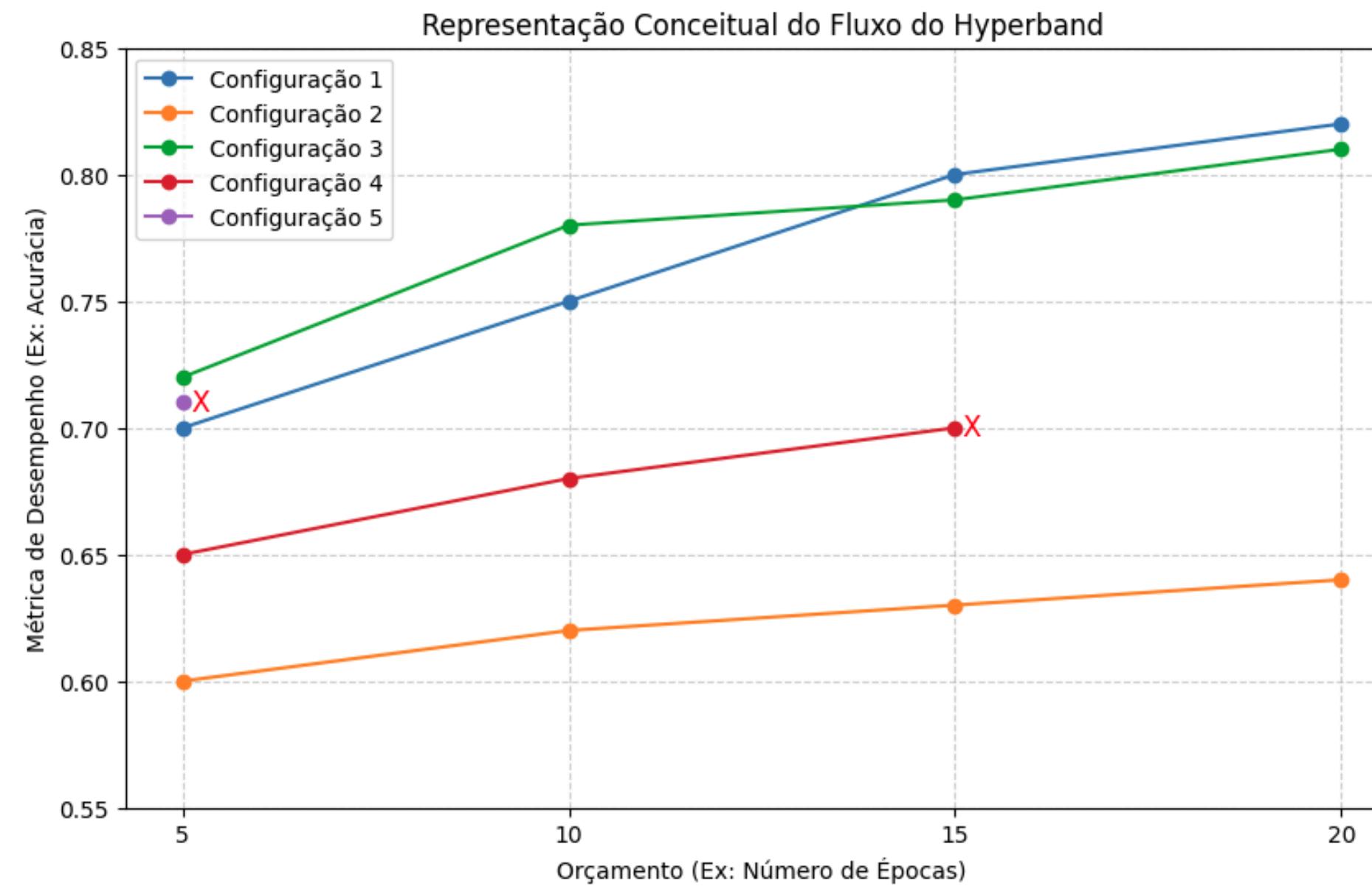
P4 - Dobra o recurso ($2r$) para as sobreviventes e repete até que reste uma única configuração;

P5 O Hyperband repete esse processo com diferentes valores iniciais de n para mitigar o dilema "poucas configurações bem treinadas vs. muitas mal treinadas".

