线性模型和梯度下降

这是神经网络的第一课，我们会学习一个非常简单的模型，线性回归，同时也会学习一个优化算法-梯度下降法，对这个模型进行优化。线性回归是监督学习里面一个非常简单的模型，同时梯度下降也是深度学习中应用最广的优化算法，我们将从这里开始我们的深度学习之旅。

一元线性回归

一元线性模型非常简单，假设我们有变量 x_i 和目标 y_i，每个 i 对应于一个数据点，希望建立一个模型

$$\hat{y}_i = wx_i + b \tag{1}$$

$\hat{y}_i$ 是我们预测的结果，希望通过 $\hat{y}_i$ 来拟合目标 y_i，通俗来讲就是找到这个函数拟合 y_i 使得误差最小，即最小化

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

那么如何最小化这个误差呢?

这里需要用到**梯度下降**，这是我们接触到的第一个优化算法，非常简单，但是却非常强大，在深度学习中被大量使用，所以让我们从简单的例子出发了解梯度下降法的原理

梯度下降法

在梯度下降法中，我们首先要明确梯度的概念，随后我们再了解如何使用梯度进行下降。

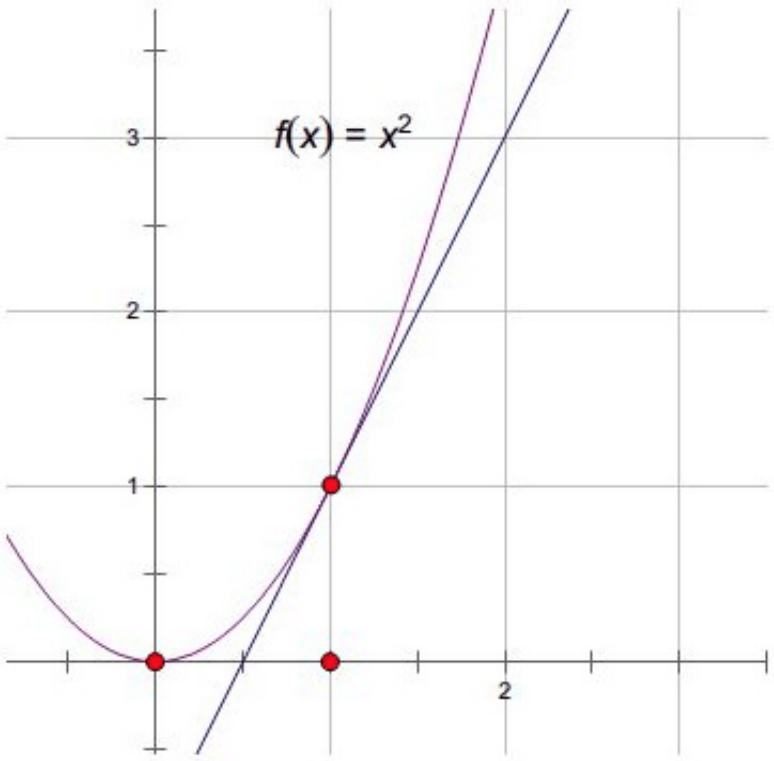
梯度

梯度在数学上就是导数，如果是一个多元函数，那么梯度就是偏导数。比如一个函数f(x, y)，那么 f 的梯度就是

$$(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}) \tag{3}$$

可以称为 grad f(x, y) 或者 $\nabla f(x, y)$。具体某一点 (x_0, y_0) 的梯度就是 $\nabla f(x_0, y_0)$。

下面这个图片是 $f(x) = x^2$ 这个函数在 x=1 处的梯度



梯度有什么意义呢？从几何意义来讲，一个点的梯度值是这个函数变化最快的地方，具体来说，对于函数 f(x, y)，在点 (x_0, y_0) 处，沿着梯度 $\nabla f(x_0, y_0)$ 的方向，函数增加最快，也就是说沿着梯度的方向，我们能够更快地找到函数的极大值点，或者反过来沿着梯度的反方向，我们能够更快地找到函数的最小值点。

