

MBAUSP
ESALQ

DEEP LEARNING I

Prof. Dr. Jeronimo Marcondes

Amilcar Ferraz Farina 188.708.208-58

MBAUSP ESALA

A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização.

Lei nº 9610/98

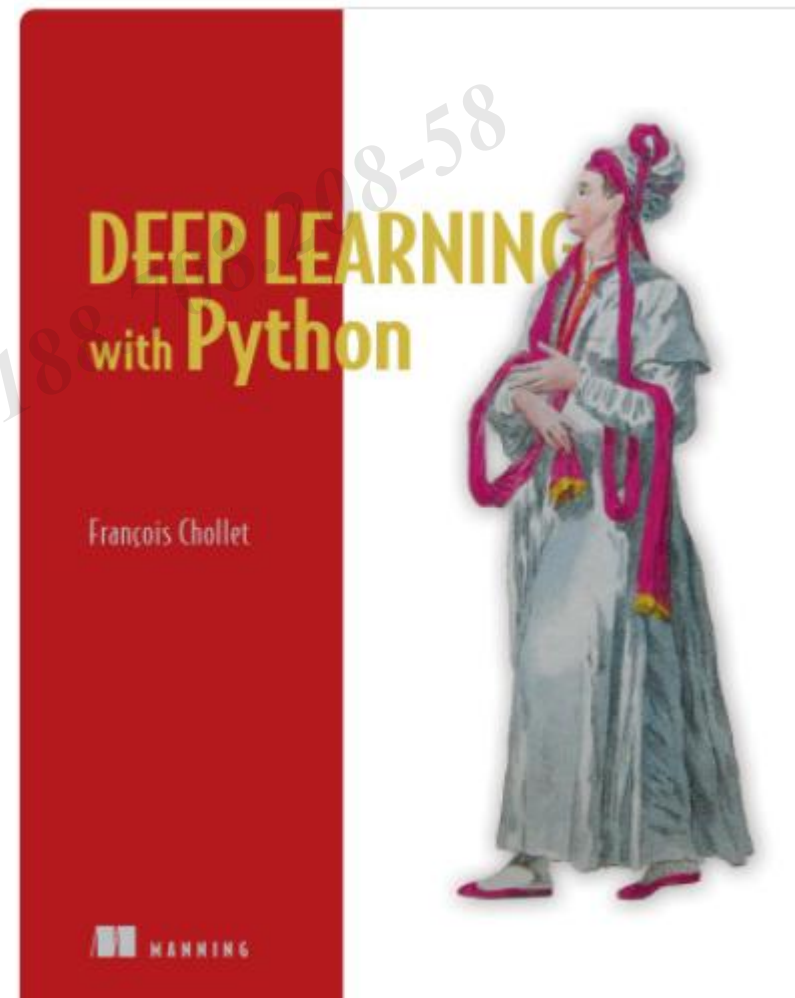
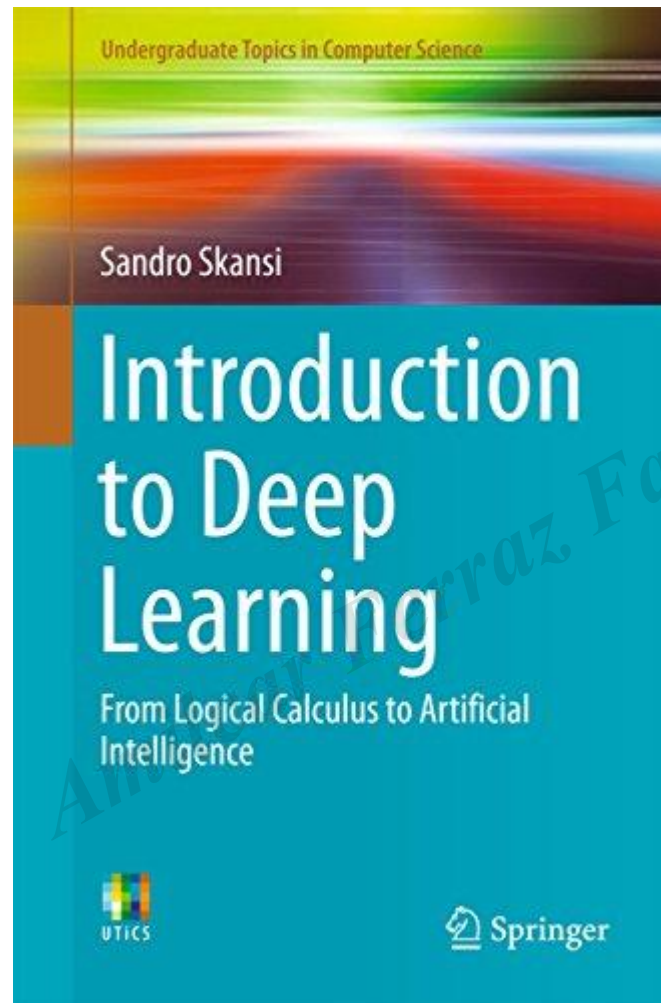
Introdução

Plano de ataque:

1. O papel das redes neurais multicamadas;
2. O problema do overfitting;
3. O problema de definição de hiperparâmetros;
4. O problema de encontrar solução ótima.

