学习率（learning rate）：参数调优过程中选择的步长。

梯度下降法：<https://www.jianshu.com/p/c7e642877b0e>

首先了解梯度的概念：

1. 在单变量函数中，梯度是函数的微分，代表函数在某个给定点的切线的斜率。
2. 在多变量函数中，梯度是一个向量，向量的方向就指出了函数在给定点的上升（反方向为下降）最快的方向。

线性回归中的梯度下降法：

1. 批量梯度下降（Batch Gradient Descent, BGD）

首先定义代价函数(Cost Function)，选取均方误差函数：



其中：为数据集个数

为真实值。

为预测值，代表参数总共有个，其中

代价函数的梯度为：对各个参数分别求导





参数更新：



伪代码：

for j=0,..n



优点：

从参数更新公式中可以看出，每一个参数的更新都是对所有样本进行计算，精度高，一定能得到全局最优。

缺点：

如果样本数目非常大，每更新一个参数都会很耗时，训练过程就会很慢。

1. 随机梯度下降（Stochastic Gradient Descent, SGD）

所有参数每次更新都使用一个样本来进行更新。





其中代表第个样本，代表样本的第个分量。

伪代码：

for i=1,…m

for j=0,…n



优点：

速度快，如果样本量很大，可能只使用一小部分样本就可以满足我们设置的条件，就不需要在进行迭代了。

缺点：

准确度下降，可能会收敛到局部最优，不能代表全部样本的趋势。

小批量梯度下降（Mini-Batch Gradient Descent, MBGD）：

批量梯度下降和随机梯度下降的折中。

所有参数每次更新都采用一定数目的样本，设为batch\_size

伪代码：假设batch\_size=10

优点

for i=1,11,21…

for j=0,…n

