Stochastic Gradient Descent

August 29, 2019

1 Stochastic Gradient Descent

Stochastic gradient descent (often abbreviated SGD) is an optimization algorithm developed from gradient descent. In this section, gradient descent is introduced as the first part of the explanation of SGD.

1.1 Gradient Descent

For convex optimization problems, gradient descent is a first-order optimization algorithm to find the local minimum of a function.

To solve the minimization problem

$$\min_{\boldsymbol{x}} L(\boldsymbol{x})$$

where $L : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is convex, differentiable and its gradient is Lipschitz continuous with constant $L > 0$.

Geometrically, the gradient $\nabla L(\boldsymbol{x}_0)$ points to the direction of the steepest ascent on $L(\cdot)$ from the point $\boldsymbol{x}_0$. By taking a small step in the direction of the negative gradient, the function value is decreased in the direction of the steepest descent. That is,

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \alpha \nabla L(\boldsymbol{x}_0)$$

for a small enough stepsize $\alpha \in \mathbb{R}_+$, then $L(\boldsymbol{x}_1) \leq L(\boldsymbol{x}_0)$. That means, compared with $L(\boldsymbol{x}_0)$, $L(\boldsymbol{x}_1)$ is closer to the local minimum.

With this observation comes the idea of gradient descent: an iterative "tour" on $L(\cdot)$ from a point towards the local minimum by following small steps of negative gradient. Let $\boldsymbol{x}_0$ be the guess of a starting point, then if

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla L(\boldsymbol{x}_k), k \geq 0$$

Then we have $L(\boldsymbol{x}_0) \geq L(\boldsymbol{x}_1) \geq L(\boldsymbol{x}_2) \geq \cdots$ with suitable α_k. The convergence of the sequence $(\boldsymbol{x}_n)$ to the local minimum is guaranteed[reference].

1.2 Stochastic Gradient Descent

SGD can be considered as a stohcastic appoximation of gradient descent optimization, when the objective function $L(\cdot)$ can be written as a sum of differentiable functions. Consider the objective

function is in the form:

$$L(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} L_k(\boldsymbol{x})$$

where the summand function L_k is usually the loss function of the kth observation among K data points.

Then by following the idea of gradient descent, the $\boldsymbol{x}$ is updated according to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla L(\boldsymbol{x}_k) = \boldsymbol{x}_k - \alpha_k \frac{1}{K}\sum_{k=1}^{K} \nabla L_k(\boldsymbol{x}_k)$$

where each α_k is a suitable stepsize. The calculation of $\sum_{k=1}^{K} \nabla L_k(\boldsymbol{x}_k)$ can be expensive, especially when the amount of summand functions is huge, or when the individual gradients are hard to compute.

To reduce the consumption of calculation, an estimation of the true gradient of $L(\boldsymbol{x})$ is taken: the true gradient $\frac{1}{K}\sum_{k=1}^{K} \nabla L_k(\boldsymbol{x}_k)$ is replaced by the gradient of a single observation $\nabla L_k(\boldsymbol{x}_k)$. So the update of the parameter $\boldsymbol{x}$ becomes

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla L_k(\boldsymbol{x}_k)$$

where α_k is a suitable stepsize.

The convergence of SGD has been proved as well[reference].

(and should I explain more about why stochastic gradient descent works?)