

Redes Neurais



Material Teórico



Arquitetura das Redes Neurais

Responsável pelo Conteúdo:

Prof. Dr. Alberto Messias da Costa Souza

Revisão Textual:

Prof. Me. Luciano Vieira Francisco

UNIDADE

Arquitetura das Redes Neurais



- Organização em Camadas;
- Convenções de Nomenclatura;
- Definindo o Número de Camadas e os seus Tamanhos.



OBJETIVO DE APRENDIZADO

- Conhecer o conceito de camadas de redes neurais, bem como dimensionar o tamanho ou o número de camadas nas redes neurais.



Orientações de estudo

Para que o conteúdo desta Disciplina seja bem aproveitado e haja maior aplicabilidade na sua formação acadêmica e atuação profissional, siga algumas recomendações básicas:



Assim:

- ✓ Organize seus estudos de maneira que passem a fazer parte da sua rotina. Por exemplo, você poderá determinar um dia e horário fixos como seu “momento do estudo”;
- ✓ Procure se alimentar e se hidratar quando for estudar; lembre-se de que uma alimentação saudável pode proporcionar melhor aproveitamento do estudo;
- ✓ No material de cada Unidade, há leituras indicadas e, entre elas, artigos científicos, livros, vídeos e sites para aprofundar os conhecimentos adquiridos ao longo da Unidade. Além disso, você também encontrará sugestões de conteúdo extra no item **Material Complementar**, que ampliarão sua interpretação e auxiliarão no pleno entendimento dos temas abordados;
- ✓ Após o contato com o conteúdo proposto, participe dos debates mediados em fóruns de discussão, pois irão auxiliar a verificar o quanto você absorveu de conhecimento, além de propiciar o contato com seus colegas e tutores, o que se apresenta como rico espaço de troca de ideias e de aprendizagem.

Organização em Camadas

Redes neurais como neurônios em gráficos, as redes neurais são modeladas como coleções de neurônios conectados em um gráfico acíclico. Em outras palavras, as saídas de alguns neurônios podem se tornar entradas para outros neurônios. Ciclos não são permitidos, pois implicariam em *loops* infinitos na passagem direta de uma rede.

Em vez de uma bolha amorfia de neurônios conectados, os modelos da rede neural são frequentemente organizados em camadas distintas de neurônios. Para redes neurais regulares, a forma de camada mais comum é a totalmente conectada, na qual os neurônios entre duas camadas adjacentes são completamente conectados em pares, mas os neurônios em uma única camada não compartilham conexões – a seguir temos dois exemplos de topologias de redes neurais que usam uma pilha de camadas totalmente conectadas:

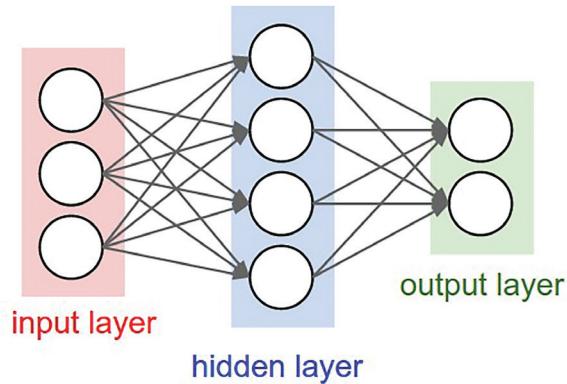


Figura 1 – Exemplo de rede neural com camadas

A Figura 1 ilustra uma rede neural de duas camadas (uma oculta de quatro neurônios (ou unidades) e uma camada de saída com dois neurônios) e três entradas.

