

BACKPROPAGATION

El método de mínimos cuadrados ordinarios no puede ser aplicado directamente a un perceptrón multicapa. Para entrenar este modelo se emplea la técnica conocida como *retropropagación* (*backpropagation*), que se vale del método de descenso por gradiente y se aplica secuencialmente a las capas de la red, desde la salida hasta la entrada, gracias a la técnica matemática de la *regla de la cadena*, que permite calcular la derivada de una composición de funciones. Si $y = g(x)$ y $z = f(y) = f(g(x))$:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}.$$

El método de retropropagación comienza calculando el error de las predicciones en la capa de salida, E , y ajustando los pesos de las unidades de dicha capa por descenso por gradiente, para pasar a la capa inmediatamente anterior, realizar en sus pesos los ajustes correspondientes, pasar a la capa anterior, etc.

Si i , j y k son los índices de unidades en tres capas consecutivas, w_{ji} es el peso que conecta la unidad i y la unidad j , $s_j = \sum_i w_{ji} z_i$ con z_i la salida de la unidad i , y F es la función de activación de las capas ocultas, entonces la dirección en la que se actualiza cada peso viene dada por:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial s_j} \frac{\partial s_j}{\partial w_{ji}} = \delta_j \frac{\partial s_j}{\partial w_{ji}} = \delta_j z_i,$$

donde

$$\begin{aligned}\delta_j &= \frac{\partial E}{\partial s_j} \\ &= \sum_k \frac{\partial E}{\partial s_k} \frac{\partial s_k}{\partial s_j} \\ &= \sum_k \delta_k \frac{\partial s_k}{\partial z_j} \frac{\partial z_j}{\partial s_j} \\ &= \sum_k \delta_k w_{kj} \frac{\partial F}{\partial s_j}\end{aligned}$$

es lo que se conoce como *error generalizado*. Este error generalizado se puede calcular directamente para las unidades de la capa de salida, y para el resto, como puede verse, se calcula a partir del error generalizado de las unidades de la capa siguiente. Al igual que en el caso del modelo lineal, las actualizaciones de los pesos calculadas con este método se ponderan con el parámetro de la tasa de aprendizaje.