<https://medium.com/data-hackers/neural-network-deep-learning-parte-4-introdu%C3%A7%C3%A3o-te%C3%B3rica-db45ebd77029>

<https://medium.com/data-hackers/neural-network-deep-learning-parte-3-introdu%C3%A7%C3%A3o-te%C3%B3rica-24bd2ef10202>

<https://www.deeplearningbook.com.br/?s=cap%C3%ADtulo+1>

<https://www.deeplearningbook.com.br/?s=loss>

Função de ativação: <https://www.deeplearningbook.com.br/?s=Fun%C3%A7%C3%A3o+de+ativa%C3%A7%C3%A3o>

* <https://iaexpert.academy/2020/05/25/funcoes-de-ativacao-definicao-caracteristicas-e-quando-usar-cada-uma/>
* <https://www.deeplearningbook.com.br/funcao-de-ativacao/#:~:text=As%20fun%C3%A7%C3%B5es%20de%20ativa%C3%A7%C3%A3o%20s%C3%A3o,fornecida%20ou%20deve%20ser%20ignorada>.

Loss function:

* <https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/>
* <https://www.datarobot.com/blog/introduction-to-loss-functions/>
* <https://towardsdatascience.com/importance-of-loss-function-in-machine-learning-eddaaec69519>

Otimizadores:

* <https://www.ime.unicamp.br/~mac/db/2019-2S-157490.pdf>

Capítulo 15 – Algoritmo Backpropagation Parte 2 – Treinamento de Redes Neurais <https://www.deeplearningbook.com.br/algoritmo-backpropagation-parte-2-treinamento-de-redes-neurais/>

**1- Instância**: linha, registro ou observação em seu conjunto de dados, sobre a qual você deseja fazer uma previsão. Por exemplo, cada instância pode ser um registro contendo informações sobre pacientes como idade, peso e altura e informação se o paciente desenvolveu diabetes. Queremos prever a ocorrência de diabetes com base nas características do paciente.

**2- Rótulo (Label)**: uma resposta para uma tarefa de previsão, seja a resposta produzida por um sistema de aprendizado de máquina ou a resposta correta fornecida nos dados de treinamento. Por exemplo, o rótulo da instância definido no item anterior seria “diabetes”, indicando se o paciente desenvolveu ou não a doença.

**3- Classe**: uma categoria de um conjunto de valores de destino enumerados para um rótulo. Por exemplo, em um modelo de classificação binária que detecta spam, as duas classes são spam e não spam. Em um modelo de classificação multiclasse que identifica raças de cães, as classes seriam poodle, beagle, pug e assim por diante.

**4- Recurso (Feature)**: uma propriedade de uma instância usada em uma tarefa de previsão. Por exemplo, uma instância com dados de pacientes teria um recurso indicando a idade.

**5- Recurso Denso**: um recurso em que a maioria dos valores é diferente de zero, normalmente um tensor de valores de ponto flutuante.

**6- Recurso Esparso**: vetor de recursos cujos valores são predominantemente zero ou vazios. Por exemplo, um vetor contendo um único valor 1 e um milhão de valores 0 é esparso. Como outro exemplo, as palavras em uma consulta de pesquisa também podem ser um recurso esparso – há muitas palavras possíveis em um determinado idioma, mas apenas algumas delas ocorrem em uma determinada consulta.

**7- Dados Categóricos**: recursos com um conjunto discreto de valores possíveis representando informação qualitativa. Por exemplo, considere um recurso categórico chamado estilo da casa, que possui um conjunto discreto de três valores possíveis: apartamento, casa e chalé.

**8- Dados Numéricos**: características representadas como números inteiros ou números reais. Por exemplo, em um modelo imobiliário, você provavelmente representaria o tamanho de uma casa (em metros quadrados) como dados numéricos. Representar um recurso como dados numéricos indica que os valores do recurso têm uma relação matemática entre si e possivelmente com o rótulo.

**9- Exemplo**: uma instância (com seus recursos) e um rótulo.

**10- Modelo**: uma representação estatística de uma tarefa de previsão. Você treina um modelo em exemplos e depois usa o modelo para fazer previsões.

**11- Baseline**: um modelo usado como ponto de referência para comparar o desempenho de outro modelo (normalmente mais complexo). Por exemplo, um modelo de regressão logística pode servir como uma boa linha de base para um modelo de Deep Learning. Para um problema específico, a linha de base ajuda os [Cientistas de Dados](https://www.datascienceacademy.com.br/bundle/formacao-cientista-de-dados) a quantificar o desempenho mínimo esperado que um novo modelo deve alcançar para que o novo modelo seja útil.