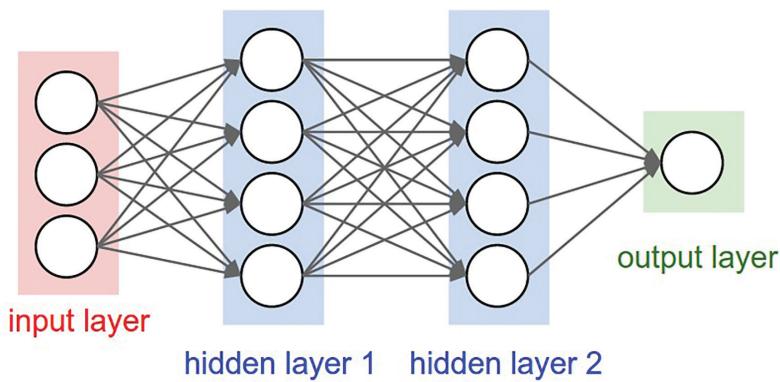


Figura 2 – Exemplo de rede neural com camadas ocultas

Já a Figura 2 ilustra uma rede neural de três camadas com três entradas, duas camadas ocultas de quatro neurônios, cada, e uma camada de saída. Observe que nos dois casos existem conexões (sinapses) entre neurônios nas camadas, mas não dentro de uma camada.

Convenções de Nomenclatura

Observe que, quando dizemos **rede neural de camada N**, não contamos a camada de entrada. Portanto, uma rede neural de camada única descreve uma rede sem camadas ocultas (entrada mapeada diretamente para a saída). Nesse sentido, às vezes você pode ouvir as pessoas dizerem que a regressão logística ou SVM é simplesmente um caso especial de redes neurais de camada única.

Em algumas literaturas as redes neurais podem ser denominadas **redes neurais artificiais** (ANN) ou **perceptrons de várias camadas** (MLP). Muitas pessoas não gostam das analogias entre redes neurais e cérebros reais e preferem se referir aos neurônios como unidades.

Camada de Saída

Ao contrário de todas as camadas de uma rede neural, os neurônios da camada de saída geralmente não têm função de ativação (ou você pode pensar nesses como tendo função de ativação de identidade linear). Isso ocorre porque a última camada de saída geralmente é usada para representar as pontuações da classe (por exemplo, na classificação), que são números arbitrários com valor real ou algum tipo de alvo com valor real (por exemplo, em regressão).

Dimensionando Redes Neurais

As duas métricas que as pessoas geralmente usam para medir o tamanho das redes neurais são o número de neurônios ou, mais comumente, o número de parâmetros.

Então, trabalhando com as duas redes das figuras anteriores, temos a:

- Primeira rede, da Figura 1, possuindo $4 + 2 = 6$ neurônios (sem contar as entradas), $[3 \times 4] + [4 \times 2] = 20$ pesos e $4 + 2 = 6$ desvios, para um total de 26 parâmetros aprendíveis;
- Segunda rede, da Figura 2, possui $4 + 4 + 1 = 9$ neurônios, $[3 \times 4] + [4 \times 4] + [4 \times 1] = 12 + 16 + 4 = 32$ pesos e $4 + 4 + 1 = 9$ preconceitos, para um total de 41 parâmetros aprendíveis.

Para contextualizar, as redes convolucionais modernas contêm ordens de 100 milhões de parâmetros e geralmente são compostas por aproximadamente de 10 a 20 camadas (portanto, aprendizado profundo). No entanto, como veremos, o número de conexões efetivas é significativamente maior devido ao compartilhamento de parâmetros.

Computação Feed-Forward

Multiplicações de matriz repetidas e entrelaçadas com função de ativação, uma das principais razões pelas quais as redes neurais são organizadas em camadas é que essa estrutura torna simples e eficiente avaliar redes neurais usando operações de vetores matriciais.

Trabalhando com a rede neural de três camadas de exemplo no diagrama anterior, a entrada seria um vetor $[3 \times 1]$. Todos os pontos fortes da conexão para uma camada podem ser armazenados em uma única matriz. Por exemplo, os pesos da primeira camada ocultam W_1 e seriam do tamanho $[4 \times 3]$ e os desvios para todas as unidades estariam no vetor b_1 , do tamanho $[4 \times 1]$. Aqui, cada neurônio tem os seus pesos seguidos de W_1 , de modo que a multiplicação do vetor da matriz $np.dot(W_1, x)$ avalia as ativações de todos os neurônios nessa camada. Similarmente, W_2 seria uma matriz $[4 \times 4]$ que armazena as conexões da segunda camada oculta e W_3 seria uma matriz $[1 \times 4]$ para a última camada (saída).