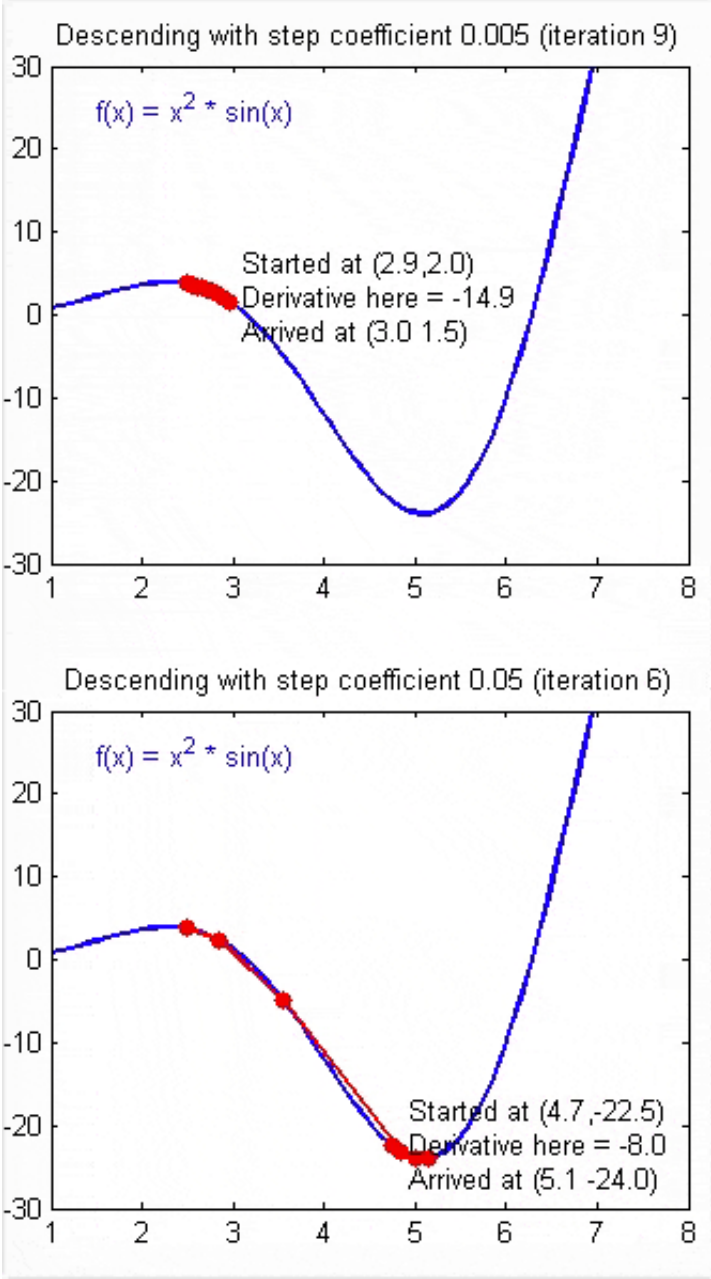
梯度下降法

有了对梯度的理解，我们就能了解梯度下降发的原理了。上面我们需要最小化这个误差，也就是需要找到这个误差的最小值点，那么沿着梯度的反方向我们就能够找到这个最小值点。

我们可以来看一个直观的解释。比如我们在一座大山上的某处位置，由于我们不知道怎么下山，于是决定走一步算一步，也就是在每走到一个位置的时候，求解当前位置的梯度，沿着梯度的负方向，也就是当前最陡峭的位置向下走一步，然后继续求解当前位置梯度，向这一步所在位置沿着最陡峭最易下山的位置走一步。这样一步步的走下去，一直走到觉得我们已经到了山脚。当然这样走下去，有可能我们不能走到山脚，而是到了某一个局部的山峰低处。

类比我们的问题，就是沿着梯度的反方向，我们不断改变 w 和 b 的值，最终找到一组最好的 w 和 b 使得误差最小。

在更新的时候，我们需要决定每次更新的幅度，比如在下山的例子中，我们需要每次往下走的那一步的长度，这个长度称为学习率，用 η 表示，这个学习率非常重要，不同的学习率都会导致不同的结果，学习率太小会导致下降非常缓慢，学习率太大又会导致跳动非常明显，可以看看下面的例子