Técnicas para Ajustes Hiperparâmetros

Hyperband

8.5.3 Representação Conceitual



Fonte: Autoria Própria

Técnicas para Ajustes Hiperparâmetros

Hyperband

8.5.4 Vantagens

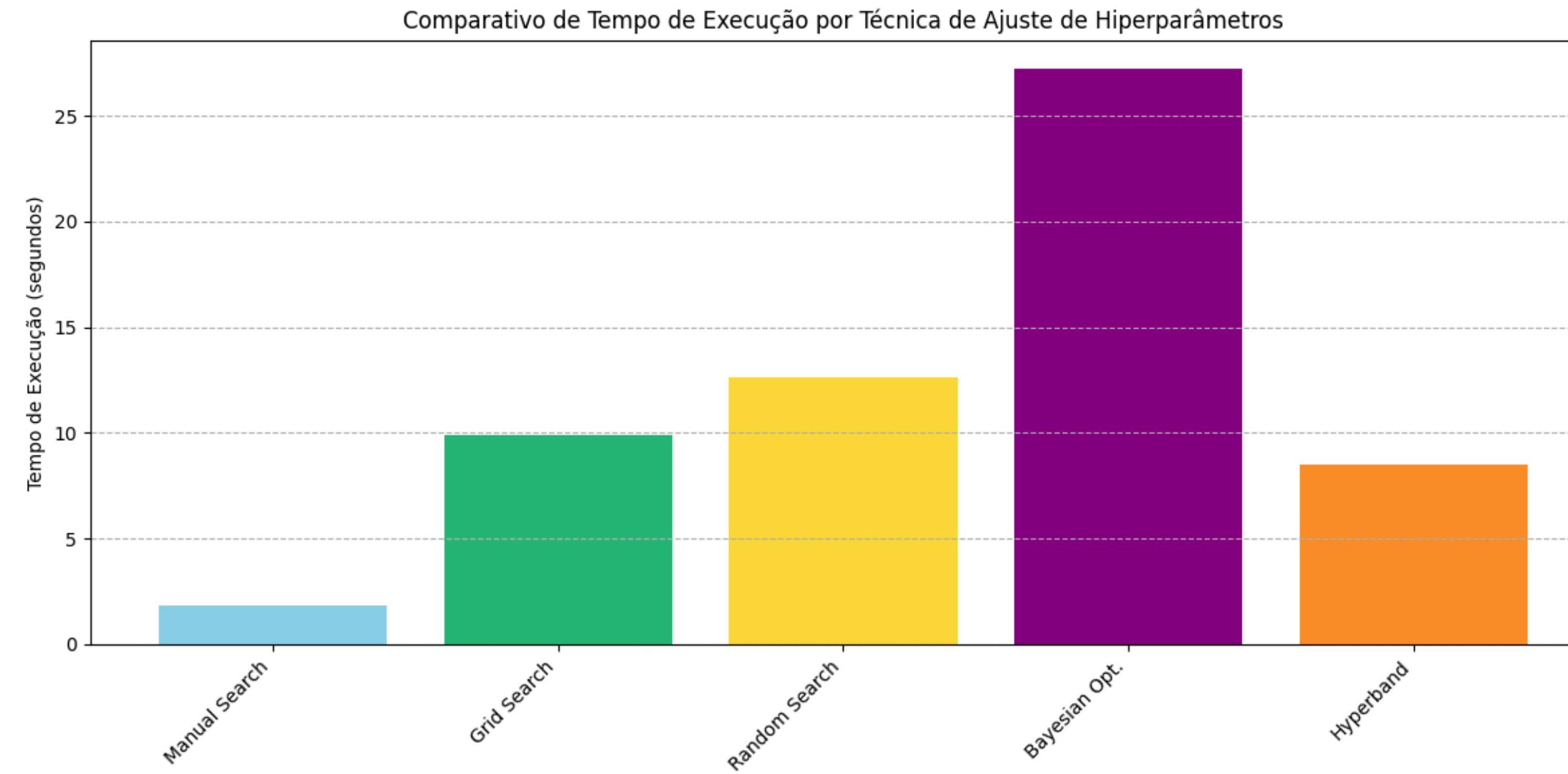
- Velocidade e Eficiência;
- Simplicidade Conceitual;
- Ideal para Deep Learning.

8.5.5 Desvantagens

- Não "Aprende" a Gerar Configurações;
- Pode Eliminar "Azarões";
- Requer Múltiplas Fidelidades.

