Laboratorio 4: SGD & Redes Neuronales

Inteligencia Artificial - CC3085

José Antonio Mérida Castejón

11 de febrero de 2026

Task 1 - Teoría

Responda con criterio y anàlisis de ingeniero. No se esperan definiciones de libro, si no que analice las consecuencias de las decisiones de diseño.

El Dilema del "Step Size"

Durante la clase hablamos sobre el tamaño del paso. Suponga que está entrenando una red neuronal profunda con SGD (Estocástico) y nota que la función de pérdida (Loss) disminuye rápidamente al inicio, pero luego empieza a oscilar violentamente sin llegar nunca a un valor mínimo estable.

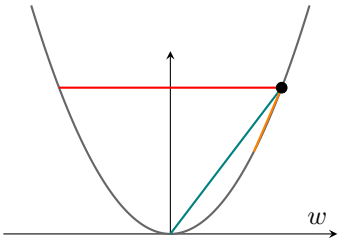
1. *Explique que está sucediendo geométricamente en la superficie de error*

   En términos simples, cada uno de nuestros 'saltos' nos está mandando hacia 'el otro lado' del mínimo que queremos encontrar.

   En términos más matemáticos, la gradiente nos indica la dirección dónde la función es más creciente (y su negativo, la dirección de descenso más pronunciado). Es decir, la información con la que contamos es únicamente una dirección, sin saber la distancia exacta a la que se encuentra el mínimo que buscamos. Esta distancia que se 'recorre' en cada iteración se determina únicamente a través del Learning Rate (tamaño del paso).

   Teniendo un Learning Rate demasiado alto, el tamaño del paso calculado por la red supera la distancia real al mínimo. Esto provoca que nuestros pasos jamás sean tan precisos como para ubicar un mínimo, resultando en un rebote constante entre las "paredes"de la función de error.

   A continuación, podemos visualizar este comportamiento en una superficie de error simple en 2 dimensiones ($f(x) = x^2$):

   Gradiente en Cero

Durante la clase hablamos sobre como evitar las gradientes cero. En base a eso responda

a) *Si usted diseña una red profunda utilizando únicamente la función Sigmoide en todas las capas ocultas, es muy probable que sufra el problema del Vanishing Gradient (Desvanecimiento del Gradiente). Explique matemáticamente por qué ocurre esto al hacer Backpropagation (piense en la derivada máxima de la sigmoide).*

b) *¿Por qué la función ReLU mitiga este problema en la parte positiva del eje?*

Capacidad del Modelo

Si comparamos una Red Neuronal con 1 capa oculta de 100 neuronas contra una Red Neuronal con 10 capas ocultas de 10 neuronas cada una (ambas tienen un número similar de parámetros).

a) *¿Cuál de las dos tiene mayor capacidad para modelar funciones no lineales complejas y composicionales? ¿Por qué? Base se respuesta con lo visto en clase ("¿Por qué redes profundas?").*