Descenso de gradiente

Ya abordamos el descenso de gradiente en el capítulo Principales algoritmos de Machine Learning dedicado a las estadísticas, pero estamos seguros de que es necesario volver a mencionarlo ahora porque se trata de un concepto importante para la comprensión del funcionamiento de las redes neuronales.

El objetivo del descenso de gradiente es minimizar la función de error ajustando poco a poco los parámetros de aprendizaje representados por los distintos pesos.

Retomamos la imagen del descenso de la montaña. Usted se sitúa en el punto más alto de la montaña y quiere alcanzar la llanura que se encuentra más abajo. Sin embargo, es una noche oscura y no puede ver a dónde va.

Entonces avanzará poco a poco con pasos cortos hasta alcanzar la parte inferior del valle.

El descenso del gradiente corresponde a esta metáfora y se lleva a cabo mediante el ajuste de los distintos pesos de la red neuronal hasta obtener una convergencia, es decir, un mínimo de errores. Este ajuste se hace mediante pasos cortos, usando un hiperparámetro denominado tasa de aprendizaje (*learning rate*).

La diferencia entre parámetros y hiperparámetros es que los primeros se definen a nivel del modelo de aprendizaje (valor de los pesos) y los segundos se definen a nivel del algoritmo (por ejemplo: tasa de aprendizaje o cantidad de capas de neuronas).

No obstante, hay que prestar atención para no elegir una tasa de aprendizaje demasiado pequeña porque entonces tendrá que esperar mucho tiempo antes de que se produzca la convergencia. Si eligiera una tasa de aprendizaje demasiado grande, la consecuencia podría ser que se sobrepasara el error mínimo y se produjera un aumento de la convergencia. Este aumento impediría encontrar la solución correcta.

Imagen que contiene Diagrama

Descripción generada automáticamente

*Descenso de gradiente y convergencia*

Diagrama

Descripción generada automáticamente

*Descenso de gradiente. Mínimo sobrepasado*

A veces, para llegar al valle que está más abajo, hay que superar una montaña que se encuentra en nuestro camino porque el punto más bajo que queremos alcanzar se halla detrás de esta montaña.

Si nos detenemos antes de ascender por la montaña pequeña, desde luego que hemos alcanzado un punto más bajo, al que llamaremos mínimo local. Pero este punto no corresponde al que queremos alcanzar, también conocido como mínimo global.

Diagrama

Descripción generada automáticamente

*Descenso de gradiente. Mínimo local y mínimo global*

Como habrá comprendido, el descenso del gradiente permite mejorar poco a poco los parámetros de la red neuronal hasta obtener una convergencia, es decir, una minimización de la función del error.

Sin embargo, la función de error que se ha de utilizar no es la relativa a una observación (error lineal), sino al conjunto de las observaciones (error global). Por lo tanto, el objetivo es encontrar parámetros que puedan minimizar el coste del error de la función MSE que hemos visto anteriormente.

¿Cómo podemos asegurarnos de que no estamos en mínimo local? La función MSE es una función convexa. Es decir, si tomamos su representación gráfica, elegimos dos puntos de la curva y trazamos un segmento entre estos dos puntos, el segmento nunca corta a la curva. Por lo tanto, no hay mínimo local, solo hay un mínimo global. En otras palabras, si buscamos minimizar esta función de error, estamos seguros de encontrar un mínimo global siempre que no hayamos elegido una tasa de aprendizaje demasiado elevada Y que esperemos bastante tiempo, es decir, realicemos una cierta cantidad de épocas (epoch).

Una etapa se define como el momento en que el conjunto de las observaciones de un conjunto de datos se envía a la red neuronal y experimenta las fases de propagación y de retropropagación. Las épocas sirven para determinar si estamos en el buen camino de minimización de la función de error global porque utilizamos esta función a la salida de una época. Recordemos que la función de error global se calcula sobre el conjunto de los errores de predicciones realizadas en el conjunto de las observaciones.

Esto significa que para minimizar la función de error necesitaremos varias épocas y que, por lo tanto, el conjunto de las observaciones deberá pasar varias veces por la red neuronal.