Técnicas para Ajustes Hiperparâmetros

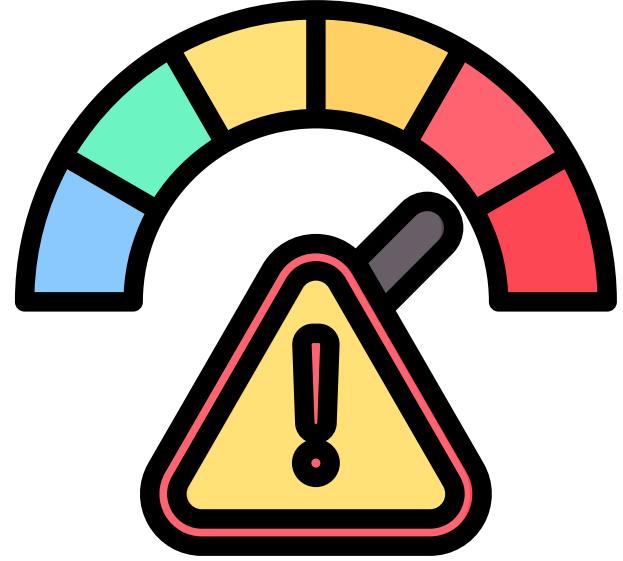
Comparativo das Técnicas



Fonte: Autoria Própria

Impacto dos Hiperparâmetros

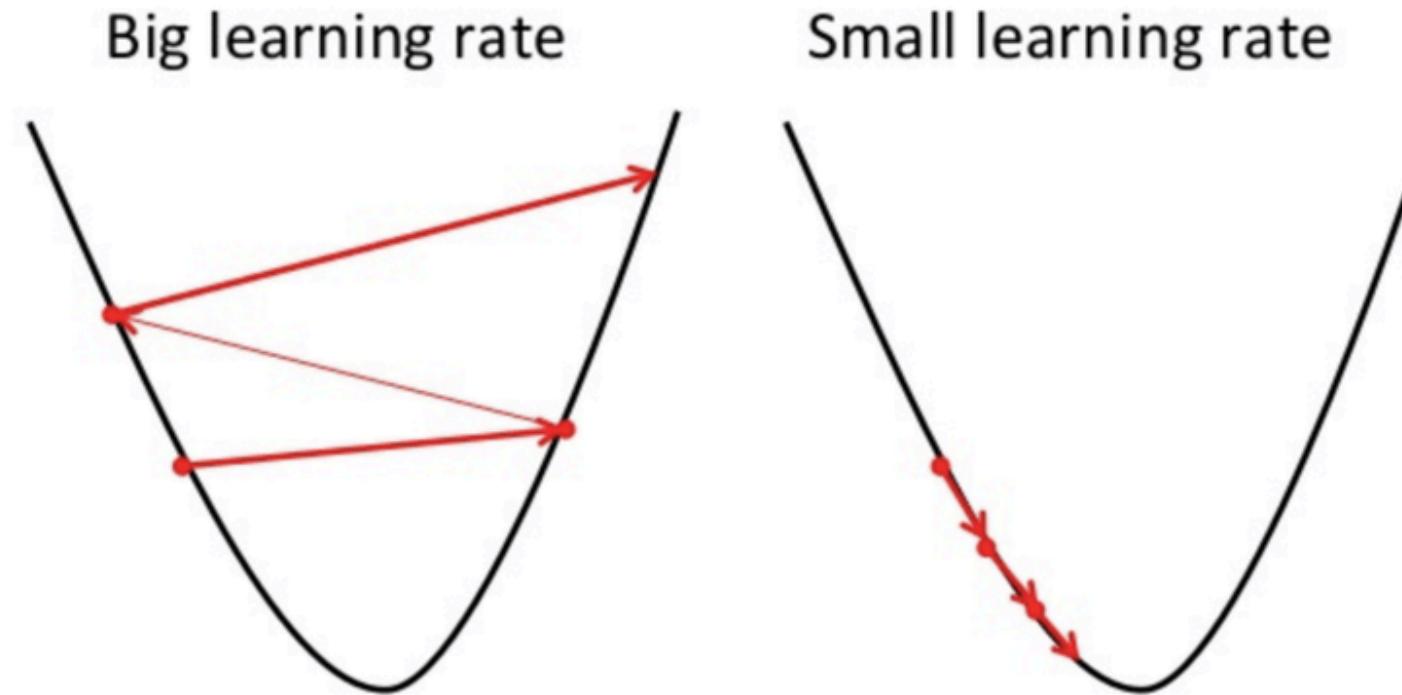
- Taxa de Aprendizado (Learning Rate, α);
- Número de Neurônios e Camadas (Capacidade do Modelo);
- Tamanho do Batch (Batch Size).



Impacto dos Hiperparâmetros

Taxa de Aprendizado (Learning Rate, α)

- O tamanho do passo que o otimizador dá na direção do gradiente negativo.
- Com α muito alta, o otimizador pode divergir (explodir) ou oscilar em torno do mínimo sem conseguir convergir;
- Com α muito baixa, a convergência extremamente lenta, risco de parar em um mínimo local ruim.

