

Luiz Guilherme Moraes da Costa Faria

APRENDIZADO DE MÁQUINA

Brasília, DF
20 de setembro de 2025

Luiz Guilherme Moraes da Costa Faria

APRENDIZADO DE MÁQUINA

Universidade de Brasília

Orientador: Nome do Orientador/Revisor (se aplicável)

Brasília, DF
20 de setembro de 2025

Sumário

Sumário	3
I	HISTÓRIA DA IA E DO COMPUTADOR 9
1	UMA BREVE HISTÓRIA DO COMPUTADOR 11
2	UMA BREVE HISTÓRIA DA INTELIGÊNCIA ARTIFICIAL . . . 13
II	CONCEITOS MATEMÁTICOS 15
3	CÁLCULO PARA APRENDIZADO DE MÁQUINA 17
3.1	Funções: A Base do Cálculo 17
3.2	Derivadas Ordinárias 17
3.3	Integrais Simples 17
3.4	Derivadas Parciais 17
4	ÁLGEBRA LINEAR PARA APRENDIZADO DE MÁQUINA . . . 19
4.1	A Unidade Fundamental: Vetores e Espaços Vetoriais 19
4.2	Organizando Dados: Matrizes e Suas Operações 19
4.3	Tensores: A Estrutura de Dados do Deep Learning 19
4.4	Resolvendo Sistemas e Encontrando Propriedades: Autovalores e Autovetores 19
4.5	Decomposição de Matrizes (SVD e PCA) 19
5	PROBABILIDADE E ESTATÍSTICA PARA APRENDIZADO DE MÁQUINA 21
5.1	Medindo a Incerteza: Probabilidade Básica e Condicional 21
5.2	O Teorema de Bayes: Aprendendo com Evidências 21
5.3	Descrevendo os Dados: Estatística Descritiva: Média, mediana, variância, desvio padrão 21
5.4	Variáveis Aleatórias e Distribuições de Probabilidade 21
5.5	A Função de Máxima Verossimilhança (Maximum Likelihood Estimation - MLE) 21

III	PILARES DAS REDES NEURAIS	23
6	O ALGORITMO DA REPROPROPAGAÇÃO E OS OTIMIZADORES BASEADOS EM GRADIENTE	25
6.1	O Método do Gradiente Descendente	25
6.1.1	Exemplo Ilustrativo	25
6.1.2	O Método em Si	25
6.1.3	Implementação em Python	25
6.2	A Retropropagação: Aprendendo com os Erros	25
6.3	Otimizadores Baseados em Gradiente	25
6.3.1	Método do Gradiente Estocástico	25
6.3.2	Método do Gradiente com Momentum	25
6.3.3	Nesterov	25
6.3.4	AdaGrad	25
6.3.5	RMSProp	25
6.3.6	Adam	25
6.3.7	Nadam	25
6.4	O Método de Newton: Indo Além do Gradiente	25
7	FUNÇÕES DE ATIVAÇÃO SIGMOIDAIS	27
7.1	Teoremas da Aproximação Universal	27
7.2	Exemplos Ilustrativo	27
7.3	A Sigmoid Logística	27
7.4	Tangente Hiperbólica	27
7.5	Softsign: Uma Sigmoidal Mais Barata	27
7.6	Hard Sigmoid e Hard Tanh: O Sacrifício da Suavidade em Prol do Desempenho	27
7.7	O Desaparecimento de Gradientes	27
7.8	Comparativo de Desempenho das Sigmoidais	27
8	FUNÇÕES DE ATIVAÇÃO RETIFICADORAS	29
8.1	Exemplo Ilustrativo	29
8.2	Rectified Linear Unit e Revolução Retificadora	29
8.3	Dying ReLUs Problem	29
8.4	Corrigindo o Dying ReLUs Problem: As Variantes com Vazamento	29
8.4.1	Leaky ReLU	29
8.4.2	Parametric ReLU	29
8.4.3	Randomized Leaky ReLU	29
8.5	Em Busca da Suavidade	29
8.5.1	Exponential Linear Unit	29