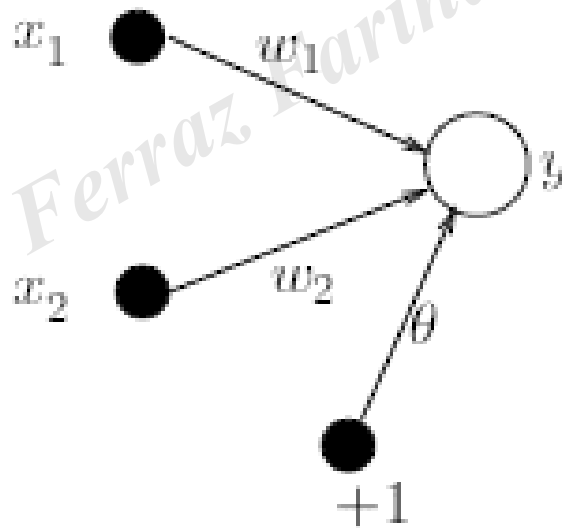
Amilcar Ferraz Farina 188.708.208-58

Introdução

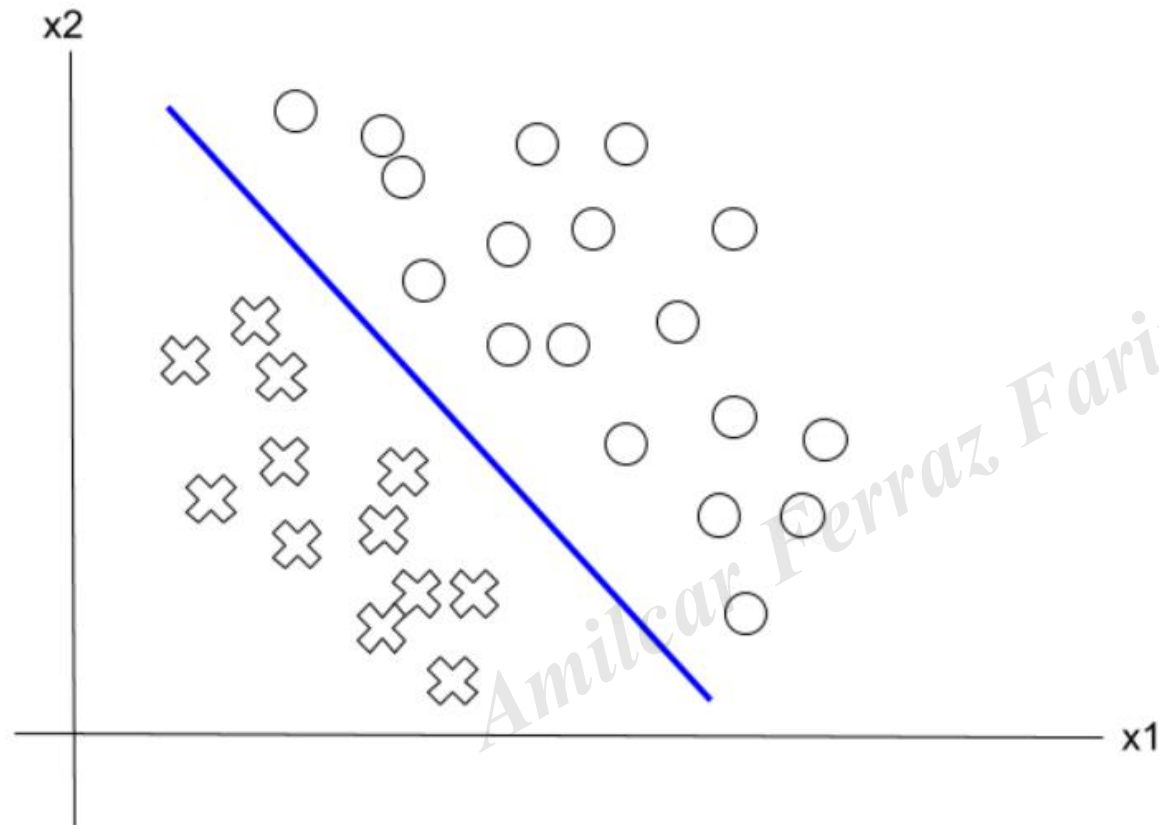


Introdução

- Problema inicial como a função XOR nas redes neurais.



Introdução



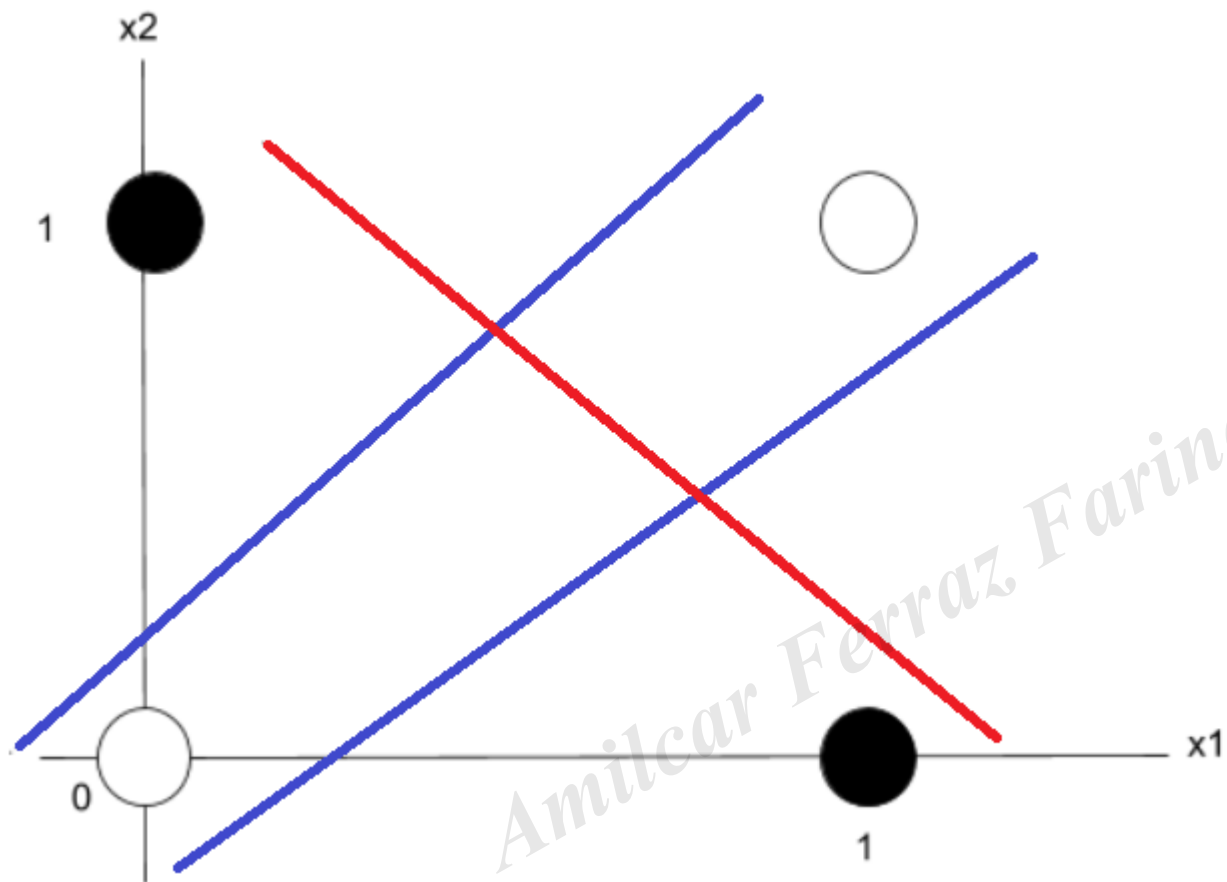
<https://automaticaddison.com/linear-separability-and-the-xor-problem/>