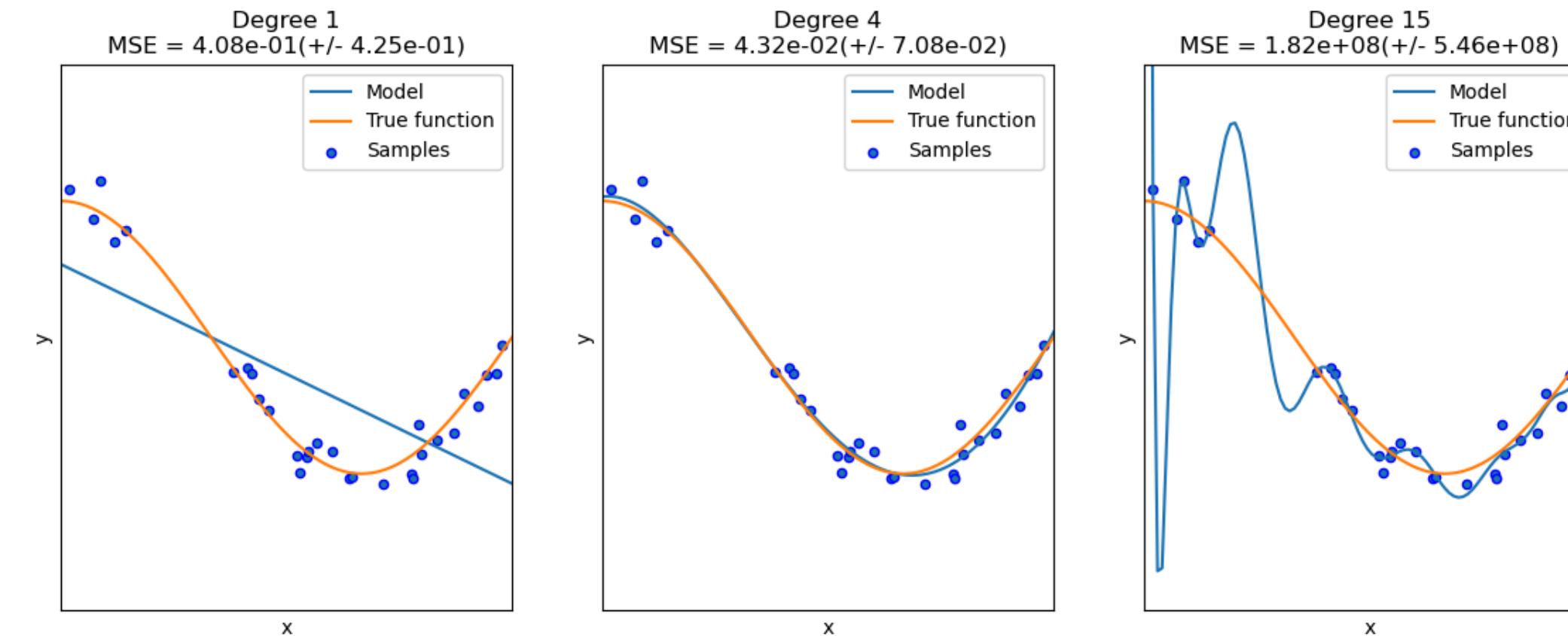


Fonte: ResearchGate

Impacto dos Hiperparâmetros

Número de Neurônios e Camadas (Capacidade do Modelo)

- Determina a complexidade e a capacidade de aprendizado da rede;
- Com Poucos neurônios/camadas (baixa capacidade), pode levar ao underfitting;
- Com Muitos neurônios/camadas (alta capacidade), pode levar ao overfitting;