8.5.2	Scaled Exponential Linear Unit	29
8.5.3	Noisy ReLU	29
8.6	O Problema dos Gradientes Explosivos	29
8.7	Comparativo de Desempenho das Funções Retificadoras	29
9	FUNÇÕES DE ATIVAÇÃO MODERNAS E OUTRAS FUNÇÕES DE ATIVAÇÃO	31
10	FUNÇÕES DE PERDA PARA CLASSIFICAÇÃO BINÁRIA	33
10.1	A Intuição da Perda: Medindo o Erro do Modelo	33
10.2	Entropia Cruzada Binária (Binary Cross-Entropy): A função de perda padrão	33
10.3	Perda Hinge (Hinge Loss)	33
10.4	Comparativo Visual e Prático	33
11	FUNÇÕES DE PERDA PARA CLASSIFICAÇÃO MULTILABEL	35
11.1	Softmax e a Distribuição de Probabilidades	35
11.2	Entropia Cruzada Categórica (Categorical Cross-Entropy)	35
11.3	Entropia Cruzada Categórica Esparsa (Sparse Categorical Cross-Entropy)	35
12	METAHEURÍSTICAS: OTIMIZANDO REDES NEURAIS SEM O GRADIENTE	37
12.1	Algoritmos Evolutivos	37
12.2	Inteligência de Enxame	37
IV	APRENDIZADO DE MÁQUINA CLÁSSICO	39
13	REGRESSÃO LINEAR E LOGÍSTICA	41
13.1	Exemplo Ilustrativo	41
14	ÁRVORES DE DECISÃO E FLORESTAS ALEATÓRIAS	43
14.1	Exemplo Ilustrativo	43
15	MÁQUINAS DE VETORES DE SUPORTE	45
15.1	Exemplo Ilustrativo	45
16	ENSAMBLE	47
16.1	Exemplo Ilustrativo	47
17	DIMENSIONALIDADE	49
17.1	Exemplo Ilustrativo	49

17.2	A Maldição da Dimensionalidade	49
17.3	Seleção de Características (Feature Selection)	49
17.4	Extração de Características (Feature Extraction)	49
17.4.1	Análise de Componentes Principais (PCA)	49
17.4.2	t-SNE (t-Distributed Stochastic Neighbor Embedding) e UMAP	49
18	CLUSTERIZAÇÃO	51
18.1	Exemplo Ilustrativo	51
18.2	Aprendizado Não Supervisionado: Encontrando Grupos nos Dados	51
18.3	Clusterização Particional: K-Means	51
18.4	Clusterização Hierárquica	51
18.5	Clusterização Baseada em Densidade: DBSCAN	51
V	REDES NEURAIIS PROFUNDAS (DNNS)	53
19	PERCEPTRONS MLP - REDES NEURAIIS ARTIFICIAIS	55
20	REDES FEEDFORWARD (FFNS)	57
21	REDES DE CRENÇA PROFUNDA (DBNS) E MÁQUINAS DE BOLTZMANN RESTRITAS	59
22	REDES NEURAIIS CONVOLUCIONAIS (CNN)	62
22.1	Exemplo Ilustrativo	62
22.2	Camadas Convolucionais: O Bloco Fundamental para as CNNs . .	62
22.2.1	Implementação em Python	62
22.3	Camadas de Pooling: Reduzindo a Dimensionalidade	62
22.3.1	Max Pooling	62
22.3.2	Avg Pooling	62
22.3.3	Global Avg Pooling	62
22.3.4	Implementação em Python	62
22.4	Camada Flatten: Achatando os Dados	62
22.4.1	Implementação em Python	62
22.5	Criando uma CNN	62
22.6	Detecção de Objetos	62
22.7	Redes Totalmente Convolucionais (FCNs)	62
22.8	You Only Look Once (YOLO)	62
22.9	Algumas Arquiteturas de CNNs	62
22.9.1	LeNet-5	62
22.9.2	AlexNet	62

22.9.3	GoogLeNet	62
22.9.4	VGGNet	62
22.9.5	ResNet	62
22.9.6	Xception	62
22.9.7	SENet	62
23	REDES RESIDUAIS (RESNETS)	63
24	REDES NEURAIS RECORRENTES (RNN)	65
24.1	Exemplo Ilustrativo	65
24.2	Neurônios e Células Recorrentes	65
24.2.1	Implementação em Python	65
24.3	Células de Memória	65
24.3.1	Implementação em Python	65
24.4	Criando uma RNN	65
24.5	O Problema da Memória de Curto Prazo	65
24.5.1	Células LSTM	65
24.5.2	Conexões Peephole	65
24.5.3	Células GRU	65
25	TÉCNICAS PARA MELHORAR O DESEMPENHO DE REDES NEURAIS	67
25.1	Técnicas de Inicialização	67
25.2	Regularização L1 e L2	67
25.3	Normalização	67
25.3.1	Normalização de Camadas	67
25.3.2	Normalização de Batch	67
25.4	Clipping do Gradiente	67
25.5	Dropout: Menos Neurônios Mais Aprendizado	67
25.6	Data Augmentation	67
26	TRANSFORMERS	69
26.1	As Limitações das RNNs: O Gargalo Sequencial	69
26.2	A Ideia Central: Self-Attention (Query, Key, Value)	69
26.3	Escalando a Atenção: Multi-Head Attention	69
26.4	A Arquitetura Completa: O Bloco Transformer	69
26.5	Entendendo a Posição: Codificação Posicional	69
26.6	As Três Grandes Arquiteturas	69
26.6.1	Encoder-Only (Ex: BERT): Para tarefas de entendimento	69
26.6.2	Decoder-Only (Ex: GPT): Para tarefas de geração	69

