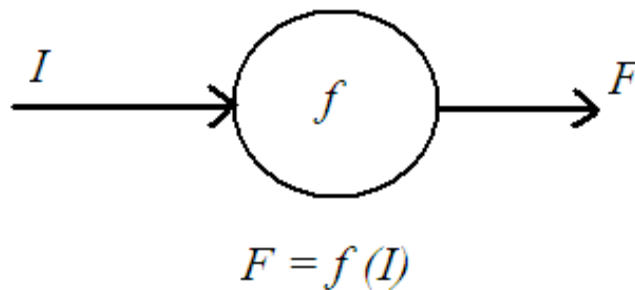


Do neurônio biológico ao neurônio das redes neurais artificiais

O objetivo desta aula é procurar justificar o modelo de neurônio usado pelas redes neurais artificiais em termos das propriedades essenciais dos neurônios biológicos apresentadas na aula de revisão sobre neurônios.

Modelando um neurônio como um elemento que recebe uma entrada (a corrente) e fornece uma saída (a frequência de disparos), a função F-I pode ser vista como a função de transferência ou ganho do neurônio, que dá a sua relação entrada-saída.



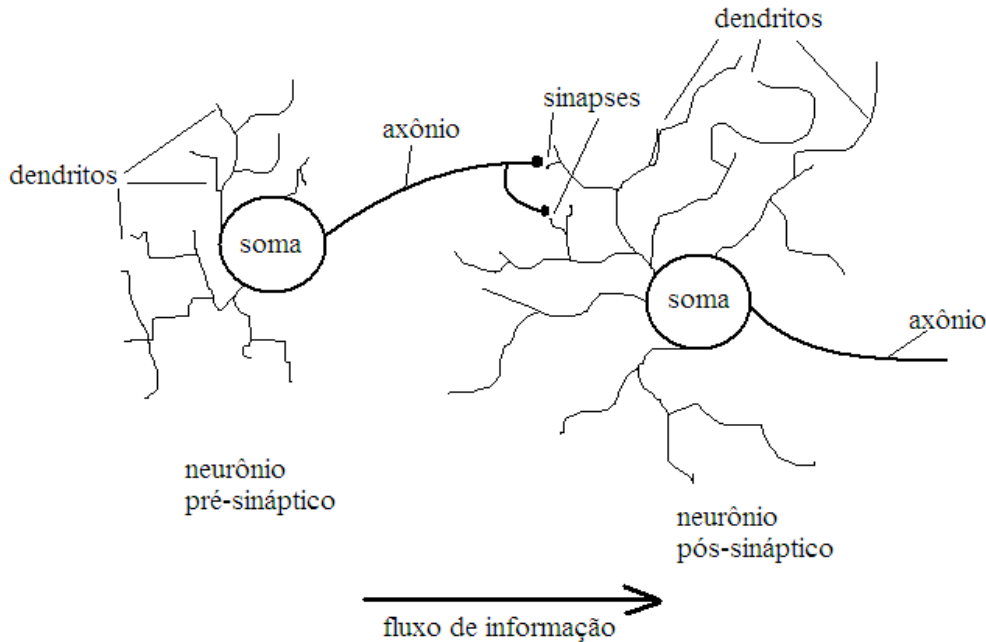
Numa situação experimental, *in vitro*, é fácil injetar corrente em um neurônio para medir sua resposta. Porém, numa situação real, *in vivo*, um neurônio está recebendo estímulos de um grande número de outros neurônios (e, às vezes, até dele mesmo).

Como fazer para interpretar a noção de função de transferência de um neurônio numa situação em que ele está inserido em uma rede?

Para isso, é necessário darmos uma olhada mais detalhada na maneira como um neurônio afeta um outro, pela sinapse.

Há dois tipos de sinapses, químicas e elétricas. Vamos considerar apenas a sinapse química, que é considerada a mais importante segundo a visão padrão que se tem sobre o sistema nervoso.

Uma sinapse química padrão conecta o axônio do neurônio que envia o estímulo, chamado de neurônio pré-sináptico, a um dendrito do neurônio que recebe o estímulo, chamado de neurônio pós-sináptico (veja a figura abaixo).