XoR

- Exclusive Or

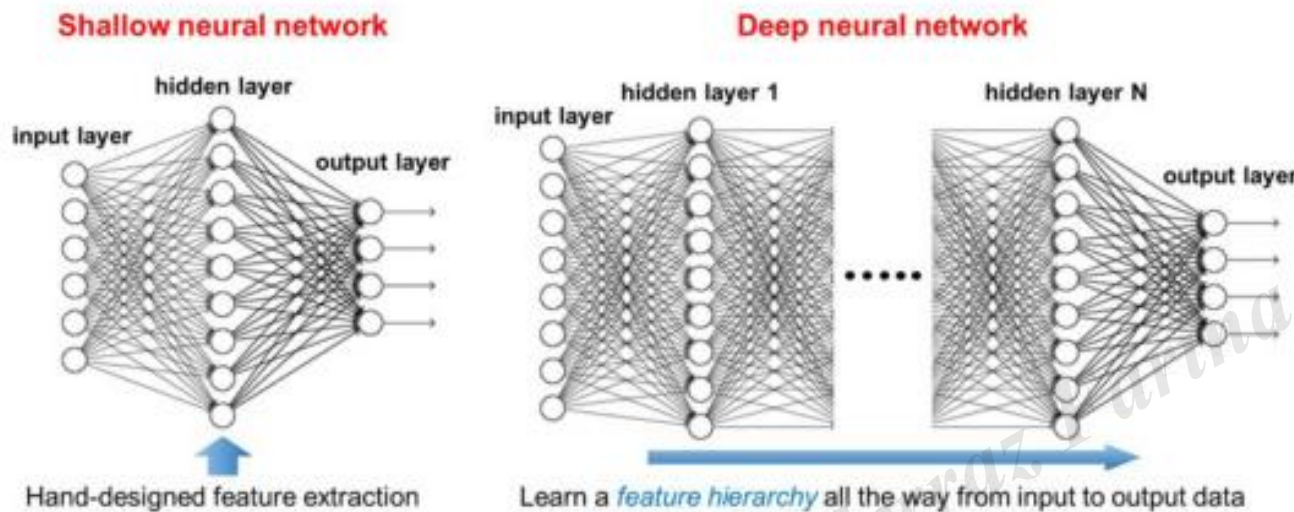
x_0	x_1	d
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

Amílcar Ferraz Farina 188.708.208-58



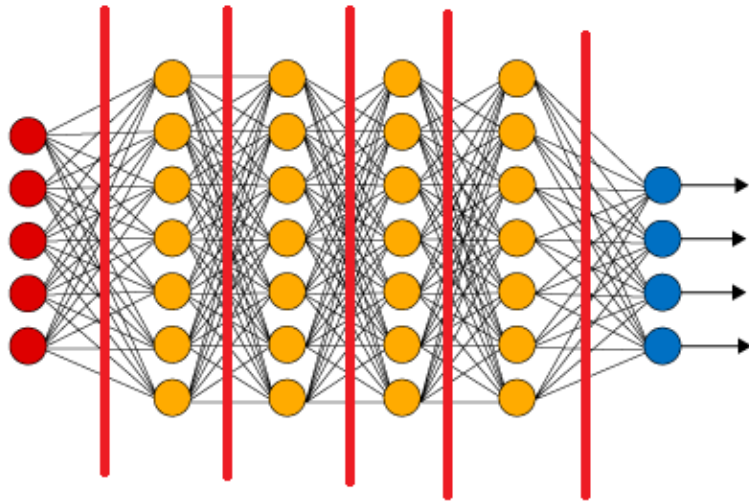
<https://automaticaddison.com/linear-separability-and-the-xor-problem/>

Rede Neural Multicamada



- Muitas camadas permitem identificar relações não lineares.
- No caso de termos varias camadas intermediarias obtém-se o que é chamado aprendizado profundo (deep learning).

Estrutura da Rede



- Camadas – grupo de neurônios em um estágio do processo.
- Camada Densa – Conecta cada neurônio em uma camada com todos os neurônios da camada anterior. Por exemplo, se a camada atual tem 5 neurônios e a camada anterior tem 3, o total de conexões é de 15.

Função Perda

- Função que permite verificar o quão assertivo é determinada previsão;
- Predicted x Actual;
- São diferentes para variáveis contínuas e categóricas.

Amilcar Ferraz Farina 188.708.208-58

Função Perda

- As variáveis categóricas contêm um número finito de categorias ou grupos distintos. Os dados categóricos podem não ter uma ordem lógica. Por exemplo, os preditores categóricos incluem gênero, tipo de material e método de pagamento.
- Variáveis contínuas são variáveis numéricas que têm um número infinito de valores entre dois valores quaisquer. Uma variável contínua pode ser numérica ou de data/hora. Por exemplo, o comprimento de uma peça ou a data e hora em que um pagamento é recebido.
- Variáveis discretas são variáveis numéricas que têm um número contável de valores entre quaisquer dois valores. Uma variável discreta é sempre numérica. Por exemplo, o número de reclamações de clientes ou o número de falhas ou defeitos.