26.6.3	Encoder-Decoder (Ex: T5): Para tarefas de tradução/sumarização	69
26.7	Além do Texto: Vision Transformers (ViT)	69
27	REDES ADVERSÁRIAS GENERATIVAS (GANS)	71
28	MIXTURE OF EXPERTS (MOE)	73
29	MODELOS DE DIFUSÃO	75
30	REDES NEURAIIS DE GRAFOS (GNNS)	77
VI	APÊNDICES	79

Parte I

História da IA e do Computador

1 Uma Breve História do Computador

O texto do seu capítulo começa aqui...

2 Uma Breve História da Inteligência Artificial

O texto do seu capítulo começa aqui...

Parte II

Conceitos Matemáticos

3 Cálculo para Aprendizado de Máquina

3.1 Funções: A Base do Cálculo

3.2 Derivadas Ordinárias

3.3 Integrais Simples

3.4 Derivadas Parciais

4 Álgebra Linear para Aprendizado de Máquina

4.1 A Unidade Fundamental: Vetores e Espaços Vetoriais

4.2 Organizando Dados: Matrizes e Suas Operações

4.3 Tensores: A Estrutura de Dados do Deep Learning

4.4 Resolvendo Sistemas e Encontrando Propriedades: Autovalores e Autovetores

4.5 Decomposição de Matrizes (SVD e PCA)

5 Probabilidade e Estatística para Aprendizado de Máquina

5.1 Medindo a Incerteza: Probabilidade Básica e Condicional

5.2 O Teorema de Bayes: Aprendendo com Evidências

5.3 Descrevendo os Dados: Estatística Descritiva: Média, mediana, variância, desvio padrão

5.4 Variáveis Aleatórias e Distribuições de Probabilidade

5.5 A Função de Máxima Verossimilhança (Maximum Likelihood Estimation - MLE)

Parte III

Pilares das Redes Neurais

6 O Algoritmo da Retropropagação e Os Otimizadores Baseados em Gradiente

6.1 O Método do Gradiente Descendente

6.1.1 Exemplo Ilustrativo

6.1.2 O Método em Si

6.1.3 Implementação em Python

6.2 A Retropropagação: Aprendendo com os Erros

6.3 Otimizadores Baseados em Gradiente

6.3.1 Método do Gradiente Estocástico

6.3.2 Método do Gradiente com Momentum

6.3.3 Nesterov

6.3.4 AdaGrad

6.3.5 RMSProp

6.3.6 Adam

6.3.7 Nadam

6.4 O Método de Newton: Indo Além do Gradiente

7 Funções de Ativação Sigmoidais

7.1 Teoremas da Aproximação Universal

7.2 Exemplos Ilustrativo

7.3 A Sigmoid Logística

7.4 Tangente Hiperbólica

7.5 Softsign: Uma Sigmoidal Mais Barata

7.6 Hard Sigmoid e Hard Tanh: O Sacrifício da Suavidade em Prol do Desempenho

7.7 O Desaparecimento de Gradientes

7.8 Comparativo de Desempenho das Sigmoidais

8 Funções de Ativação Retificadoras

8.1 Exemplo Ilustrativo

8.2 Rectified Linear Unit e Revolução Retificadora

8.3 Dying ReLUs Problem

8.4 Corrigindo o Dying ReLUs Problem: As Variantes com Vazamento

8.4.1 Leaky ReLU

8.4.2 Parametric ReLU

8.4.3 Randomized Leaky ReLU

8.5 Em Busca da Suavidade

8.5.1 Exponential Linear Unit

8.5.2 Scaled Exponential Linear Unit

8.5.3 Noisy ReLU

8.6 O Problema dos Gradientes Explosivos

8.7 Comparativo de Desempenho das Funções Retificadoras

9 Funções de Ativação Modernas e Outras Funções de Ativação

O texto do seu capítulo começa aqui...

10 Funções de Perda para Classificação Binária

10.1 A Intuição da Perda: Medindo o Erro do Modelo

10.2 Entropia Cruzada Binária (Binary Cross-Entropy): A função de perda padrão

10.3 Perda Hinge (Hinge Loss)

10.4 Comparativo Visual e Prático

11 Funções de Perda para Classificação Multilabel

11.1 Softmax e a Distribuição de Probabilidades

11.2 Entropia Cruzada Categórica (Categorical Cross-Entropy)

11.3 Entropia Cruzada Categórica Esparsa (Sparse Categorical Cross-Entropy)

12 Metaheurísticas: Otimizando Redes Neurais Sem o Gradiente

O texto do seu capítulo começa aqui...