**12- Aprendizado de Máquina Supervisionado**: técnica para treinar um modelo a partir de dados de entrada e seus rótulos correspondentes. O aprendizado de máquina supervisionado é análogo a um aluno aprendendo um assunto estudando um conjunto de perguntas e suas respostas correspondentes. Depois de dominar o mapeamento entre perguntas e respostas, o aluno pode fornecer respostas para novas perguntas (nunca antes vistas) sobre o mesmo tema.

**13- Aprendizado de Máquina Não Supervisionado**: técnica para treinar um modelo para encontrar padrões em um conjunto de dados, normalmente um conjunto de dados não rotulado. O uso mais comum do aprendizado de máquina não supervisionado é agrupar dados em grupos de exemplos semelhantes. Por exemplo, um algoritmo de aprendizado de máquina não supervisionado pode agrupar músicas com base em várias propriedades da música. Os clusters resultantes podem se tornar uma entrada para outros algoritmos de aprendizado de máquina (por exemplo, para um serviço de recomendação de música). O clustering pode ser útil em domínios onde os rótulos verdadeiros são difíceis de obter. Por exemplo, em domínios como antiabuso e fraude, os clusters podem ajudar os humanos a entender melhor os dados. Outro exemplo de aprendizado de máquina não supervisionado é a análise de componentes principais (PCA). Por exemplo, a aplicação de PCA em um conjunto de dados contendo o conteúdo de milhões de carrinhos de compras pode revelar que carrinhos de compras contendo limões frequentemente também contêm antiácidos.

**14- Métrica**: medida da performance do seu modelo. Cada tipo de modelo pode ter uma ou mais métricas, que usamos para comparar a performance entre versões do modelo.

**15- Função Objetivo**: a fórmula matemática ou métrica que um modelo visa otimizar. Por exemplo, a função objetivo para regressão linear geralmente é o erro ao quadrado. Portanto, ao treinar um modelo de regressão linear, o objetivo é minimizar o erro quadrática. Em alguns casos, o objetivo é maximizar a função objetivo. Por exemplo, se a função objetivo é precisão, o objetivo é maximizar a precisão.

**16- AUC (Área sob a Curva ROC)**: uma métrica de avaliação que considera todos os limites de classificação possíveis. A área sob a curva ROC é a probabilidade de um classificador estar mais confiante de que um exemplo positivo escolhido aleatoriamente é realmente positivo, do que um exemplo negativo escolhido aleatoriamente é positivo.

**17- Acurácia**: a fração de previsões que um modelo de classificação acertou. Em um modelo com acurácia de 82% dizemos que o modelo acerta 82 previsões a cada 100 previsões realizadas.

**18- Precisão Média**: uma métrica para resumir o desempenho de uma sequência classificada de resultados. A precisão média é calculada tomando a média dos valores de precisão para cada resultado relevante.

**19- DataFrame**: popular estrutura de dados para representar conjuntos de dados. Um DataFrame é análogo a uma tabela. Cada coluna do DataFrame tem um nome (um cabeçalho) e cada linha é identificada por um número.

**20- Função de Ativação**: uma função (por exemplo, ReLU ou sigmoid) que recebe a soma ponderada de todas as entradas da camada anterior em um modelo de rede neural e, em seguida, gera e passa um valor de saída (normalmente não linear) para a próxima camada.

**21- Backpropagation**: o algoritmo primário para executar gradiente descendente em redes neurais. Primeiro, os valores de saída de cada nó são calculados (e armazenados em cache) em uma passagem direta. Então, a derivada parcial do erro em relação a cada parâmetro é calculada em uma passagem para trás pelo grafo computacional, conforme estudado em detalhes neste livro online.

**22- Batch**: o conjunto de exemplos usados em uma iteração (ou seja, uma atualização de gradiente) do treinamento do modelo.

**23- Batch Normalization**: normalização da entrada ou saída das funções de ativação em uma camada oculta. A normalização em lote pode fornecer os seguintes benefícios: torna as redes neurais mais estáveis protegendo contra pesos discrepantes, habilita taxas de aprendizado mais altas e reduz o sobreajuste (overfitting).