<https://support.minitab.com/>

Algumas funções de erro

- Erro quadrático Médio:

$$EQM = \frac{1}{n} \sum_{n=1}^k (Actual - Predicted)^2$$

	Nota	
	Actual	Predicted
A	7	8
B	10	9
C	5	10
D	8	8

Algumas funções de erro

- Erro Absoluto Médio

$$EAM = \frac{1}{n} \sum_{n=1}^k |Actual - Predicted|$$

	Nota	
	Actual	Predicted
A	7	8
B	10	9
C	5	10
D	8	8

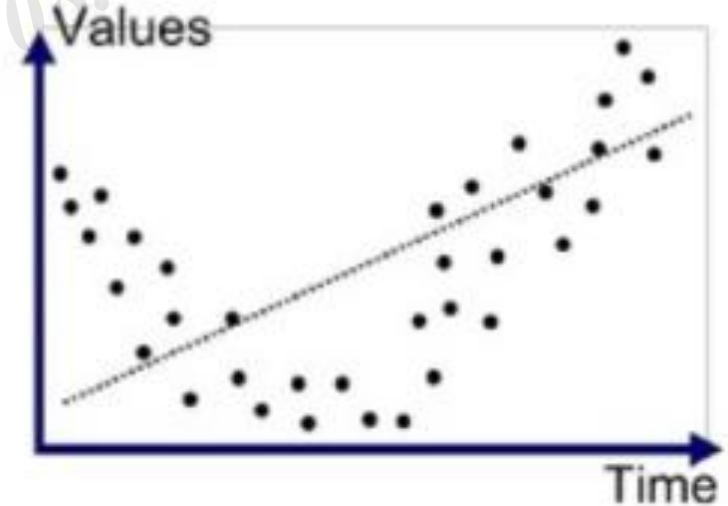
Algumas funções de erro

- Funções para output contínuo;
- Há necessidade de outras métricas para classificação, como 1 e 0;
- Binary cross-entropy;
- Categorical cross-entropy.

Amilcar Ferraz Farina 188.708.208-58

Viés x Variância

- O que estamos buscando?
- O que pode ocorrer?
- Importância da Generalização – teoria do mapa.



Underfitted

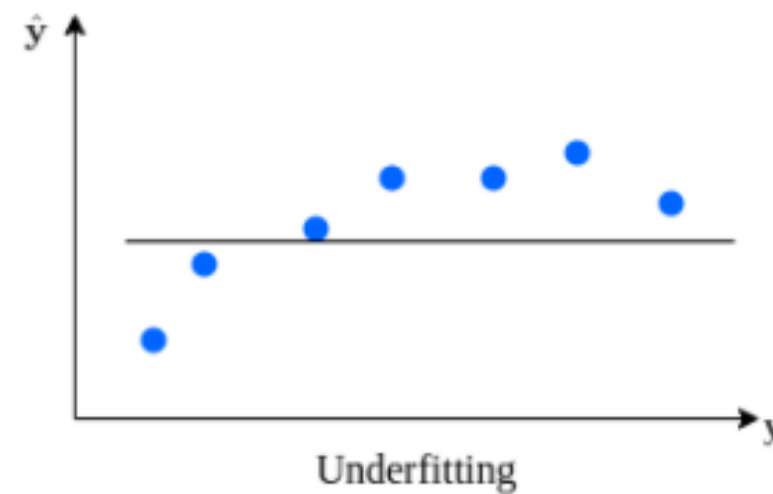
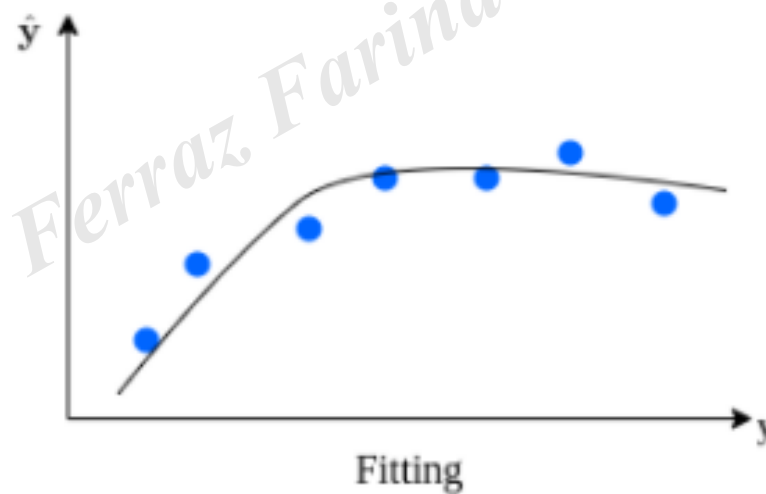
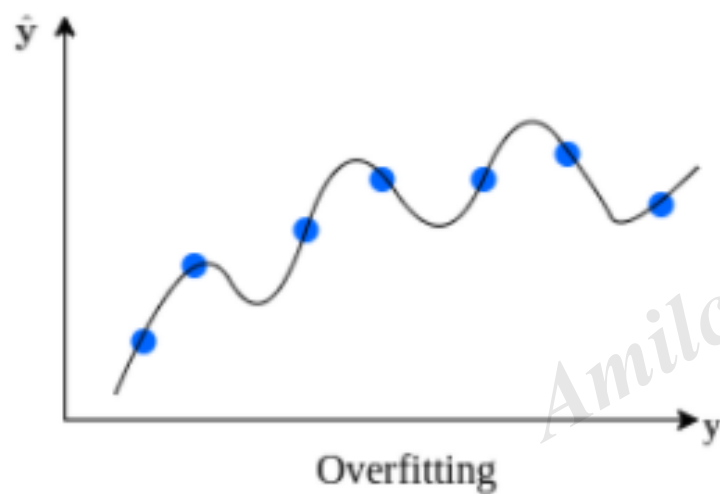
Viés x Variância

- Ausência de Viés - na média você acerta.
- Redução de Variância – tiro ao alvo.

Amilcar Ferraz Farina 188.708.208-58

Viés x Variância

- Preocupação com a variância.



<https://www.baeldung.com/cs/epoch-neural-networks>

Fitting

- UnderFitting - Seu modelo está subjugando os dados de treinamento quando o modelo tem um desempenho insatisfatório nos dados de treinamento. Isso ocorre porque o modelo é incapaz de capturar a relação entre os exemplos de entrada (geralmente chamados de X) e os valores de destino (geralmente chamados de Y).