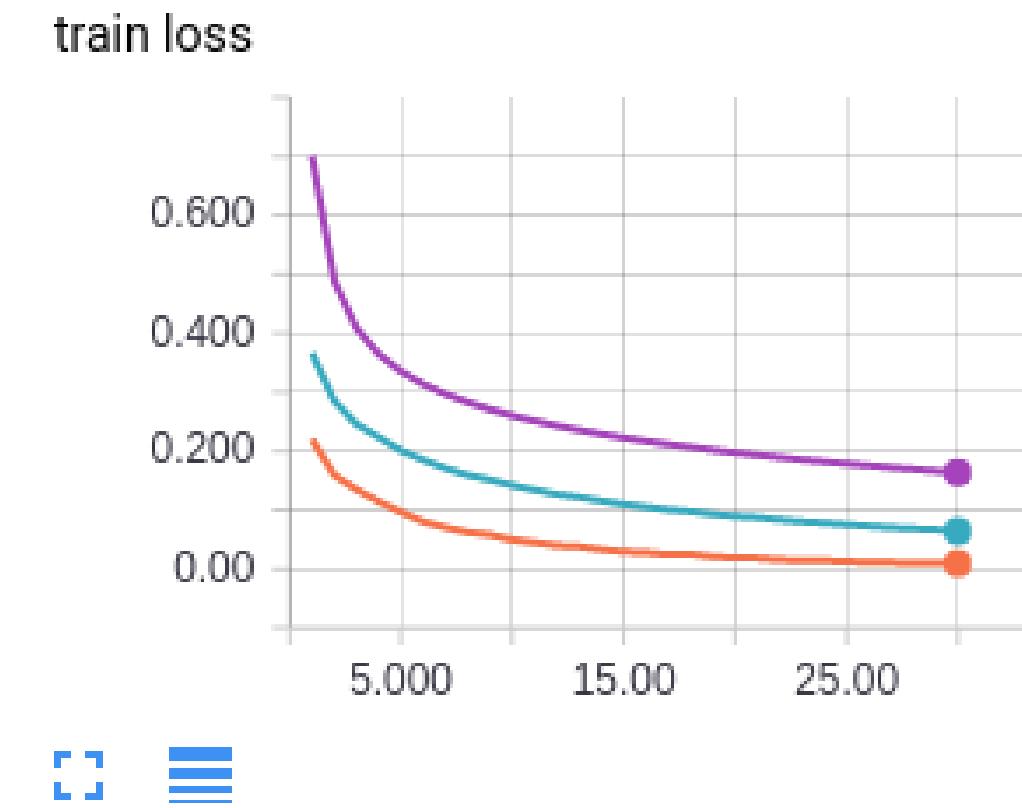
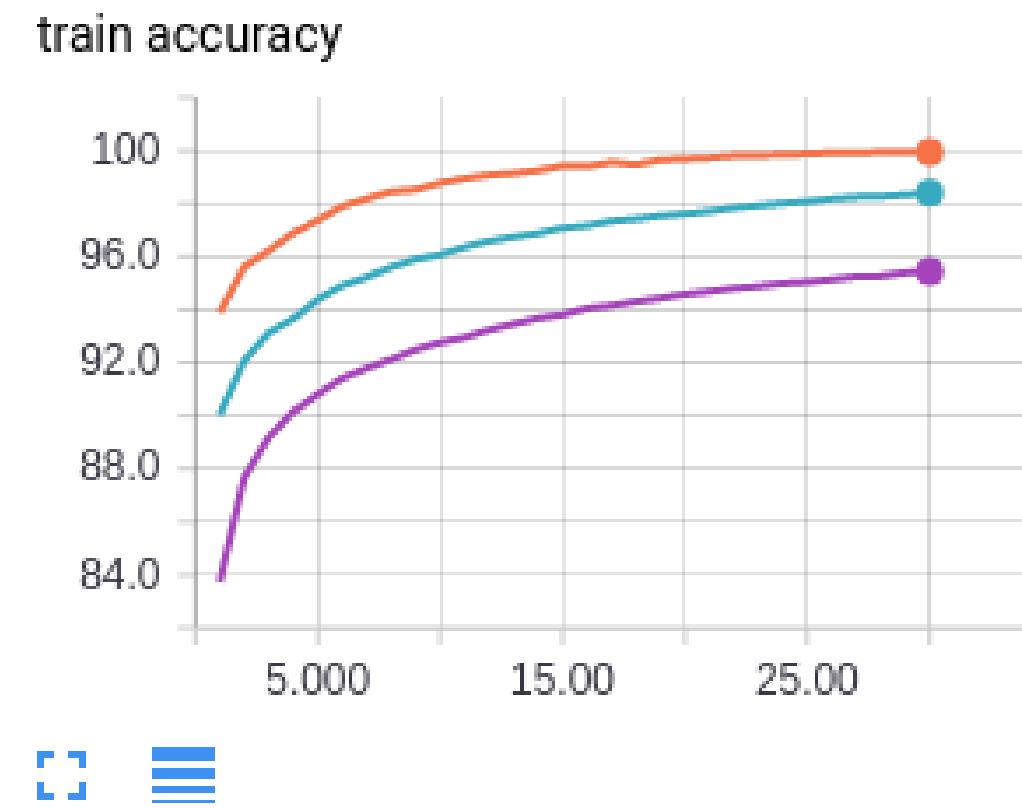


Fonte: Scikit-learn.