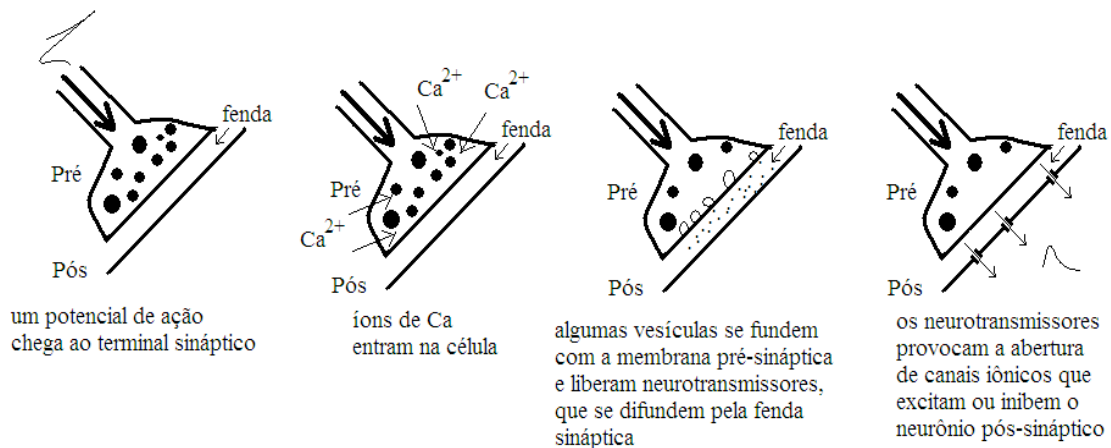


Quando um potencial de ação chega ao terminal do axônio do neurônio pré-sináptico, uma série de eventos acontece:

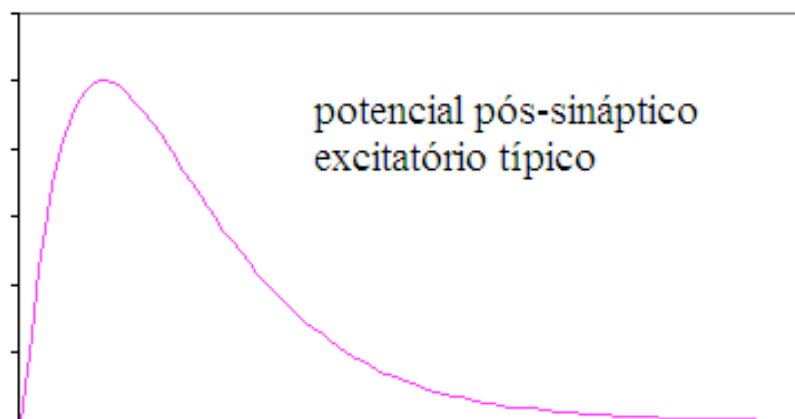
- Canais de cálcio na membrana do terminal pré-sináptico se abrem e íons de Ca^{2+} entram na célula pré-sináptica;
- Os íons de Ca^{2+} provocam a fusão de vesículas que contêm neurotransmissores com a membrana pré-sináptica, liberando esses neurotransmissores na fenda sináptica;
- Os neurotransmissores se difundem pela fenda sináptica e se ligam a receptores na membrana do dendrito do neurônio pós-sináptico;
- Dependendo do tipo de neurotransmissor, abrem-se canais iônicos na membrana do dendrito pós-sináptico que permitem a entrada de íons que, ou provocam uma pequena despolarização local na membrana, ou provocam uma pequena hiperpolarização local na membrana;

Uma despolarização local na membrana é chamada de potencial pós-sináptico excitatório e uma hiperpolarização local é chamada de potencial pós-sináptico inibitório;

Um neurônio pré-sináptico sempre libera o mesmo tipo de neurotransmissor: quando ele provoca uma despolarização local, o neurônio pré-sináptico é chamado de excitatório e a sinapse é dita excitatória; quando ele provoca uma hiperpolarização local, o neurônio pré-sináptico é chamado de inibitório e a sinapse é dita inibitória.



Os potenciais pós-sinápticos (excitatórios ou inibitórios) têm durações muito maiores que a duração de um potencial de ação. Um potencial pós-sináptico típico tem uma fase de subida que leva de 1 a 2 ms e um tempo de decaimento mais lento, que leva de 3 a 5 ms. A figura abaixo ilustra um potencial pós-sináptico típico.