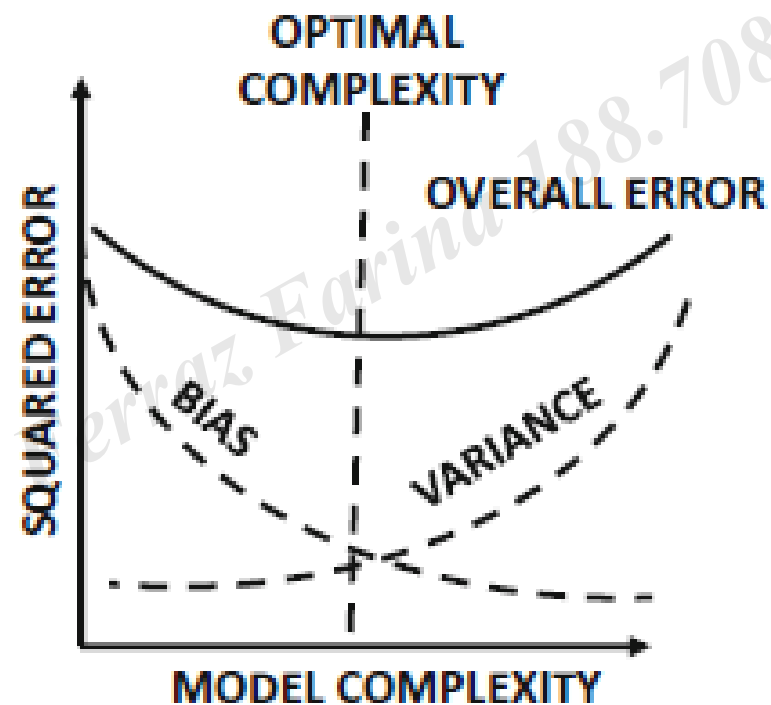
Amilcar Ferraz Farina 180-108-208-58

Fitting

- OverFitting - Seu modelo está super ajustando seus dados de treinamento quando você vê que o modelo tem um bom desempenho nos dados de treinamento, mas não nos dados de avaliação. Isso ocorre porque o modelo está memorizando os dados que viu e é incapaz de generalizar para exemplos não vistos. Muito comum em Redes Neurais.

Amilcar Ferraz Farina 188-108-208-58

Fitting



Redução do overfitting

- Como reduzir overfitting?
- Reduzir número de camadas;
- Mais parâmetros = + capacidade de memorização;
- Regularização.

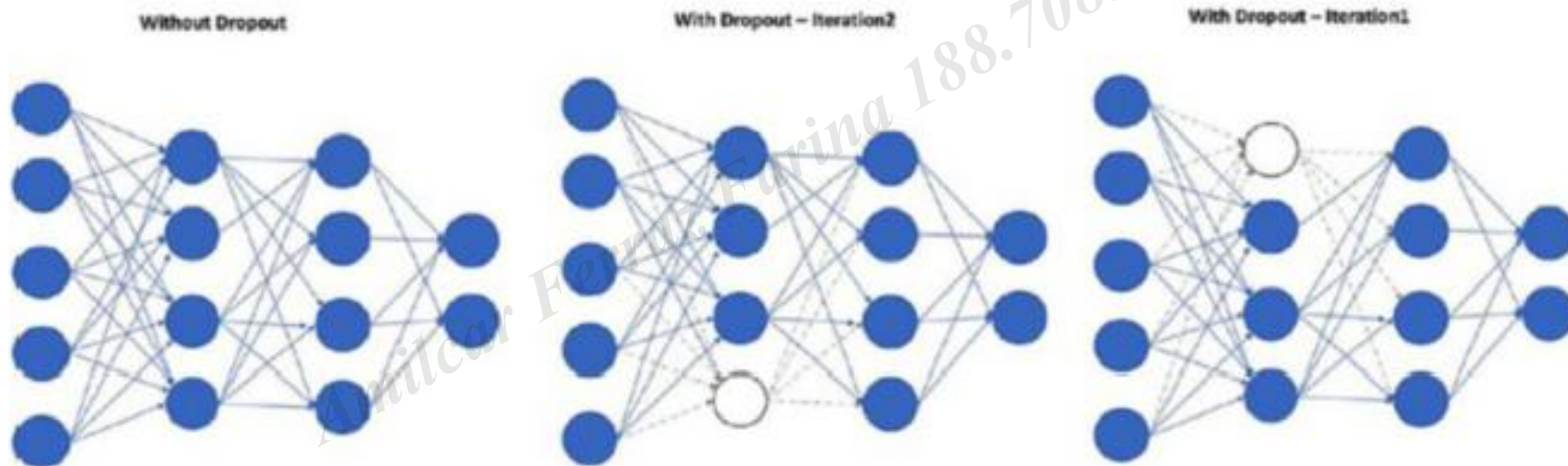
Amilcar Ferreira Farina 188.708.208-58

Regularização

- Introdução de “barulho”.
- Dropout

O modelo descarta ou desativa arbitrariamente alguns neurônios para uma camada durante cada iteração. Em cada iteração, o modelo olha para uma estrutura ligeiramente diferente de si mesma para otimizar (como um casal de neurônios e as conexões seriam desativadas). Diga que temos duas camadas sucessivas, H1 e H2, com 15 e 20 neurônios, respectivamente. A aplicação da técnica de eliminação entre essas duas camadas resultaria em descartar aleatoriamente alguns neurônios (com base em uma porcentagem definida) para H1, o que, portanto, reduz as conexões entre H1 e H2. Esse processo se repete para cada iteração com aleatoriedade, portanto, se o modelo tiver que aprender para um lote e atualizou os pesos, o próximo lote pode ter um conjunto bastante diferente de pesos e conexões para treinar.

Dropout



L1

- Ajuste dos pesos feito por meio da regularização.
- O foco não é mais no neurônio, mas no valor do peso.
- O peso é ajustado até o valor de zero dependendo da sua importância.

Amilcar Ferreira Farina 188.708.208-58

L2

- Ajuste dos pesos feito por meio da regularização;
- O foco não é mais no neurônio, mas no valor do peso;
- O peso é ajustado até valor próximo de zero dependendo da sua importância.

Amilcar Ferreira Farina 188.708.208-58

Early Stopping

- Final de cada época – verificar melhoria.
- Melhoria não está mais ocorrendo? Pare!
- Isso evita o overfitting.