Impacto dos Hiperparâmetros

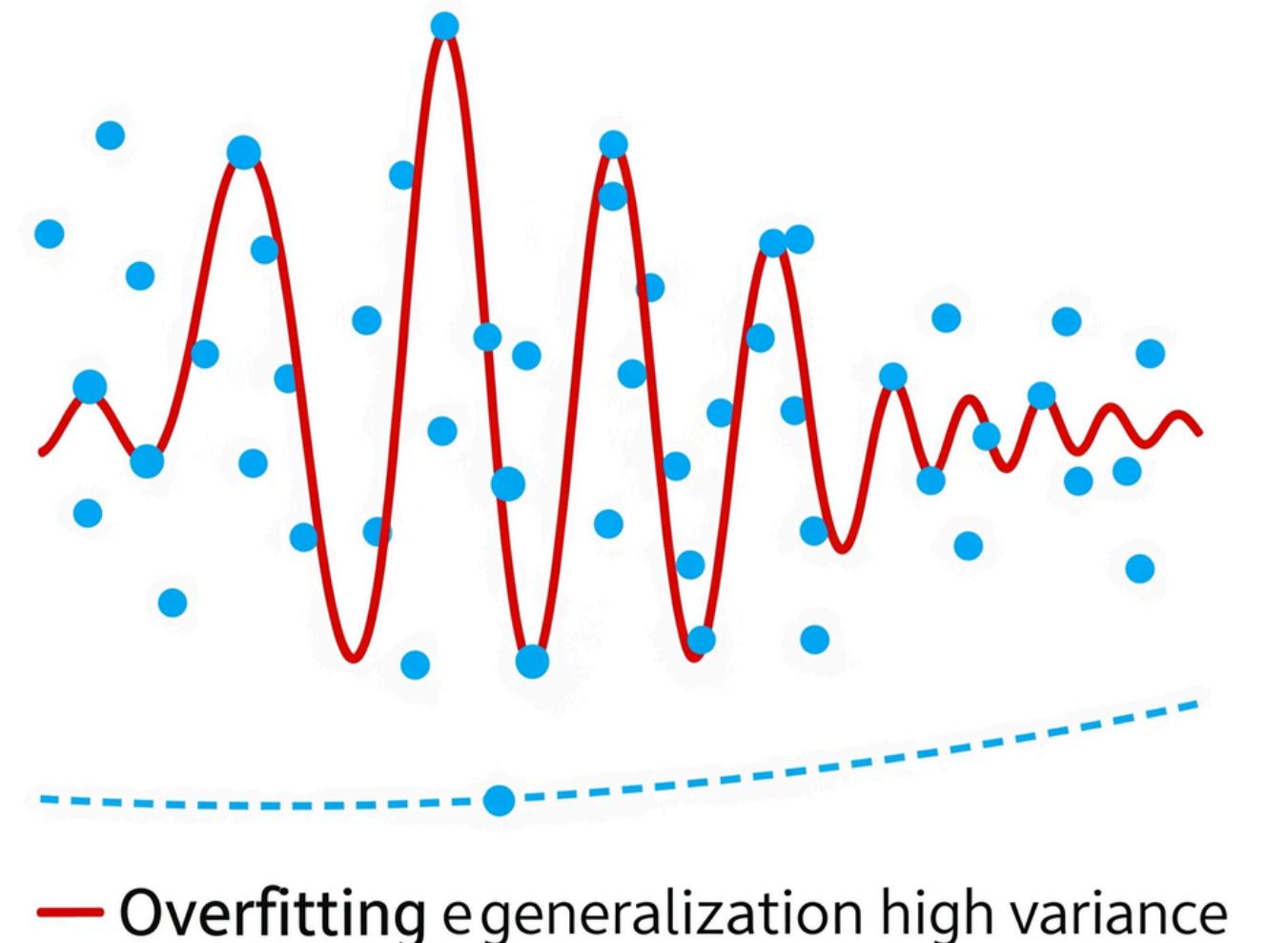
Tamanho do Batch (Batch Size)

- Número de exemplos de treino utilizados em uma única iteração (atualização dos pesos);
- Lotes pequenos levam a atualizações mais ruidosas, mas podem ajudar na generalização;
- Lotes grandes são computacionalmente eficientes, mas podem convergir para mínimos "mais nítidos" (pior generalização).