A passagem para frente completa dessa rede neural de três camadas é simplesmente três multiplicações de matrizes, entrelaçadas com a aplicação da seguinte função de ativação:

```
# passagem direta de uma rede neural de 3 camadas:  
f = lambda x: 1.0 / (1.0 + np.exp (-x)) # função de ativação (use sigmoid)  
x = np.random.random (3, 1) # vetor de entrada aleatória de três números (3 x 1)  
h1 = f(np.dot (W1, x) + b1) # calcula as ativações da primeira camada oculta (4 x 1)  
h2 = f(np.dot (W2, h1) + b2) # calcula as ativações da segunda camada oculta (4 x 1)  
out = np.dot (W3, h2) + b3 # neurônio de saída (1 x 1)
```

Neste pseudocódigo, W_1 , W_2 , W_3 , b_1 , b_2 , b_3 são os parâmetros aprendíveis da rede.

Observe também que, em vez de ter um único vetor de coluna de entrada, a variável x poderia conter um lote inteiro de dados de treinamento (onde cada exemplo de entrada seria uma coluna de x) e, em seguida, todos os exemplos seriam eficientemente avaliados em paralelo.

Note ainda que a camada final da rede neural geralmente não possui função de ativação (por exemplo, representa uma pontuação de classe (com valor real) em uma configuração de classificação).

Poder Representacional

Uma maneira de observar as redes neurais com camadas totalmente conectadas é que essas definem uma família de funções que são parametrizadas pelos pesos da rede.

Assim, questões naturais colocadas são as seguintes: **Qual é o poder representacional dessa família de funções? Em particular, existem funções que não podem ser modeladas com uma rede neural?**

Acontece que as redes neurais com, pelo menos, uma camada oculta são códigos de aproximação universais. Ou seja, pode ser mostrado através desta explicação intuitiva que é, na verdade, uma função contínua $f(X)$ e alguns $\varepsilon > 0$, existindo uma rede neural $g(X)$ com uma camada oculta (com uma escolha razoável de não

linearidade, por exemplo, sigmoide) tal que $\forall x, |f(x) - g(x)| < \epsilon$. Em outras palavras, a rede neural pode se aproximar de qualquer função contínua, a partir de suas iterações de aprendizagem.

Se uma camada oculta é suficiente para aproximar qualquer função, por que usar mais camadas e ir mais fundo? A resposta é que o fato de uma rede neural de duas camadas ser um aproximador universal é, embora matematicamente atraente, uma declaração relativamente fraca e inútil na prática. Em uma dimensão, a função de soma dos saltos dos indicadores, ou de descida do gradiente é: $g(x) = \sum_i c_i I(a_i < x < b_i)$, onde a , b e c são vetores de parâmetros. Igualmente, trata-se de um aproximador universal, mas ninguém sugeriria que usássemos essa forma funcional no *machine learning*.

As redes neurais funcionam bem na prática porque expressam compactamente funções agradáveis e suaves que se encaixam bem nas propriedades estatísticas dos dados que encontramos na prática e são fáceis de aprender usando os nossos algoritmos de otimização (por exemplo, descida em gradiente). Da mesma forma, o fato de redes mais profundas (com várias camadas ocultas) poderem funcionar melhor do que redes de camada única oculta é uma observação empírica, apesar de seu poder representacional ser igual.

Como aparte, na prática, geralmente é o caso de redes neurais de três camadas superarem as redes de duas camadas, mas ir ainda mais fundo (em quatro, cinco ou seis camadas) raramente ajuda. Isso contrasta fortemente com as redes convolucionais, onde a profundidade foi considerada um componente extremamente importante para um bom sistema de reconhecimento (por exemplo, na ordem de dez camadas de aprendizagem).

Um argumento para tal observação é que as imagens contêm estrutura hierárquica (por exemplo, as faces são compostas de olhos, estes que são constituídos de bordas etc.); portanto, várias camadas de processamento fazem sentido intuitivo para esse domínio de dados.

Definindo o Número de Camadas e os seus Tamanhos

Como decidimos sobre qual arquitetura usar quando confrontados com um problema prático? Não devemos usar camadas ocultas? Uma camada oculta? Duas camadas ocultas? Qual é o tamanho de cada camada?

