



Data Science Academy

www.datascienceacademy.com.br

Matemática Para Machine Learning

No Problema de Regressão em Machine Learning, Por Que o Mínimo Local é Calculado Para Uma Função Derivada em vez da Função Real?



No Problema de Regressão em Machine Learning, Por Que o Mínimo Local é Calculado Para Uma Função Derivada em vez da Função Real?

A descida de gradiente é um algoritmo de otimização iterativa de primeira ordem para encontrar o mínimo de uma função. Para encontrar um mínimo local de uma função usando gradiente descendente, tomamos medidas proporcionais ao negativo do gradiente (ou gradiente aproximado) da função no ponto atual. Se, em vez disso, adotarmos medidas proporcionais ao positivo do gradiente, uma delas se aproxima de um máximo local dessa função; o procedimento é então conhecido como subida do gradiente.

A descida de gradiente é também conhecida como descida mais acentuada. No entanto, gradiente descendente não deve ser confundido com o método de descida mais íngreme para aproximar integrais.

O algoritmo de descida de gradiente é aplicado para encontrar um mínimo local de uma função. Por exemplo:

Função A (Função Real) $\rightarrow f(x) = x^4 - 3x^3 + 2$

Função B (Função Derivada) $\rightarrow f'(x) = 4x^3 - 9x^2$

Aqui para encontrar o mínimo local usando o algoritmo de gradiente descendente para a função (A) usamos a função derivada de (A) que é a função (B).

A razão é que, como a função é côncava (ou convexa, se você estiver fazendo a maximização - esses problemas são equivalentes), você sabe que existe um mínimo (e máximo). Isso significa que existe um único ponto em que o gradiente é igual a zero. Existem técnicas que usam a função em si, mas se você puder calcular o gradiente, poderá convergir muito mais rápido, pois pode pensar no gradiente que fornece informações sobre o quanto você está longe do ideal.

Assim como o Gradient Descent, existe um método de otimização conhecido como método de Newton, que requer o cálculo da segunda derivada (o Hessian na otimização multivariada). Isso converge ainda mais rápido, mas requer que você seja capaz de inverter o Hessian, o que não é viável se você tiver muitos parâmetros.

Portanto, existem métodos para contornar isso, que calculam uma aproximação de memória limitada do Hessian. Esses métodos convergem mais rápido ainda porque estão usando informações sobre a curvatura do gradiente: é uma troca simples, em que quanto mais você souber sobre a função que está tentando otimizar, mais rápido poderá encontrar a solução. Estudaremos isso em detalhes mais a frente aqui mesmo no curso!