Fonte: Deep Learning Book

Técnicas para Evitar Overfitting Hiperparâmetros

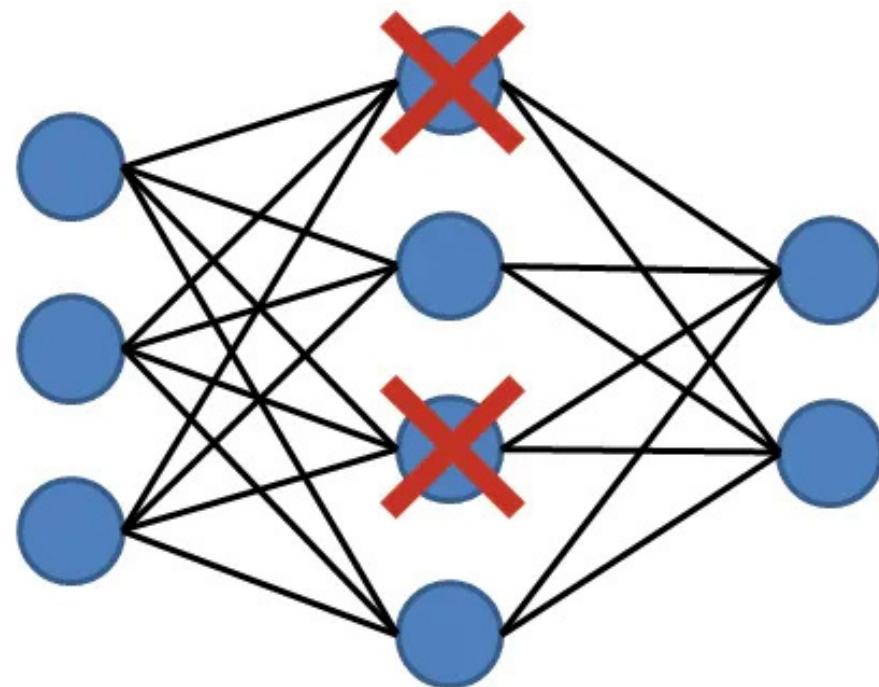


Fonte: Reaserch Gate

Técnicas para Evitar Overfitting (Hiperparâmetros)

Dropout

- Uma técnica de regularização onde a contribuição de neurônios é temporariamente removida da rede durante o treino;
- O Hiperparâmetro Principal é a Taxa de Dropout.

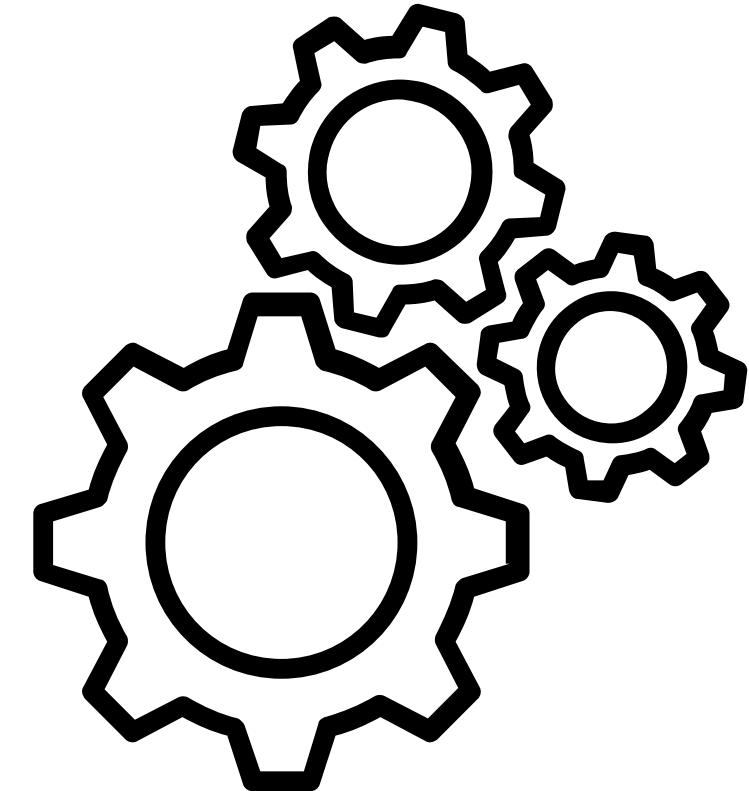


Fonte: Medium

Técnicas para Evitar Overfitting (Hiperparâmetros)