Primeiro, observe que, à medida que aumentamos o tamanho e número de camadas em uma rede neural, a capacidade da rede aumenta. Ou seja, o espaço das funções representáveis aumenta, pois os neurônios podem colaborar para expressar muitas funções diferentes. Por exemplo, suponha que tenhamos um problema de classificação binária em duas dimensões: poderíamos treinar três redes neurais

separadas, cada qual com uma camada oculta de algum tamanho e obter os seguintes classificadores:



Figura 3 – Resultado da classificação com uma rede com três neurônios ocultos



Figura 4 – Resultado da classificação com uma rede com seis neurônios ocultos

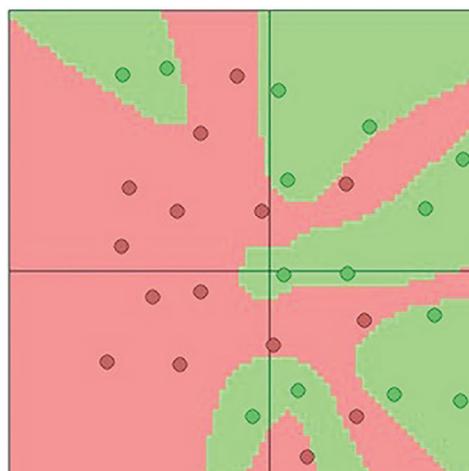


Figura 5 – Resultado da classificação com uma rede com vinte neurônios ocultos

Redes neurais maiores podem representar funções mais complicadas. Os dados são mostrados como círculos coloridos por sua classe e as regiões de decisão por uma rede neural treinada conforme o observado nas figuras 3, 4 e 5; a precisão da classificação aumenta à medida que se amplia o número de neurônios ou camadas ocultas.

No diagrama anterior, podemos ver que redes neurais com mais neurônios podem expressar funções mais complicadas. No entanto, isso é um benefício, dado que podemos aprender a classificar dados mais complicados; assim como uma dificuldade, pois é mais fácil superestimar os dados de treinamento.

O sobreajuste ocorre quando um modelo com alta capacidade ajusta o ruído nos dados em vez do relacionamento subjacente (assumido). Por exemplo, o modelo com vinte neurônios ocultos se encaixa em todos os dados de treinamento, mas com o custo de segmentar o espaço em várias regiões vermelhas e verdes de decisão.

O modelo com três neurônios ocultos tem apenas o poder representacional para classificar os dados em traços amplos; modela os dados como dois grupos e interpreta os poucos pontos vermelhos dentro do *cluster* verde como *outliers* (pontos fora da curva). Na prática, isso pode levar a uma melhor generalização no conjunto de testes.

Com base em nossa discussão, parece que redes neurais menores podem ser preferidas se os dados não forem complexos o suficiente para impedir o ajuste excessivo. No entanto, isso está incorreto – existem muitas outras maneiras preferidas de evitar o ajuste excessivo nas redes neurais e que discutiremos adiante (tais como a regularização L2, desistência e o ruído de entrada). Na prática, é sempre melhor usar esses métodos para controlar a superadaptação, em vez do número de neurônios.

A razão sutil por trás disso é que redes menores são mais difíceis de treinar com métodos locais, tal como o *gradient descent* ou a descida de gradiente: é claro que as suas funções de perda têm relativamente poucos mínimos locais, mas acontece que muitos desses mínimos são mais fáceis de convergir e são ruins (ou seja, apresentam alta perda). Por outro lado, redes neurais maiores contêm mínimos significativamente mais locais, contudo, melhores em termos de perda real.

Como as redes neurais não são convexas, é difícil estudar essas propriedades matematicamente, mas algumas tentativas de entender essas funções objetivas foram feitas. Na prática, o que você descobre é que se treina uma pequena rede, a perda final pode mostrar uma boa variação – em alguns casos, você tem sorte e converge para um bom lugar; todavia, em outros, fica preso em um dos mínimos ruins.

Ademais, se você treinar uma rede grande, começará a encontrar muitas soluções diferentes, mas a variação na perda final alcançada será consideravelmente menor. Em outras palavras, todas as soluções são igualmente boas e dependem menos da sorte da inicialização aleatória.