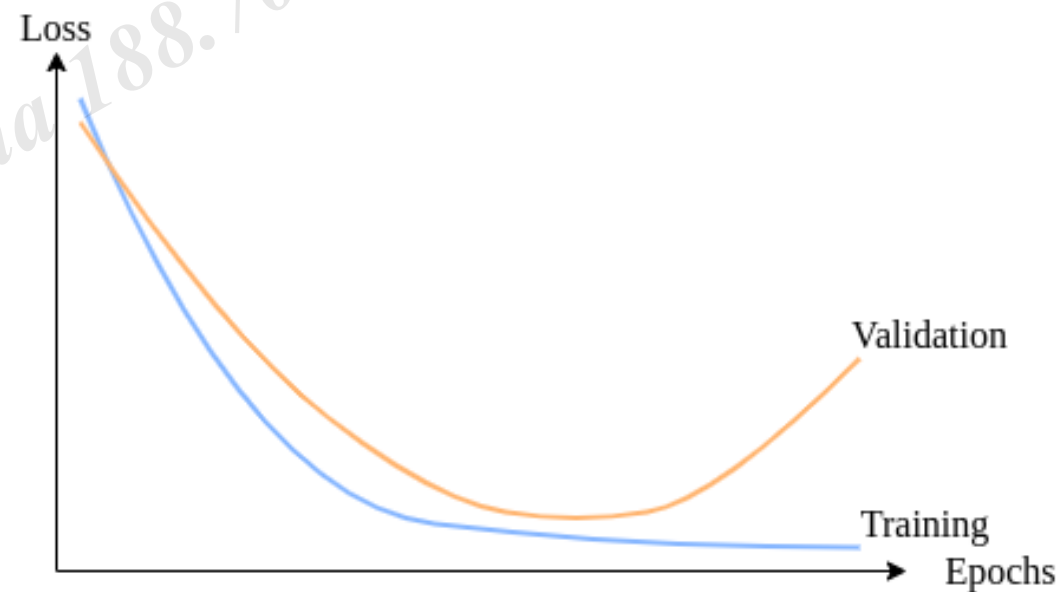
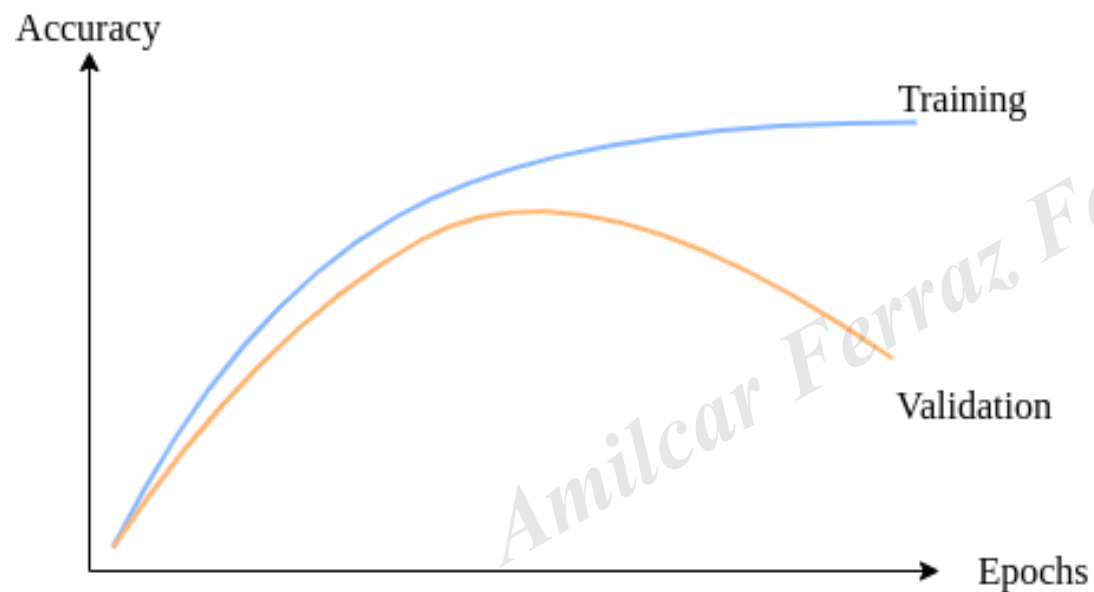
Amilcar Ferraz Farina 188.708.208-58

Épocas

- Como verificar se temos overfitting?
- Quantas vezes os dados passam pela rede.
- Cada época é um passo em direção ao resultado “ótimo”.
- Convergência.

Amilcar Ferraz Farina 188.708.208-58

Épocas



<https://www.baeldung.com/cs/epoch-neural-networks>

Batch

- Como realizar o treinamento?
- Técnicas que podem ajudar no processo de otimização.
- Uso de batch (lotes).

Amilcar Ferraz Farina 188.708.208-58

Batch

Normalmente, o treinamento seria feito em lotes devido a restrições de memória no sistema. Um lote é uma coleção de amostras de treinamento de toda a entrada. A rede atualiza seus pesos depois de processar todas as amostras em um lote. Isso é chamado de iteração (ou seja, uma passagem bem-sucedida de todas as amostras em um lote, seguido por uma atualização de peso na rede). A computação de todas as amostras de treinamento fornecidas nos dados de entrada com atualizações de peso lote por lote é chamado de época. Em cada iteração, a rede aproveita a função do otimizador para fazer uma pequena mudança para seus parâmetros de peso (que foram inicializados aleatoriamente no início) para melhorar a previsão final, reduzindo a função de perda. Passo a passo, com várias iterações e, em seguida, várias épocas, a rede atualiza seu pondera e aprende a fazer uma previsão correta para o treinamento dado amostras.

Hiperparâmetros

- O que são?
- Qualquer número utilizado pela rede que não é aprendido.
- Como determinar esses valores?
- Cross validation? Grid Search?

Amilcar Ferraz Farina 188.708.208-58

Hiperparâmetros

- Vamos lembrar de alguns?