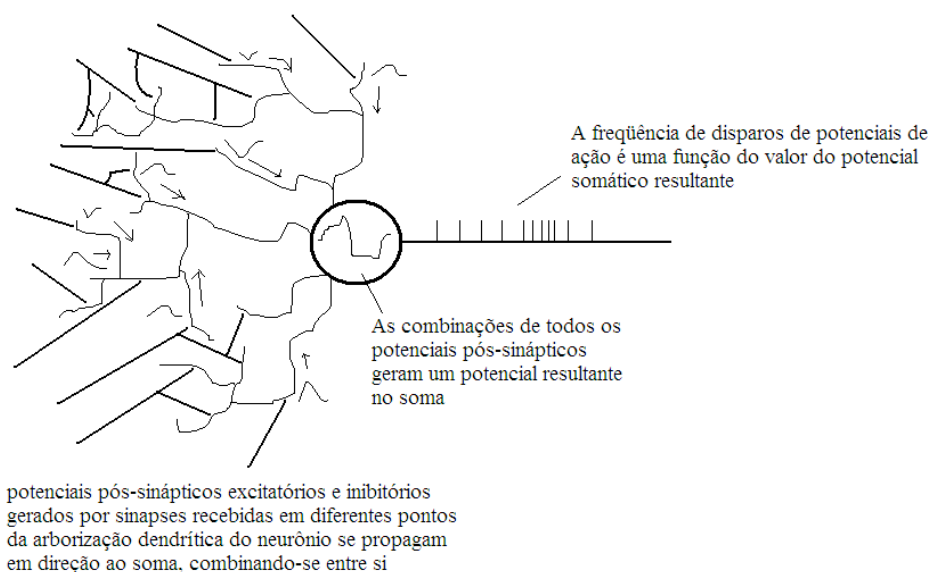
Um neurônio típico recebe um grande número de sinapses de outros neurônios, algumas excitatórias e outras inibitórias, através da sua arborização dendrítica. Cada uma delas gera um potencial pós-sináptico excitatório ou inibitório.

É muito raro encontrar um potencial de ação gerado em um dendrito. Em geral, eles são produzidos em uma estrutura denominada cone axônico, que é uma região de alta densidade de canais de sódio onde o axônio se conecta ao corpo celular.

A cada instante de tempo, os diversos potenciais pós-sinápticos produzidos nos diferentes pontos da ramificação dendrítica de um neurônio, causados pelas sinapses que ele recebe de outros neurônios, propagam-se através da arborização dendrítica em direção ao corpo celular.

Como eles não são potenciais de ação, eles sofrem atenuações ao longo da sua propagação. Porém, quando dois potenciais pós-sinápticos excitatórios ou dois potenciais pós-sinápticos inibitórios se encontram, eles podem se amplificar; já quando um potencial pós-sináptico excitatório se encontra com um inibitório, eles podem se cancelar.

A figura abaixo ilustra o processo de somação espaço-temporal que ocorre na arborização dendrítica de um neurônio típico.



