Redes Neuronais Artificiais/Tópicos Avan. RNA
Introdução às Redes Neuronais

1 Função Composta

Sejam $f : D_f \subseteq \mathbb{R} \to \mathbb{R}$ e $g : D_g \subseteq \mathbb{R} \to \mathbb{R}$ duas funções. Definimos a **composição** de f com g por

$$(f \circ g) = f\Big(g(x)\Big).$$

Exemplo 1.1. *A função* $h(x) = \sqrt{x^2 - 4}$ *resulta da composição de* $f(x) = \sqrt{x}$ *com* $g(x) = x^2 - 4$.

2 Derivada de uma Função Composta

Se g é diferenciável em a e f é diferenciável em $g(a)$, então a função $(f \circ g)$ é uma função diferenciável em a e tem-se

$$(f \circ g)'(a) = f'(g(a)) \cdot g'(a).$$

Exemplo 2.1. *A função* $h(x) = (5 - 6x)^5$ *resulta da composição de* $f(x) = x^5$ *com* $g(x) = 5 - 6x$.

Como

$$f'(x) = 5x^4 \quad e \quad g'(x) = -6,$$

resulta que

$$h'(x) = 5(5 - 6x)^4 \cdot -6,$$

e simplificando

$$h'(x) = -30(5 - 6x)^4.$$

3 Optimização dos Erros

3.1 Pesos Entre a Última Camada Oculta e a Camada de Saída

Se denotarmos o valor esperado por y e o valor obtido pela rede por $\hat{y}$, então o erro vem:

$$E = \sum_n (y_n - \hat{y}_n)^2,$$

onde n denota o número de neurónios de saída.

Por vezes, e como o objectivo de facilitar os cálculos intermédios, o erro é definido por:

$$E = \frac{1}{2} \sum_n (y_n - \hat{y}_n)^2.$$

Para minimizarmos o erro, temos de perceber o quão sensível é o erro às mudanças nos pesos das ligações. Matematicamente, essa relação é dada por:

$$\frac{\partial E}{\partial w_{jk}},$$

onde w_{jk} denota o pesos que estão associados ao neurónio k.

Vejamos, agora, como podemos actualizar os pesos de uma rede neuronal.

Comecemos por decompor o erro:

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \sum_n (y_n - \hat{y}_n)^2.$$

Visto que a saída do neurónio de saída n (*i.e.* $\hat{y}_n$) só depende dos pesos das ligações até ele, podemos simplificar a equação anterior, retirando o somatório. Ou seja,

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k)^2.$$

Esta é uma função é uma função composta. Neste caso, a função $h(x) = (y_k - \hat{y}_k)^2$ resulta da composição de $f(x) = x^2$ com $g(x) = y_k - \hat{y}_k$.

Continuando,

$$\frac{\partial E}{\partial w_{jk}} = 2(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k).$$

Vamos por partes. Assumindo que a função de activação é a função sigmóide,

$$\frac{\partial \hat{y}_k}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} \text{sigmóide} \left(\sum_j w_{jk} \cdot o_j \right),$$

onde o_j é o valor da saída do neurónio da camada oculta anterior.

Esta é uma função é uma função composta. Neste caso, a função $h(x) = \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right)$ resulta da composição de $f(x) = \text{sigmóide}(x)$ com $g(x) = \sum_j w_{jk} \cdot o_j$.

A derivada da função sigmóide é pode ser calculada da seguinte forma:

$$\frac{\partial}{\partial x} \text{sigmóide}(x) = \text{sigmóide}(x) \cdot (1 - \text{sigmóide}(x)).$$

Continuado,

$$\frac{\partial}{\partial w_{jk}} \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) = \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) \cdot \left(1 - \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot \frac{\partial}{\partial w_{jk}} \sum_j w_{jk} \cdot o_j.$$

É importante notar que em $\frac{\partial}{\partial w_{jk}} \sum_j w_{jk} \cdot o_j$ apenas um termo depende de ∂w_{jk}, *i.e.*,, quando o j do somatório for igual ao j da derivada parcial.
Assim, essa derivada parcial pode ser reescrita da seguinte forma:

$$\frac{\partial}{\partial w_{jk}} w_{jk} \cdot o_j = o_j.$$

Finalmente,

$$\frac{\partial}{\partial w_{jk}} \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) = \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) \cdot \left(1 - \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot o_j.$$

Agora que já sabemos $\frac{\partial \hat{y}_k}{\partial w_{jk}}$, resta calcular

$$\frac{\partial y_k}{\partial w_{jk}} = 1.$$

Ou seja,

$$\frac{\partial}{\partial w_{jk}} (y_k - \hat{y}_k) = -\text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) \cdot \left(1 - \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot o_j.$$

O que nos leva a concluir que

$$\frac{\partial E}{\partial w_{jk}} = -2(y_k - \hat{y}_k)\ \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) \cdot \left(1 - \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot o_j.$$

Visto que só nos interessa a direcção, podemos simplificar a equação anterior retirando a constante 2, de tal forma que:

$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k)\ \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right) \cdot \left(1 - \text{sigmóide}\left(\sum_j w_{jk} \cdot o_j\right)\right) \cdot o_j.$$

Ou seja, o gradiente do erro para um determinado peso entre a última camada oculta e a camada de saída é dado por:

$$\frac{\partial E}{\partial w_{jk}} = -E_k \times O_k \times (1 - O_k) \times o_j,$$

onde $E_k = (y_k - \hat{y}_k)^2$, $O_k = \text{sigmóide}\,(\mathbf{w} \cdot \mathbf{i})$ e o_j é o valor de saída do neurónio anterior.

3.2 Restantes Pesos

A expressão é semelhante, e é dada por

$$\frac{\partial E}{\partial w_{jk}} = -e_k \times O_k \times (1 - O_k) \times o_j,$$

onde e_k é o erro que vem da retro-propagação.

Ou seja, a primeira parte—que antes era a diferença entre o valor real e o valor obtido—agora é o erro retro-propagado a partir dos neurónios ocultos.

As partes referentes às funções de activação sigmóides são as mesmas, mas referem-se, agora, às camadas anteriores, ou seja, a função de activação é aplicada à a soma de todas as entradas multiplicadas pelos das ligações ao neurónio k.

A última parte agora é a saída da camada anterior, que se for a última, corresponde às entradas da rede.

Contudo, no caso genérico, têm-se

$$\frac{\partial E}{\partial w_{jk}} = -e_k \times \frac{\partial}{\partial w_{jk}} \phi \left(\sum_j w_{jk} \cdot o_j \right) \times o_j,$$

onde ϕ é uma qualquer função de activação diferenciável.

3.3 Concluindo

O peso é, finalmente, actualizado da seguinte forma:

$$w'_{jk} = w_{jk} - \alpha \times \frac{\partial E}{\partial w_{jk}},$$

onde α é o *learning rate.*