1. Learning rate
2. Batch size
3. Epochs
4. Função de ativação

Amilcar Ferraz Farina 188.708.208-58

AUTOML

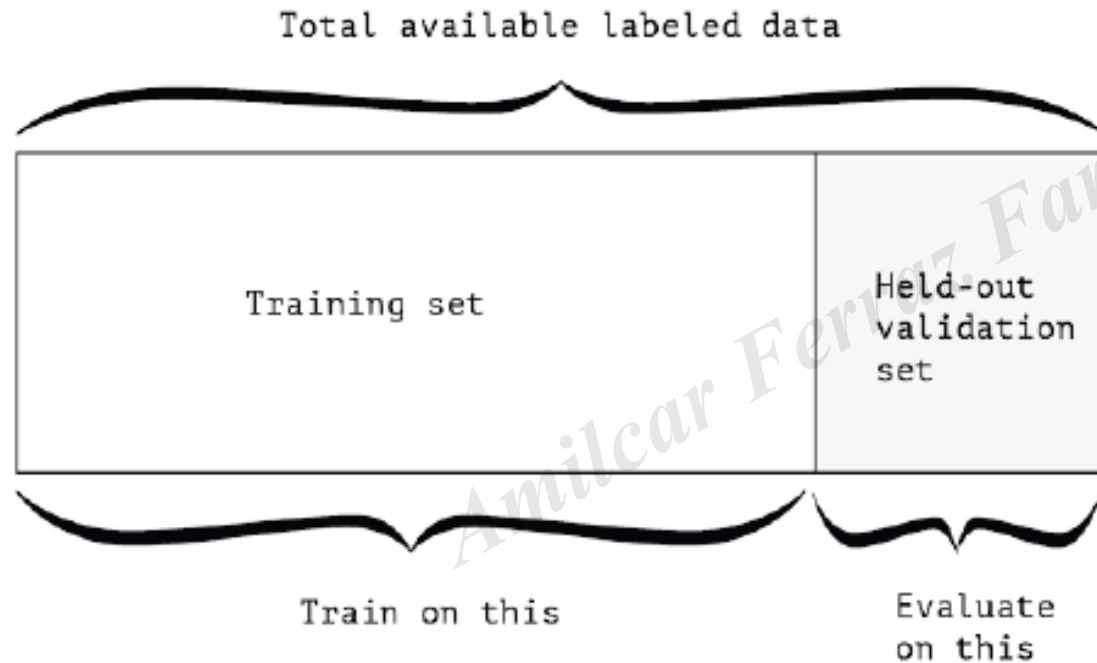
Vamos discutir um pouco sobre isso.

- AutoKeras, H2O.
- Transfer Learning.
- Grid Search

Amilcar Ferraz Farina 188.708.208-58

Cross validation

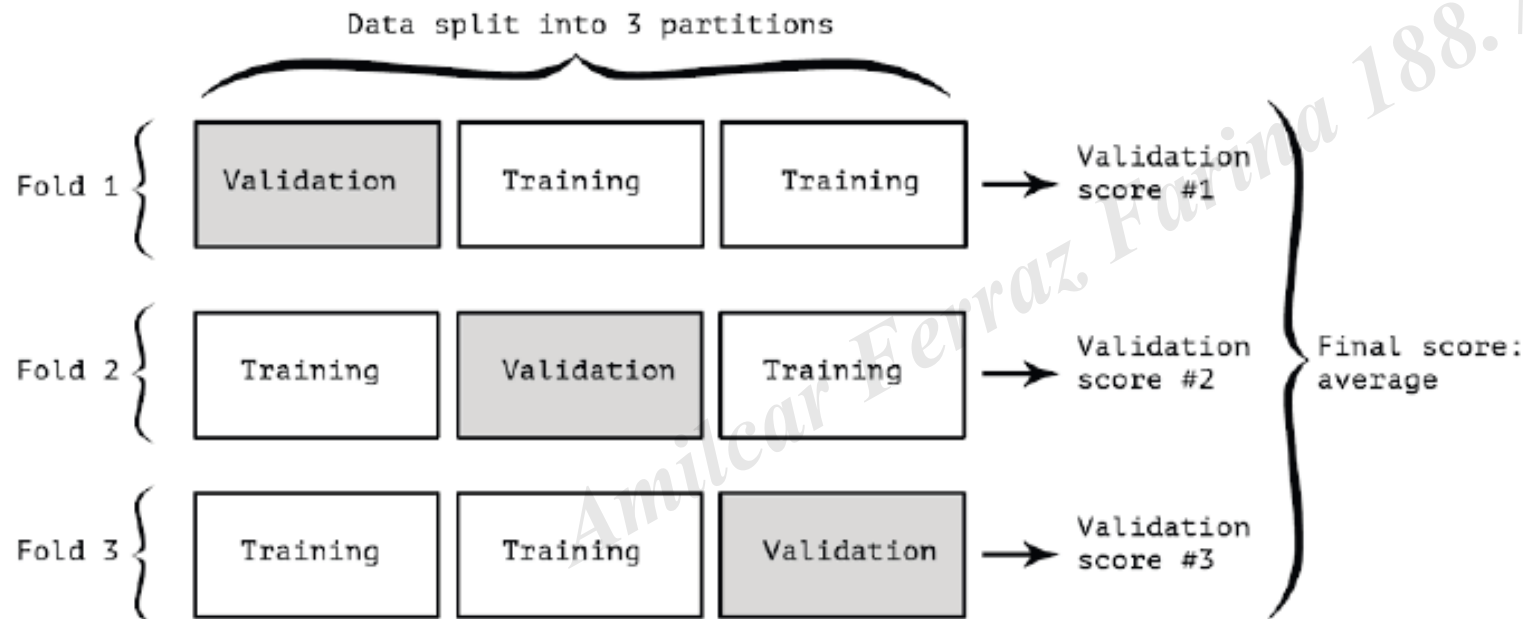
- Como funciona?
- Conjunto de treinamento, teste e validação