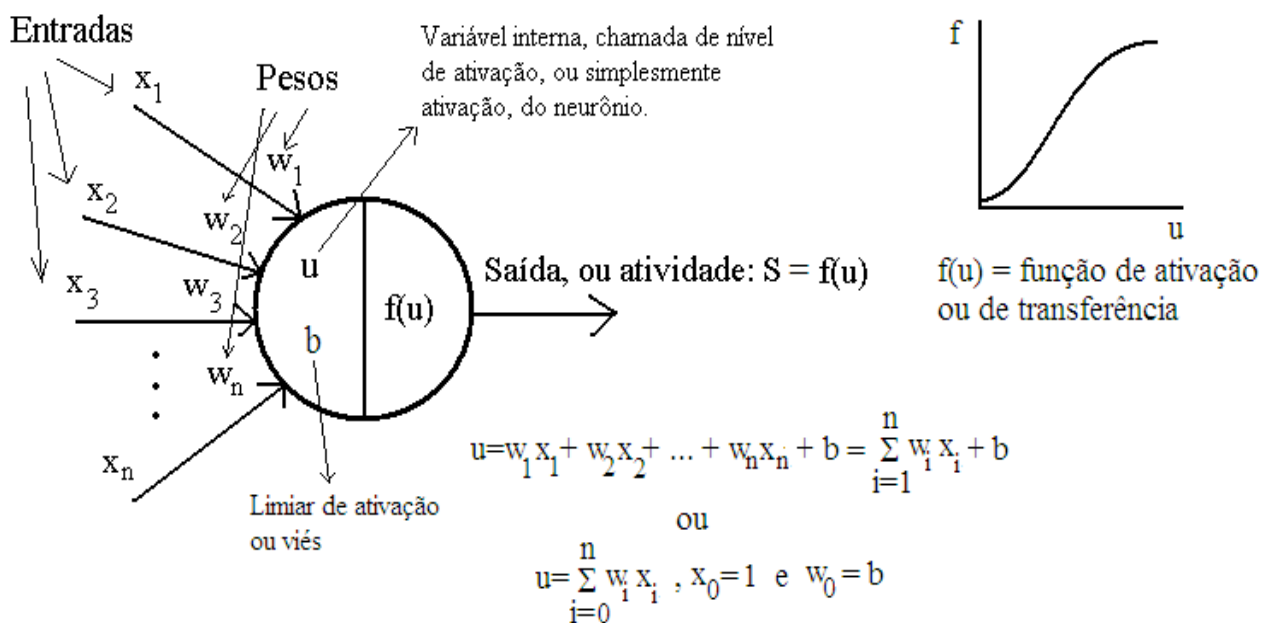
Ao se propagar em direção ao soma, os potenciais pós-sinápticos se combinam de diversas maneiras, gerando um potencial resultante no soma que pode ser visto como um “potencial filtrado” a partir dos vários potenciais pós-sinápticos recebidos num dado instante de tempo.

Segundo o modelo padrão de um neurônio, adotado pelas redes neurais artificiais, é esse potencial somático resultante que funciona como entrada para o neurônio produzir uma frequência de disparos de potenciais de ação.

O Neurônio das Redes Neurais Artificiais

O modelo de neurônio usado pelas redes neurais artificiais está baseado na concepção de neurônio biológico apresentada acima: ele é uma unidade que recebe muitas entradas, integra-as segundo alguma regra e fornece uma saída que é dada por uma função (de transferência) do valor integrado.

O modelo de neurônio utilizado pelas redes neurais artificiais tem dois estágios de operação, simbolizados pelo esquema abaixo.



No primeiro estágio, ele calcula o chamado nível de ativação, u , que faz o papel do potencial resultante no soma de um neurônio biológico.

O nível de ativação u é dado por uma combinação linear dos sinais de entrada vindos de outros neurônios, x_i , $i = 1, \dots, n$, (análogos aos potenciais pós-sinápticos) cujos coeficientes, w_i , $i = 1, \dots, n$, são os chamados pesos sinápticos. A esta combinação linear é somado um número b , positivo ou negativo, que é chamado de limiar de ativação ou viés (*bias*) do neurônio.

Os pesos sinápticos são números, positivos ou negativos, dando a força, ou eficácia, do acoplamento sináptico entre um neurônio que envia o sinal e o neurônio que o recebe. Quando o peso w de uma sinapse é positivo, ela é dita excitatória; quando ele é negativo, ela é dita inibitória.

Matematicamente, o viés de um neurônio pode ser tratado como um peso sináptico extra, $w_0 = b$, que pondera uma entrada constante $x_0 = +1$. Isto torna mais concisa e simétrica a expressão para o cálculo de u ,

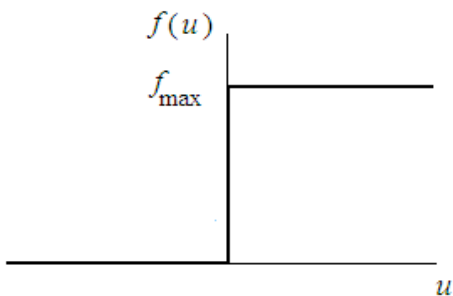
$$u = \sum_{i=1}^n w_i x_i + b = \sum_{i=0}^n w_i x_i .$$

No segundo estágio de operação, uma saída é calculada em função do nível de ativação u . Esta saída, que é um número S , é dada por uma função $f(u)$, onde $f(x)$ é chamada de função de ativação, ou de transferência, do neurônio.