Regularização L1/L2

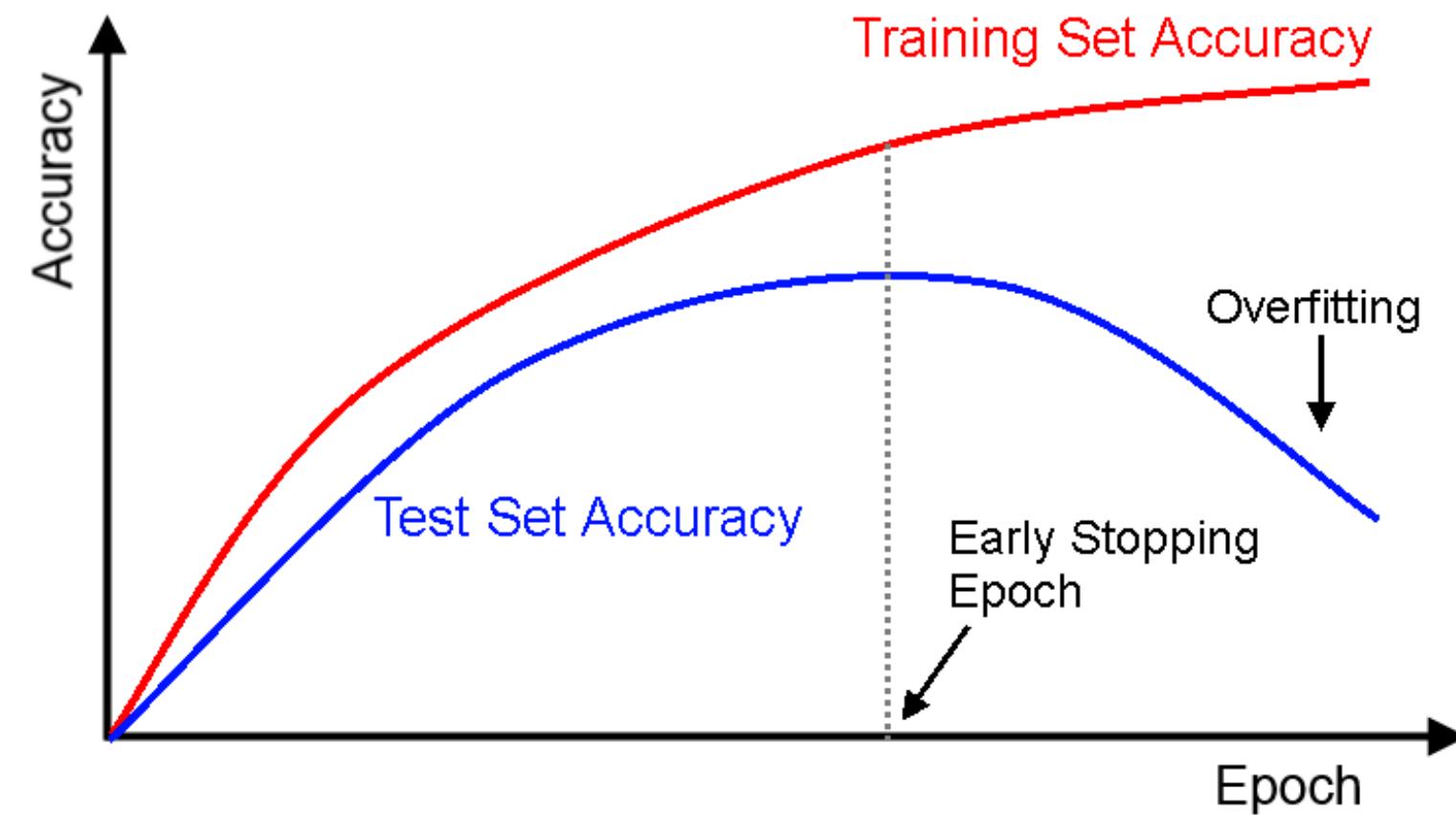
- Adiciona um termo de penalidade à função de perda, cuja intensidade é controlada por um hiperparâmetro;
- O Hiperparâmetro Principal é a Força da Regularização (λ , lambda).



Técnicas para Evitar Overfitting (Hiperparâmetros)

Early Stopping (parada antecipada)

- Técnica que interrompe o treinamento quando a performance em um conjunto de validação para de melhorar;
- O Hiperparâmetro Principal é a "Paciência" (Patience).



Fonte: Deep Learning Book

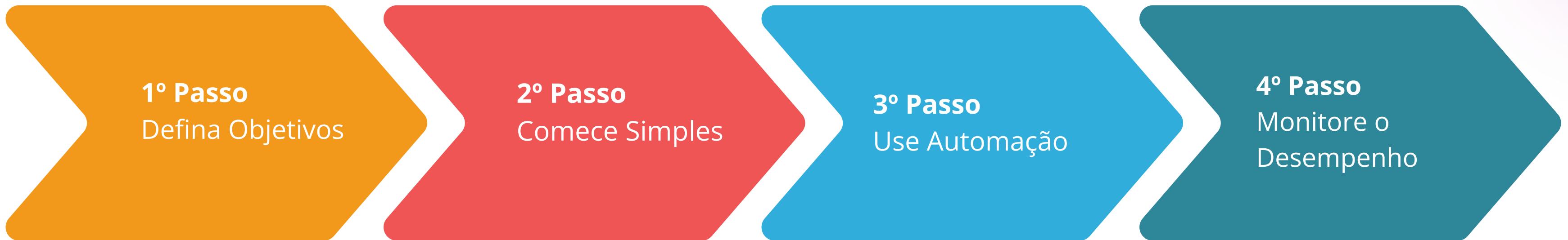
Por que Otimizar Hiperparâmetros é Importante?

- Impacto no Desempenho;
- Underfitting vs. Overfitting;
- Eficiência;
- Obtenção do Estado da Arte;

APLICAÇÕES DOS AJUSTES DE HIPERPARÂMETRO

- Classificação de Imagens;
- Processamento de Linguagem Natural (PLN);
- Análise Preditiva.

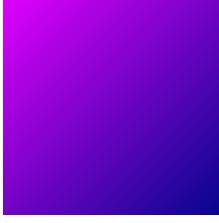
Melhores Práticas para o ajuste de Hiperparâmetros





Atividade

- **Atividade: Visualização do Desvanecimento do Gradiente.**



Referências

- HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. **Automated machine learning: methods, systems, challenges.** 1. ed. Cham: Springer, 2019. ISBN 978-3-030-05318-5.