12.1 Algoritmos Evolutivos

12.2 Inteligência de Enxame

Parte IV

Aprendizado de Máquina Clássico

13 Regressão Linear e Logística

13.1 Exemplo Ilustrativo

14 Árvores de Decisão e Florestas Aleatórias

14.1 Exemplo Ilustrativo

15 Máquinas de Vetores de Suporte

15.1 Exemplo Ilustrativo

16 Ensamble

16.1 Exemplo Ilustrativo

17 Dimensionalidade

17.1 Exemplo Ilustrativo

17.2 A Maldição da Dimensionalidade

17.3 Seleção de Características (Feature Selection)

17.4 Extração de Características (Feature Extraction)

17.4.1 Análise de Componentes Principais (PCA)

17.4.2 t-SNE (t-Distributed Stochastic Neighbor Embedding) e UMAP

18 Clusterização

18.1 Exemplo Ilustrativo

18.2 Aprendizado Não Supervisionado: Encontrando Grupos nos Dados

18.3 Clusterização Particional: K-Means

18.4 Clusterização Hierárquica

18.5 Clusterização Baseada em Densidade: DBSCAN

Parte V

Redes Neurais Profundas (DNNs)

19 Perceptrons MLP - Redes Neurais Artificiais

O texto do seu capítulo começa aqui...

20 Redes FeedForward (FFNs)

O texto do seu capítulo começa aqui...

21 Redes de Crença Profunda (DBNs) e Máquinas de Boltzmann Restritas

O texto do seu capítulo começa aqui...

22 Redes Neurais Convolucionais (CNN)

22.1 Exemplo Ilustrativo

22.2 Camadas Convolucionais: O Bloco Fundamental para as CNNs

22.2.1 Implementação em Python

22.3 Camadas de Pooling: Reduzindo a Dimensionalidade

22.3.1 Max Pooling

22.3.2 Avg Pooling

22.3.3 Global Avg Pooling

22.3.4 Implementação em Python

22.4 Camada Flatten: Achatando os Dados

22.4.1 Implementação em Python

22.5 Criando uma CNN

22.6 Detecção de Objetos

22.7 Redes Totalmente Convolucionais (FCNs)

22.8 You Only Look Once (YOLO)

22.9 Algumas Arquiteturas de CNNs

22.9.1 LeNet-5

22.9.2 AlexNet

22.9.3 GoogLeNet

22.9.4 VGGNet

22.9.5 ResNet

22.9.6 Xception

23 Redes Residuais (ResNets)

O texto do seu capítulo começa aqui...

24 Redes Neurais Recorrentes (RNN)

O texto do seu capítulo começa aqui...

24.1 Exemplo Ilustrativo

24.2 Neurônios e Células Recorrentes

24.2.1 Implementação em Python

24.3 Células de Memória

24.3.1 Implementação em Python

24.4 Criando uma RNN

24.5 O Problema da Memória de Curto Prazo

24.5.1 Células LSTM

24.5.2 Conexões Peephole

24.5.3 Células GRU

25 Técnicas para Melhorar o Desempenho de Redes Neurais

25.1 Técnicas de Inicialização

25.2 Regularização L1 e L2

25.3 Normalização

25.3.1 Normalização de Camadas

25.3.2 Normalização de Batch

25.4 Clipping do Gradiente

25.5 Dropout: Menos Neurônios Mais Aprendizado

25.6 Data Augmentation

26 Transformers

26.1 As Limitações das RNNs: O Gargalo Sequencial

26.2 A Ideia Central: Self-Attention (Query, Key, Value)

26.3 Escalando a Atenção: Multi-Head Attention

26.4 A Arquitetura Completa: O Bloco Transformer

26.5 Entendendo a Posição: Codificação Posicional

26.6 As Três Grandes Arquiteturas

26.6.1 Encoder-Only (Ex: BERT): Para tarefas de entendimento

26.6.2 Decoder-Only (Ex: GPT): Para tarefas de geração

26.6.3 Encoder-Decoder (Ex: T5): Para tarefas de tradução/sumarização

26.7 Além do Texto: Vision Transformers (ViT)

27 Redes Adversárias Generativas (GANs)

O texto do seu capítulo começa aqui...

28 Mixture of Experts (MoE)

O texto do seu capítulo começa aqui...

29 Modelos de Difusão

O texto do seu capítulo começa aqui...

30 Redes Neurais de Grafos (GNNs)

O texto do seu capítulo começa aqui...

Parte VI

Apêndices