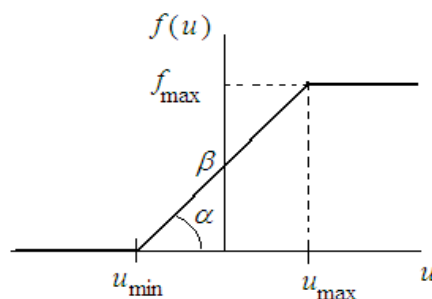
A conversão que o neurônio das redes neurais artificiais faz do nível de atividade u para a saída $S = f(u)$ é análoga à que o modelo padrão de neurônio biológico faz entre o valor de voltagem local no soma e a sua frequência de disparos de potenciais de ação.

Os três tipos básicos de função de ativação $f(u)$ usados pelos modelos de redes neurais artificiais são:

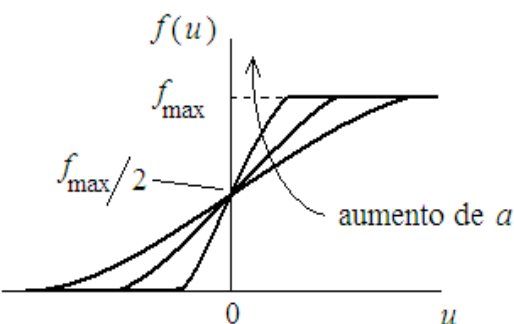
- Função degrau (função de Heaviside):

$$S = f(u) = \begin{cases} f_{\max} & \text{se } u \geq 0 \\ 0 & \text{se } u < 0 \end{cases}$$


- Função linear por partes:

$$S = f(u) = \begin{cases} f_{\max} & \text{se } u \geq u_{\max} \\ \alpha u + \beta & \text{se } u_{\min} < u < u_{\max} \\ 0 & \text{se } u \leq u_{\min} \end{cases}$$


- Função sigmoidal (o exemplo usa a chamada função logística):

$$S = f(u) = \frac{f_{\max}}{1 + \exp(-au)}$$


Esses três tipos de função de ativação capturam detalhes observados em diferentes tipos de curvas $F-I$ de neurônios biológicos. Dependendo do tipo de problema ou aplicação, uma ou outra delas pode ser mais conveniente de ser usada.

Em geral, as curvas $F-I$ dos neurônios biológicos são não-lineares saturantes e o uso de uma função de transferência sigmoideal é o mais recomendado. Do ponto de vista das aplicações, o uso de unidades cujas saídas sejam funções não-lineares das suas entradas (e diferenciáveis, como no caso da função logística) traz grandes vantagens, como veremos ao longo do curso.

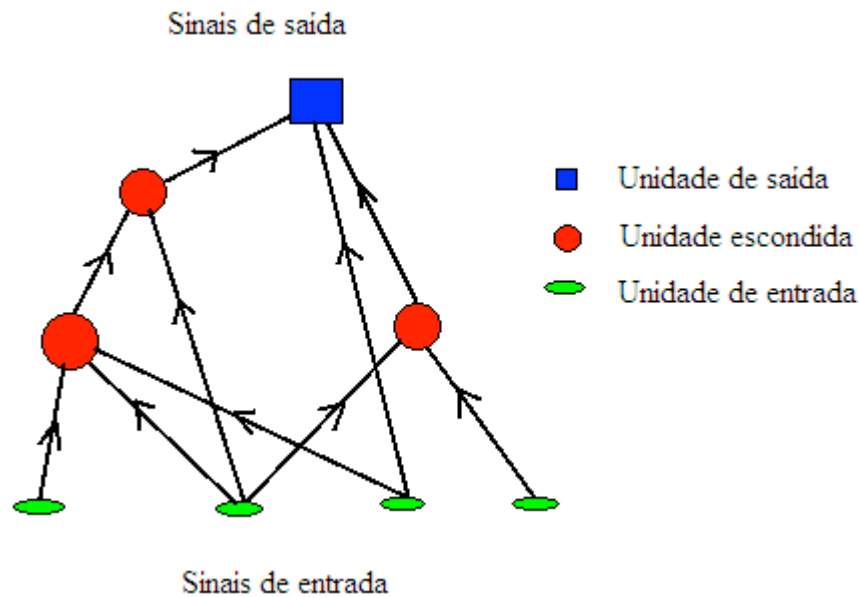
Porém, para muitas faixas de valores experimentais a resposta em frequência F de um neurônio é aproximadamente uma função linear do estímulo I . Nestes casos, uma função linear por partes pode ser usada. Ademais, em muitas aplicações práticas as unidades lineares funcionam muito bem e são mais simples de serem tratadas matematicamente.