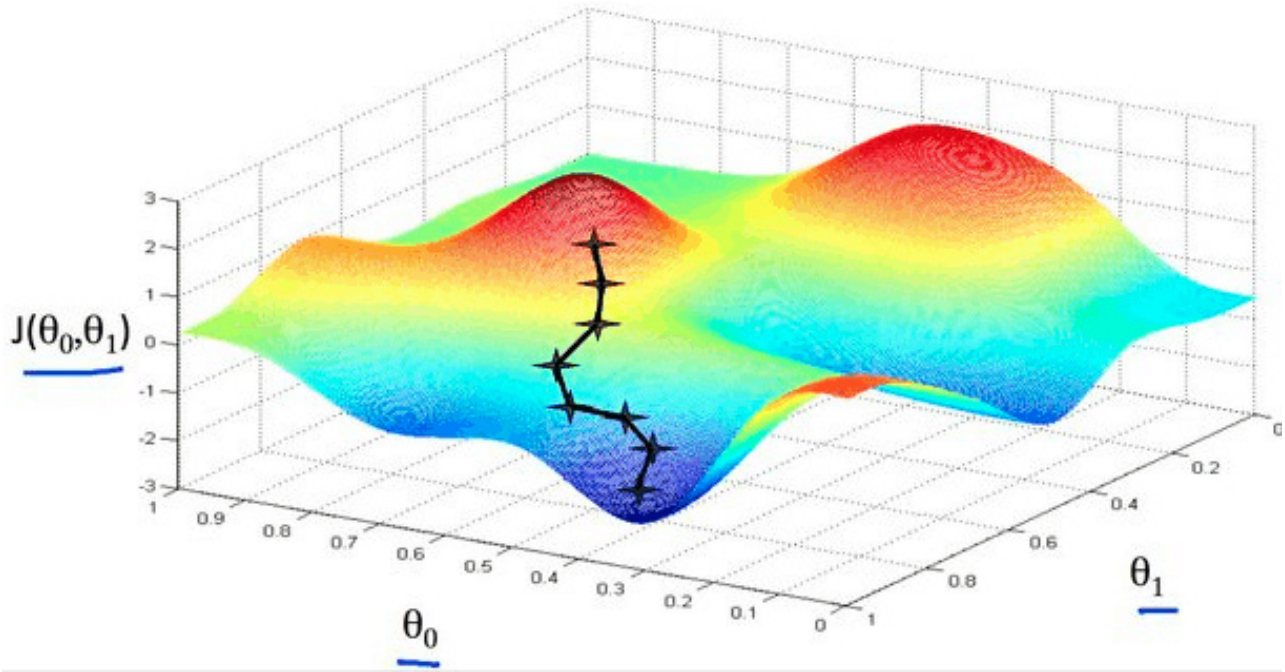
可以看到上面的学习率较为合适，而下面的学习率太大，就会导致不断跳动

最后我们的更新公式就是

$$w := w - \eta \frac{\partial f(w,\ b)}{\partial w} \tag{4}$$
$$b := b - \eta \frac{\partial f(w,\ b)}{\partial b}$$

通过不断地迭代更新，最终我们能够找到一组最优的 w 和 b，这就是梯度下降法的原理。

最后可以通过这张图形象地说明一下这个方法



多项式回归模型

下面我们更进一步，讲一讲多项式回归。什么是多项式回归呢？非常简单，根据上面的线性回归模型

$$\hat{y} = wx + b \tag{5}$$

这里是关于 x 的一个一次多项式，这个模型比较简单，没有办法拟合比较复杂的模型，所以我们可以使用更高次的模型，比如

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots \tag{6}$$

这样就能够拟合更加复杂的模型，这就是多项式模型，这里使用了 x 的更高次，同理还有多元回归模型，形式也是一样的，只是出了使用 x，还是更多的变量，比如 y、z 等等，同时他们的 loss 函数和简单的线性回归模型是一致的。