Deep Learning with R

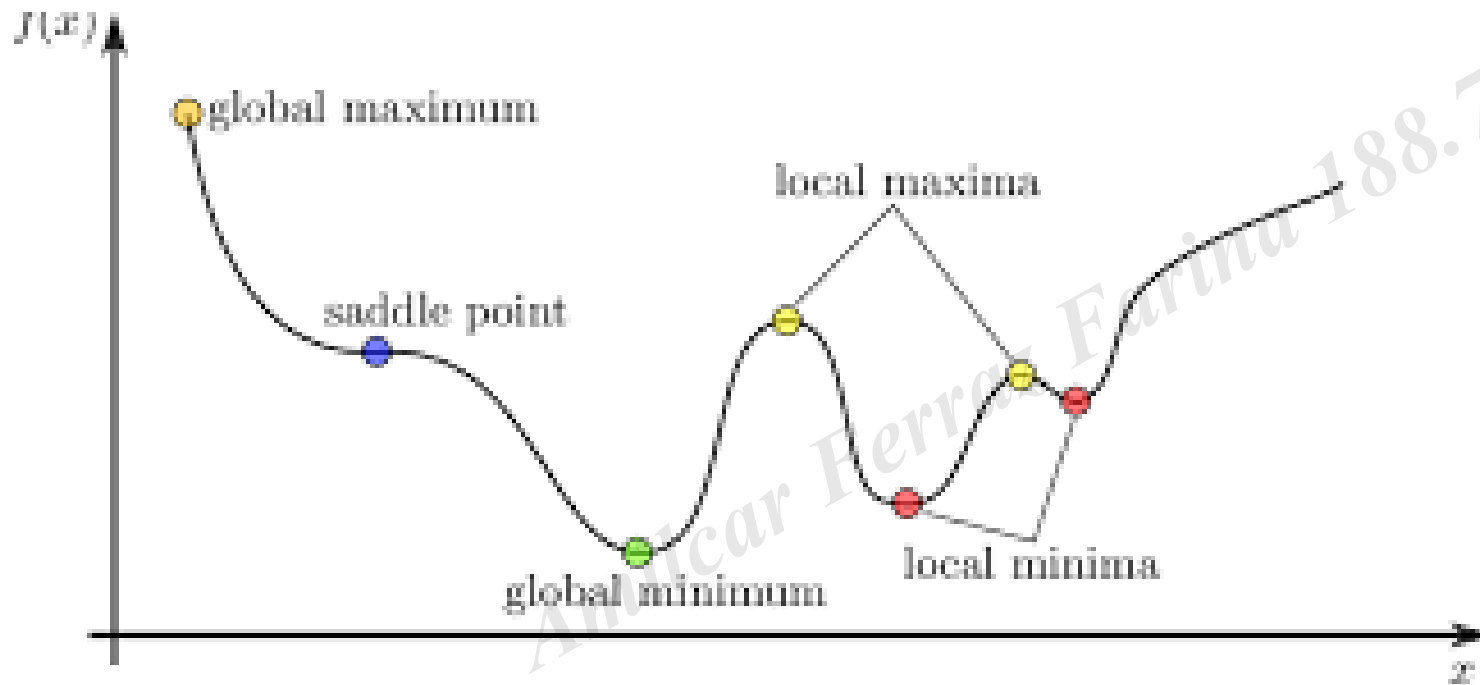
Cross validation

- K fold cross validation



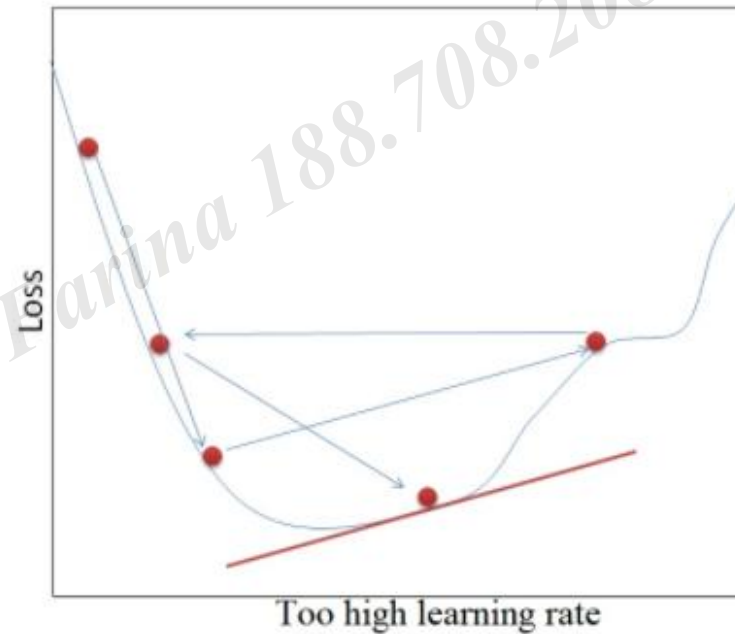
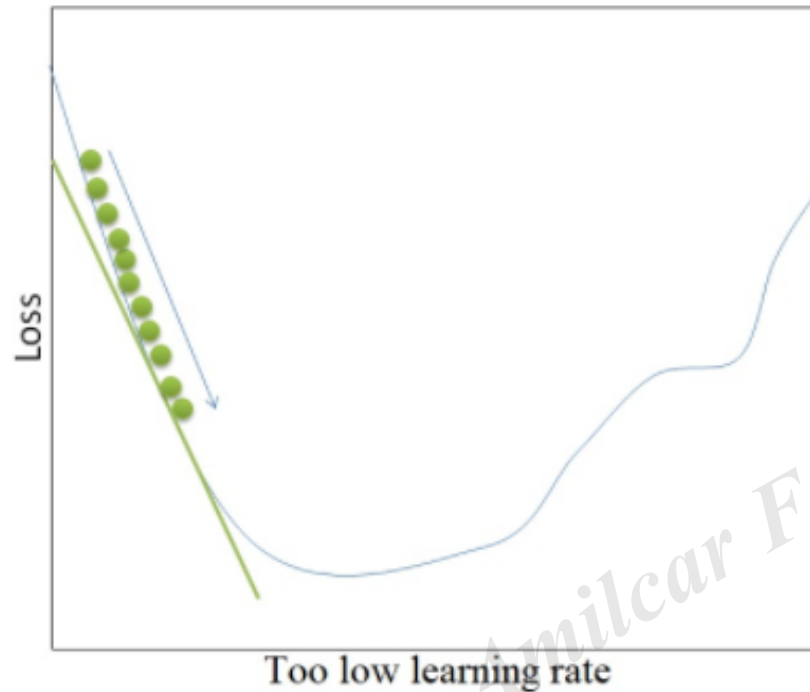
Deep Learning with R

Mínimo local e ponto de sela



<https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/>

Mínimo local e ponto de sela



<https://www.analyticsvidhya.com/blog/2021/05/tuning-the-hyperparameters-and-layers-of-neural-network-deep-learning/>

Otimizadores

- Ou seja, dada a função perda, como atualizar os pesos?
- Qual a regra?
- Nós usamos o modelo clássico de descida do gradiente, mas há outras.

Amilcar Ferraz Farina 188.708.208-58

Otimizadores

- ADAGRAD
- ADAM
- RMSPROP

Amilcar Ferraz Farina 188.708.208-58

Momentun

- Também chamada de inércia
- Conceito próximo à de inércia = atrito
- Figura acima – a bolinha com velocidade poderia passar o ponto de sela e chegar no mínimo.

Amilcar Ferraz Farina 188.708.208-58

Momentum

$$w_i^{new} = w_i^{old} - \eta \frac{\partial E}{\partial w_i^{old}} + \mu(|w_i^{old} - w_i^{older}|)$$

Introduction to deep Learning

MBAUSP ESALA

Obrigado!

<https://www.linkedin.com/in/jeronymo-marcondes-585a26186>