Já uma função do tipo degrau, embora seja a menos realista das três, enfatiza o fato de que a função de transferência de um neurônio é não-linear. Ela é a mais “radical” das funções não-lineares, levando a frequência de disparos de um neurônio instantaneamente do valor zero para o valor máximo. O seu uso assume que um neurônio só tem dois estados de disparo possíveis: (1) ausência de disparos, ou (2) disparando com a frequência máxima. O uso de um modelo de unidade binária facilita bastante a análise matemática do sistema e permite muitas analogias entre as redes neurais artificiais e sistemas físicos.

Por motivos históricos (lembre-se da aula 2), um neurônio cuja função de ativação é do tipo degrau é chamado de neurônio de McCulloch-Pitts.

As Redes Neurais Artificiais

A partir do modelo de um neurônio artificial pode-se pensar em construir uma rede neural artificial (ou simplesmente rede neural, pois a diferença entre ela e um modelo para redes de neurônios biológicos está óbvia agora). A figura abaixo dá um exemplo de uma rede neural.



As setas indicam as conexões sinápticas entre os neurônios. A cada conexão atribui-se um peso sináptico w .

Quando uma rede como a do desenho recebe um padrão de entrada, cada um dos neurônios de entrada recebe uma pequena parte desse padrão. Portanto, o padrão apresentado na entrada é representado de maneira distribuída pelas unidades de entrada. Em geral, os neurônios de entrada não executam nenhum tipo de processamento sobre os sinais que recebem, apenas passando-os adiante.

Cada neurônio de entrada envia sinais para alguns neurônios chamados de escondidos (ou ocultos). Eles são assim chamados porque são inspirados nos interneurônios do cérebro, aqueles que não estão diretamente conectados com o mundo exterior (nem pela entrada, nem pela saída). Eles estão *escondidos* do mundo exterior à rede neural.

O conjunto dos neurônios escondidos também recebe todo o padrão de entrada, só que agora ele está modificado pelos pesos que multiplicam os sinais vindos dos neurônios de entrada.

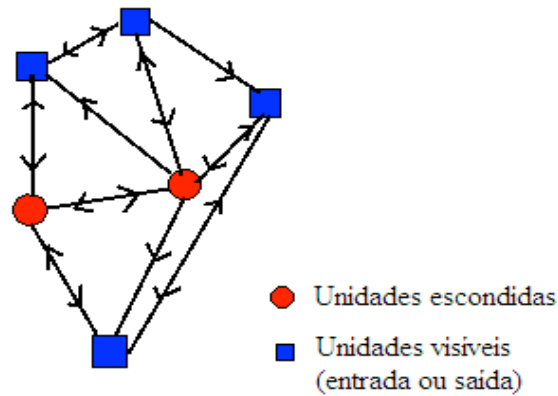
Isto faz com que o conjunto dos neurônios escondidos construa uma representação do padrão de entrada diferente da feita pelos neurônios de entrada. Essa representação é codificada pelos níveis de ativação (os valores dos *us*) dos neurônios escondidos.

As saídas dos neurônios escondidos, que são transformações (lineares ou não-lineares, dependendo das suas funções de transferência) dos seus níveis de ativação, são enviadas para os neurônios de saída e também multiplicadas por pesos.

Os neurônios de saída combinam toda a informação fornecida pelos neurônios escondidos e fornecem saídas que, como um todo, correspondem a algum tipo de operação feita pela rede neural: por exemplo, controlar um movimento, reconhecer ou classificar um padrão, prever o estado futuro de um sistema dado o estado atual, etc.