**24- Stochastic Gradient Descent (SGD)**: um algoritmo de descida de gradiente em que o tamanho do lote é um. Em outras palavras, o SGD conta com um único exemplo escolhido uniformemente de forma aleatória de um conjunto de dados para calcular uma estimativa do gradiente em cada etapa.

**25- Overfitting**: o overfiting (sobre ajuste) ocorre quando o modelo aprende os detalhes nos dados de treino. Não é isso que queremos em Machine Learning. Em aprendizado de máquina buscamos a criação de um modelo que aprende a generalização dos dados, para então fazer previsões com novos dados.

**26- Bagging**: um método para treinar um ensemble onde cada modelo constituinte é treinado em um subconjunto aleatório de exemplos de treinamento amostrados com substituição. Por exemplo, uma floresta aleatória (Random Forest) é uma coleção de árvores de decisão treinadas com bagging. O termo bagging é a abreviação de bootstrap aggregating.

**27- Boosting**: uma técnica de aprendizado de máquina que combina iterativamente um conjunto de classificadores simples e não muito precisos (referidos como classificadores “fracos”) em um classificador com alta precisão (um classificador “forte”), valorizando os exemplos que o modelo está classificando incorretamente no momento .

**28- Bag of Words**: uma representação das palavras em uma frase ou passagem, independentemente da ordem. Cada palavra é mapeada para um índice em um vetor esparso, onde o vetor tem um índice para cada palavra do vocabulário. Por exemplo, a frase “o cachorro pula” é mapeada em um vetor de características com valores diferentes de zero nos três índices correspondentes às palavras “o”, “cachorro” e “pula”.

**29- Word Embedding**: estrutura que representa cada palavra em um conjunto de palavras; ou seja, representando cada palavra como um vetor de valores de ponto flutuante entre 0,0 e 1,0. Palavras com significados semelhantes têm representações mais semelhantes do que palavras com significados diferentes. Por exemplo, cenoura, aipo e pepino teriam representações relativamente semelhantes, que seriam muito diferentes das representações de avião, óculos de sol e pasta de dente.

**30- Inteligência Artificial**: um programa ou modelo não humano que pode resolver tarefas sofisticadas. Por exemplo, um programa ou modelo que traduz texto ou um programa ou modelo que identifica doenças a partir de imagens radiológicas exibem Inteligência Artificial. Formalmente, o aprendizado de máquina é um subcampo da Inteligência Artificial. No entanto, nos últimos anos, algumas organizações começaram a usar os termos Inteligência Artificial e aprendizado de máquina de forma intercambiável.

**31- Inteligência Artificial Geral**: um mecanismo não humano que demonstra uma ampla gama de resolução de problemas, criatividade e adaptabilidade. Por exemplo, um programa que demonstra Inteligência Artificial Geral pode traduzir texto, compor sinfonias e se destacar em jogos que ainda não foram inventados.

**32- Atenção**: qualquer um de uma ampla gama de mecanismos de arquitetura de rede neural que agregam informações de um conjunto de entradas de maneira dependente de dados. Um mecanismo de atenção típico pode consistir em uma soma ponderada sobre um conjunto de entradas, onde o peso de cada entrada é calculado por outra parte da rede neural. São os blocos principais dos Transformers, estudados aqui mesmo neste livro online.

**33- Recall**: o recall é a medida do nosso modelo identificando corretamente os Verdadeiros Positivos. Assim, para todos os pacientes que realmente têm doença cardíaca, o recall nos diz quantos identificamos corretamente como tendo uma doença cardíaca.

**34- Precisão**: é a razão entre os Verdadeiros Positivos e todos os Positivos. Para nossa declaração de problema, essa seria a medida de pacientes que identificamos corretamente como tendo uma doença cardíaca entre todos os pacientes que realmente a têm.

**35- Bias (ética/justiça)**: estereótipo, preconceito ou favoritismo em relação a algumas coisas, pessoas ou grupos em detrimento de outros. Esses vieses podem afetar a coleta e interpretação de dados, o design de um sistema e como os usuários interagem com um sistema.

**36- Bias (matemática)**: uma interceptação ou deslocamento de uma origem. O bias (também conhecido como termo de viés) é referido como b ou w0 em modelos de aprendizado de máquina.

**37- Classificação Binária**: um tipo de tarefa de classificação que gera uma das duas classes mutuamente exclusivas. Por exemplo, um modelo de aprendizado de máquina que avalia mensagens de e-mail e gera “spam” ou “não spam” é um classificador binário.