Para reiterar, a força da regularização é a maneira preferida de controlar o ajuste excessivo de uma rede neural. Em nível de exemplo, podemos observar os resultados alcançados por três configurações diferentes:

$$\lambda = 0.001$$

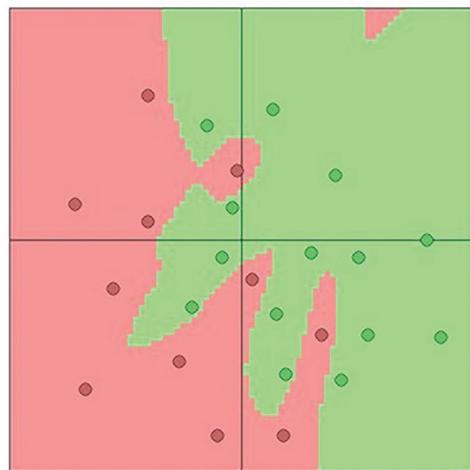


Figura 6 – Resultado do processamento da rede com menor variação

$$\lambda = 0.01$$

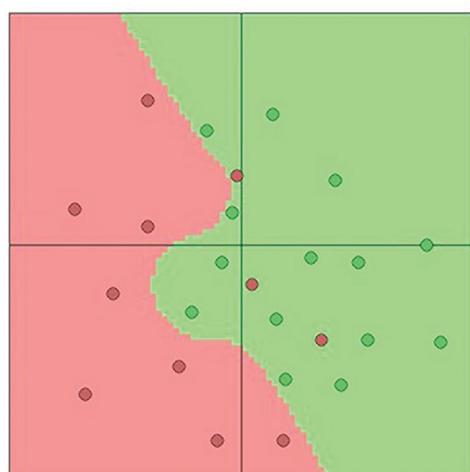


Figura 7 – Resultado do processamento da rede com variação intermediária

$$\lambda = 0.1$$



Figura 8 – Resultado do processamento da rede com maior variação

Os efeitos da força de regularização: cada rede neural aqui ilustrada possui vinte neurônios ocultos, mas alterar a força de regularização torna as suas regiões de decisão final mais suaves com maior regularização.

O ponto principal é que você não deve usar redes menores porque tem medo de se adaptar demais; em vez disso, deve usar uma rede neural tão grande quanto o seu orçamento computacional permitir, assim como adotar outras técnicas de regularização para controlar o ajuste excessivo.

É importante você pensar em ajustar o modelo de rede para cada problema a ser resolvido, ou ainda testar e refinar a rede com menor volume de dados, analisando os resultados e tempos de processamento.

Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:



Livros

Inteligência Artificial

Silva, F. M. da, *et al.* **Inteligência Artificial**. Porto Alegre: SAGAH, 2019.
Capítulo sobre as redes multicamadas (que se inicia na página 138) do livro.

Inteligência Artificial

RUSSEL, S. J.; NORVIG, P. **Inteligência artificial**. Trad. Regina Célia Smille. 3. ed. Rio de janeiro: Elsevier, 2013.
Leia o 18º capítulo do livro.



Leitura

Quantas camadas escondidas e quantos neurônios incluir em uma rede neural artificial?

<https://bit.ly/32K9rde>

Dicas para a configuração de redes neurais

<https://bit.ly/3jz3hUh>

Referências

BISPO, F., et al. **Inteligência artificial**. [S.l.]: Grupo A, 2019. Disponível em: <<https://integrada.minhabiblioteca.com.br/#/books/9788595029392>>. Acesso em: 24/01/2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT, 2016. Disponível em: <<http://www.deeplearningbook.org>>. Acesso em: 21/11/ 2019.

NORVIG, P. **Inteligência artificial**. [S.l.]: Grupo GEN, 2013. Disponível em: <<https://integrada.minhabiblioteca.com.br/#/books/9788595156104>>. Acesso em: 24/01/2020.

RICHARD, O. D.; PETER, E. H.; DAVID, G. S. **Pattern classification**. 2. ed. [S.l.]: Wiley, 2000.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. 4th ed. USA: Academic Press, 2008.



Cruzeiro do Sul
Educacional