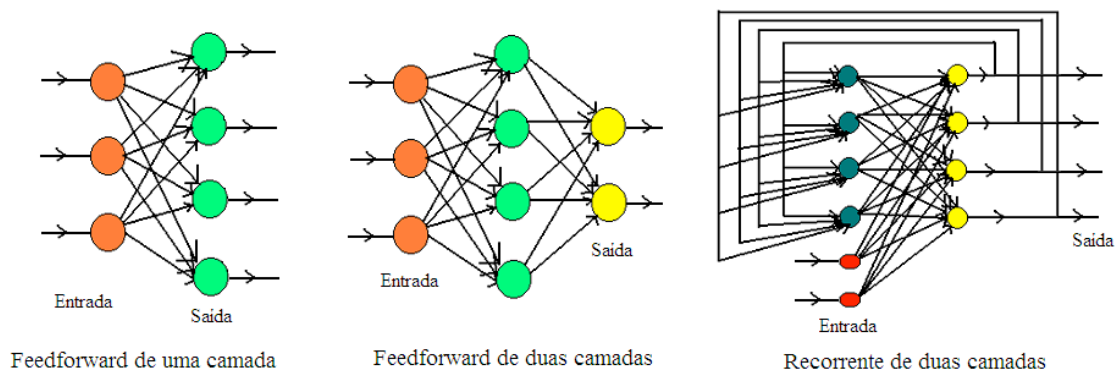
A maneira com os neurônios de uma rede neural estão conectados entre si é chamada de arquitetura da rede neural. De maneira geral, há dois tipos de arquitetura para uma rede neural.

1. Com alimentação para frente (*feedforward*): a característica básica desse tipo de rede é que ela não apresenta realimentação (*feedback*). Começando a partir de qualquer unidade da rede e seguindo as direções das conexões sinápticas chega-se a uma das unidades de saída, e esta não está conectada a outra unidade da rede. A rede neural do exemplo dado anteriormente é uma rede desse tipo.
2. Recorrente: a característica básica deste tipo de rede é que pelo menos um dos neurônios tem um laço de realimentação em que um dos caminhos sinápticos a partir da sua saída leva de volta a ele. Um exemplo de rede recorrente é dado a seguir.



Exemplo de Rede Neural Recorrente

Uma classe importante de redes neurais é a das que têm arquitetura em camadas. Elas podem ser do tipo feedforward ou do tipo recorrente (veja abaixo).



Note que a camada dos neurônios de entrada não é contada como uma das camadas da rede. Isto porque os neurônios de entrada não executam qualquer operação sobre o padrão de entrada que é apresentado à rede. Eles apenas o representam de uma maneira distribuída. A “operação” propriamente dita de uma rede neural começa na camada seguinte à de entrada, quando os sinais de entrada são multiplicados por pesos e somados para gerar os níveis de ativação dos neurônios dessa camada.

Dada uma rede neural, com uma certa arquitetura e formada por neurônios lineares ou não-lineares, o seu objetivo está em executar uma determinada função, como, por exemplo, reconhecer padrões, modelar uma ação etc.

Uma rede neural biológica (o nosso cérebro, por exemplo) não nasce sabendo executar todas as funções que ela tem potencial para executar. É necessário um período de aprendizado para tal.

Em geral, as redes neurais biológicas aprendem a partir de exemplos: vários exemplos vão sendo apresentados à rede biológica até que, em algum momento, ela consegue generalizar e aprender a regra que está por trás deles.

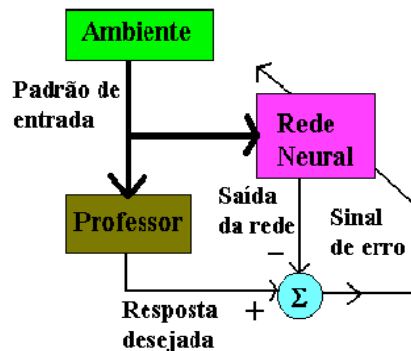
Acredita-se que um dos mecanismos responsáveis pelo aprendizado biológico esteja associado a modificações nas eficácias das sinapses ocorrendo durante o processo de aprendizagem (lembre-se da aula 2). Essas modificações se manifestariam, por exemplo, em termos de um número maior ou menor de vesículas que se fundiriam com a membrana neuronal pré-sináptica, ou em termos de um aumento ou diminuição no número de canais iônicos na membrana pós-sináptica.