**38- BLEU (Bilingual Evaluation Understudy)**: uma pontuação entre 0,0 e 1,0, inclusive, indicando a qualidade de uma tradução entre duas línguas humanas (por exemplo, entre inglês e russo). Uma pontuação BLEU de 1,0 indica uma tradução perfeita; uma pontuação BLEU de 0,0 indica uma tradução terrível. Medida usada em modelos de Processamento de Linguagem Natural.

**39- Normalização**: o processo de conversão de um intervalo real de valores em um intervalo padrão de valores, normalmente -1 a +1 ou 0 a 1. Por exemplo, suponha que o intervalo natural de um determinado recurso seja de 800 a 6.000. Por meio de subtração e divisão, você pode normalizar esses valores no intervalo -1 a +1.

**40- Normalização Z-score**: uma técnica de normalização que substitui um valor de recurso bruto por um valor de ponto flutuante que representa o número de desvios padrão da média desse recurso.

**41- Entropia Cruzada**: uma generalização de Log Loss para problemas de classificação multiclasse. A entropia cruzada quantifica a diferença entre duas distribuições de probabilidade.

**42- Matriz de Confusão**: uma tabela NxN que agrega as suposições corretas e incorretas de um modelo de classificação. Um eixo de uma matriz de confusão é o rótulo que o modelo previu e o outro eixo é a verdade básica. N representa o número de classes. Por exemplo, N=2 para um modelo de classificação binária.

**43- Validação**: um processo usado, como parte do treinamento, para avaliar a qualidade de um modelo de aprendizado de máquina usando o conjunto de validação. Como o conjunto de validação é separado do conjunto de treinamento, a validação ajuda a garantir que o desempenho do modelo se generalize além do conjunto de treinamento.

**44- Estacionaridade**: uma propriedade de dados em um conjunto de dados, na qual a distribuição de dados permanece constante em uma ou mais dimensões. Mais comumente, essa dimensão é o tempo, o que significa que os dados que exibem estacionariedade não mudam com o tempo. Por exemplo, os dados que exibem estacionariedade não mudam de setembro a dezembro. Termo muito usado em análise de séries temporais.

**45- Não Estacionaridade**: uma propriedade de dados cujos valores mudam em uma ou mais dimensões, geralmente o tempo. Por exemplo, o número de maiôs vendidos em uma determinada loja demonstra não estacionaridade porque esse número varia com a estação. Como segundo exemplo, a quantidade de uma determinada fruta colhida em uma determinada região normalmente mostra uma acentuada não estacionariedade ao longo do tempo. Termo muito usado em análise de séries temporais.

**46- Convergência**: informalmente, muitas vezes se refere a um estado alcançado durante o treinamento no qual a perda de treinamento e a perda de validação mudam muito pouco ou nada a cada iteração após um certo número de iterações. Em outras palavras, um modelo atinge a convergência quando o treinamento adicional nos dados atuais não melhora o modelo. No aprendizado profundo, os valores de perda às vezes permanecem constantes por muitas iterações antes de finalmente descer, produzindo temporariamente uma falsa sensação de convergência.

**47- Desbalanceamento de Classe:** um problema de classificação binária em que os rótulos para as duas classes têm frequências significativamente diferentes. Por exemplo, um conjunto de dados de doenças em que 0,0001 dos exemplos tem rótulos positivos e 0,9999 têm rótulos negativos é um problema de desequilíbrio de classe, mas um preditor de jogo de futebol no qual 0,51 dos exemplos rotula um time vencedor e 0,49 rotula o outro time vencedor não é um problema de classe desequilibrada.

**48- Limite (Threshold) de Classificação**: um critério de valor escalar que é comparado com a pontuação prevista de um modelo para separar a classe positiva da classe negativa. Usado ao mapear resultados de regressão logística para classificação binária. Por exemplo, considere um modelo de regressão logística que determina a probabilidade de uma determinada mensagem de email ser spam. Se o limite de classificação for 0,9, os valores de regressão logística acima de 0,9 são classificados como spam e os abaixo de 0,9 são classificados como não spam.

**49- Checkpoint**: captura o estado dos parâmetros de um modelo em um determinado momento. Os pontos de verificação permitem exportar pesos de modelo ou realizar treinamento em várias sessões. Os pontos de verificação também permitem que o treinamento continue com erros anteriores.

















