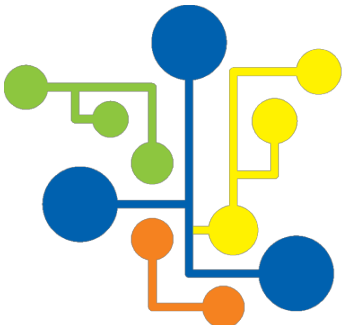


Data Science Academy

www.datascienceacademy.com.br

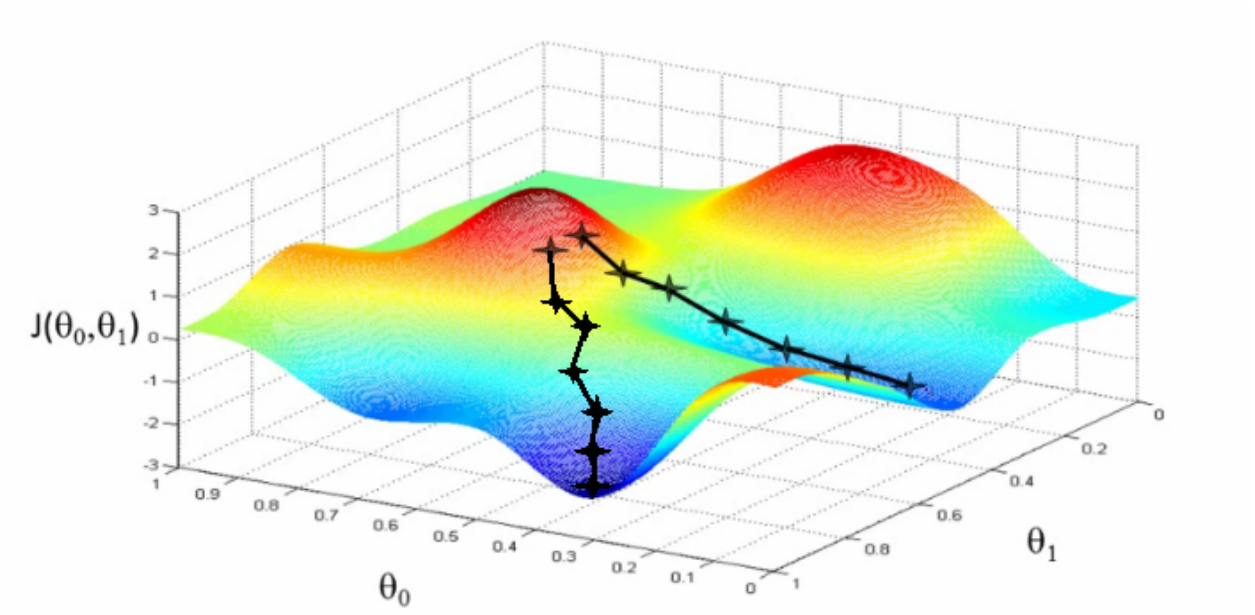
Matemática Para Machine Learning

Descrevendo o Processo de Descida do Gradiente

A otimização é o ingrediente essencial na receita dos algoritmos de aprendizado de máquina. Ele começa com a definição de algum tipo de função de perda / função de custo e termina com a minimização do uso de uma ou outra rotina de otimização. A escolha do algoritmo de otimização pode fazer a diferença entre obter uma boa precisão em horas ou dias. As aplicações de otimização são ilimitadas e são amplamente pesquisadas por profissionais em todo mundo.

Descida Estocástica do Gradiente

A Descida Estocástica do Gradiente (SGD – Stochastic Gradient Descent) é o algoritmo de otimização mais simples usado para encontrar parâmetros que minimizam a função de custo dada. Aparentemente, para a descida de gradiente convergir para o mínimo ideal, a função de custo deve ser convexa. Para o propósito de demonstração, imagine a representação gráfica da descida estocástica do gradiente na figura abaixo.