O nome que se dá a uma modificação na eficácia de uma sinapse é **plasticidade sináptica**.

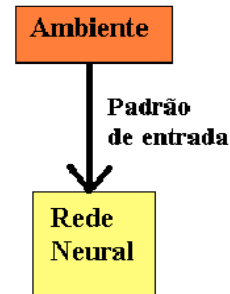
No caso das redes neurais artificiais, as eficácias sinápticas são representadas pelos pesos w_{ij} conectando os neurônios entre si.

A maneira como se faz para que uma rede neural artificial execute uma dada tarefa com um desempenho satisfatório é, por analogia com o mecanismo biológico, através de um período de aprendizado a partir de exemplos. Durante esse período de aprendizado, os pesos sinápticos w_{ij} são modificados de acordo com alguma regra (um algoritmo) visando o aprimoramento do desempenho da rede neural na execução da tarefa.

Há basicamente dois tipos de aprendizado em redes neurais: supervisionado e não-supervisionado. A figura a seguir ilustra estes dois tipos de aprendizado.



Aprendizado supervisionado
(há um professor que conhece
exemplos de entrada-saída)



Aprendizado não-supervisionado
ou **auto-organizado**
(sem professor)

No aprendizado supervisionado, há um período de treinamento ou aprendizado em que padrões retirados de uma “base de treinamento” são apresentados à rede neural, um por vez. Existe também um “professor”, que recebe os mesmos padrões que a rede neural. O professor conhece as respostas corretas para cada padrão de entrada, chamadas de saídas desejadas da rede neural. Para cada padrão apresentado em sua entrada, a rede neural fornece uma resposta na saída. Essa resposta é comparada com a resposta desejada fornecida pelo professor. Se as duas forem iguais, nada é feito e passa-se para o próximo padrão. Porém, se houver diferença entre a resposta desejada e a saída da rede, um sinal de erro é enviado à rede neural e os seus pesos sinápticos são modificados em função desse sinal de erro.

Este processo é repetido para cada um dos padrões da base de treinamento. Quando todos os padrões da base de treinamento tiverem sido apresentados à rede neural, dizemos que uma “época” de treinamento foi concluída.

Em geral, várias épocas de treinamento são necessárias para que a rede neural “aprenda” a fornecer as respostas corretas para os padrões de treinamento.

Com o processo de aprendizado supervisionado, podemos dizer que a rede neural cria um “modelo” para reproduzir as regras de associação entre padrões de entrada e desejados ensinadas pelo professor. A rede cria um “modelo do professor”. Como esse modelo é aprendido a partir dos casos existentes na base de treinamento, ele será tanto melhor quanto mais representativos do ambiente forem os casos escolhidos.

Sob este ponto de vista, uma rede neural pode ser vista como uma ferramenta estatística para o aprendizado das regras de associação entre os padrões de entrada do ambiente e as respostas desejadas propostas pelo professor.

Por outro lado, no aprendizado não-supervisionado a rede neural não tem um professor para lhe forçar a dar as respostas que ele considera desejáveis. Na ausência de um modelo a seguir, a rede vai alterando os seus pesos sinápticos segundo regras pré-estabelecidas que se baseiam apenas nos padrões de entrada e nos valores dos seus pesos. Dizemos que a rede neural se auto-organiza. Um exemplo de regra de aprendizado não-supervisionado é a regra de Hebb, vista na aula 2.

No caso não-supervisionado, a auto-organização da rede neural pode levá-la ou não a fornecer respostas adequadas aos padrões de entrada vindos do ambiente. Como uma rede neural é, em geral, vista como uma ferramenta para gerar respostas úteis para o ser humano, a adequação das respostas da rede será medida pela utilidade delas em alguma tarefa específica.

Pode-se também estar interessado em estudar as respostas de uma rede neural treinada de uma maneira não-supervisionada para ver se elas são similares às observadas em experimentos de aprendizado com animais e humanos. Neste caso, não se estará dando tanta importância à utilidade prática das respostas, mas sim à capacidade das regras de auto-organização da rede em capturar elementos das regras de aprendizado existentes nos seres vivos (por exemplo, pré-definidas geneticamente).