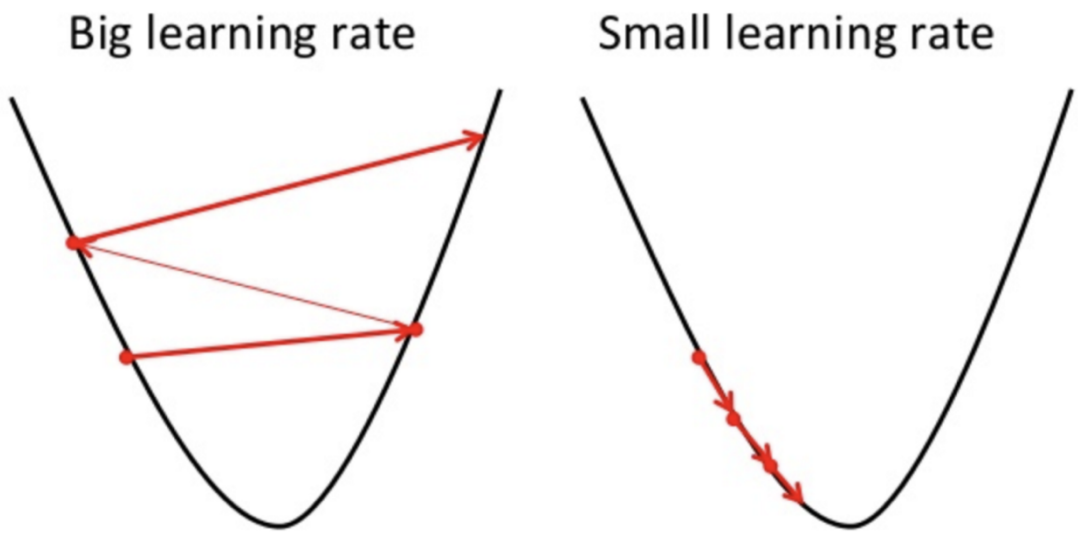


Começamos com a definição de alguns valores iniciais aleatórios para os parâmetros. O objetivo do algoritmo de otimização é encontrar valores de parâmetros que correspondam ao valor mínimo da função de custo. Especificamente, a descida de gradiente começa com o cálculo de gradientes (derivadas) para cada um dos parâmetros da função de custo. Esses gradientes nos dão um ajuste numérico que precisamos fazer em cada parâmetro para minimizar a função de custo. Esse processo continua até atingirmos o mínimo local / global. Matematicamente, teríamos isso em Python:

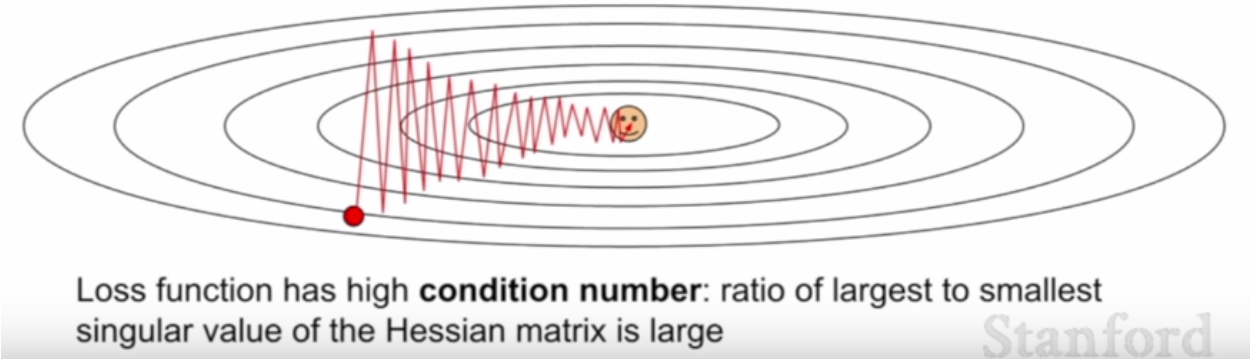
```
for i in range(iterations_count):
    param_gradients = evaluate_gradients(loss_function,
                                         data,
                                         params)

    params -= learning_rate * param_gradients
```

A learning rate (taxa de Aprendizado) define quantos parâmetros devem ser alterados em cada iteração. Em outras palavras, ele controla o quão rápido ou devagar devemos convergir para o mínimo. Por um lado, uma pequena taxa de aprendizado pode levar as iterações a convergir. Uma grande taxa de aprendizado pode exceder o mínimo, como você pode ver na figura abaixo.



Embora fácil de aplicar na prática, esse processo tem algumas desvantagens quando se trata de redes neurais profundas, pois essas redes têm um grande número de parâmetros para se encaixar. Para ilustrar problemas com gradiente descendente, vamos supor que temos uma função de custo com apenas dois parâmetros. Suponha que a função de custo seja muito sensível a alterações em um dos parâmetros, por exemplo, na direção vertical e menos a outro parâmetro, ou seja, a direção horizontal (isso significa que a função de custo tem um número de condição alto).



Se executarmos a descida estocástica do gradiente nessa função, obteremos um tipo de comportamento em ziguezague. Em essência, a SGD está fazendo um progresso lento em direção a uma direção menos sensível e mais voltada para uma alta sensibilidade e, portanto, não alinha na direção do mínimo. Na prática, a rede neural profunda poderia ter milhões de parâmetros e, portanto, milhões de direções para acomodar ajustes de gradiente e, assim, agravar o problema. Treinar um modelo e Deep Learning não é tarefa simples.

Outro problema com o SGD é o problema do mínimo local ou dos *pontos de sela*. Os pontos de sela são pontos em que o gradiente é zero em todas as direções. Consequentemente, o nosso SGD ficará preso apenas lá. Por outro lado, os mínimos locais são pontos que são mínimos, mas não o global mínimo. Como o gradiente será zero no mínimo local, nossa descida do gradiente o reportará como um valor mínimo quando o mínimo global estiver em outro lugar.

Para corrigir os problemas com o SGD padrão, vários algoritmos avançados de otimização foram desenvolvidos nos últimos anos. Vamos estudar os principais agora.

Referências:

http://cs229.stanford.edu/notes/cs229-notes1.pdf

https://stanford.edu/~rezab/classes/cme323/S15/notes/lec11.pdf