# 线性回归（Linear Regression）

## 一、梯度下降法（Gradient Descend Algorithm）

一元情况下的Taylor展开式：

推广到多元情况下的Taylor展开式，即,时：

其中，，.

梯度下降法的目的就是求取的极值。我们知道当，且时(这里表示应为正定矩阵，正定矩阵的充要条件为它的所有特征值都大于零或者各阶主子式大于零),存在唯一极值点。

对于一阶Taylor展开式，当足够小的时候有以下的近似式：

从而

如果已经有一个，希望使小于，那么可以令。此时就有

在Taylor展开中，通常需要对有一定的限制，在梯度下降法中一般取**，**其中为学习率或步长。

如果考虑二阶Taylor展开，则：

注：有很多书和论文用， 来代替书写。

要求解的最小值，可以看作求使得

最小化。对求导可得

下面给出进行梯度下降法求解的一般步骤：

①给定初值，，学习率，给定阈值；

②求，令；

③求；

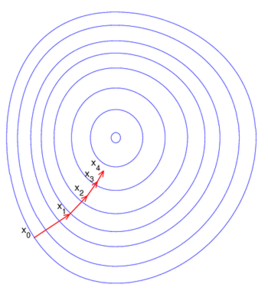
④若，转向⑤，否则，，转向②

⑤输出.

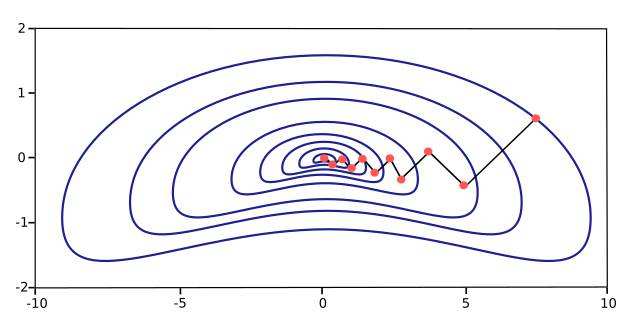
梯度下降法收敛速度慢与关于的Hessian矩阵的“条件”有关。梯度下降法第k次的迭代过程与之前迭代过程之间的关系可以通过下式来表示：

其中，是关于的Hessian矩阵（用表示）最小特征值与最大特征值的比值。从上式我们可以看出当较小或者的条件数较大的时候，梯度下降法的收敛速度就可能会较慢。经典的梯度下降法是线性收敛的，而牛顿法是二次收敛的。

另外，梯度下降法容易出现锯齿效应。考虑目标函数的等高线，如图1的情况，梯度下降法收敛速度也会很快。



图一 梯度下降法收敛较快的情形的示意图



图二 梯度下降法产生锯齿效应情形的示意图

可以证明梯度的方向与目标函数值构成的等高线垂直，如图2就容易产生锯齿效应收敛的很慢。

## 二、线性回归（Linear Regression）问题的求解

假设我们有一个数据集包含个数据对，其中是自变量，，是因变量，。可以认为是我们在生产过程中获得的一些传感器和仪表的数据构成的向量，比如的各个维度分别对应压力，温度，浓度等等，表示采样的时刻，是对应的某个指标，比如产品的质量指标等，它是连续的一个数值。我们希望有一个模型可以预测未来的，即（其中，）。简单地说，这个模型应该“非常好”地适合于已有的数据，我们令误差。线性回归的目标就是要求取来极小化的平方和

上式即为线性回归中的**目标函数.**

通常，将进行简化，最简单的，可以看作是一个超平面，即，其中。则有方程

令，，，,则有

的作用主要是产生一个偏置，如，而时，那么在轴上就会产生一个截距，这是线性回归问题优化时要考虑的。

求解的最小值有几种方法，下面用求导数的方式来求解，即用梯度下降法来求解。

因此线性回归问题可用梯度下降法来解。下面我们给出梯度下降法进行求解线性回归问题的一般步骤：

①给定初值, ，学习率，给定阈值；；

②求 ；

③若，转向④，否则，，转向②

④输出.

其中

考虑牛顿法，即

则，

可以发现即为最优解，这是因为目标函数为的二次函数，Taylor展开到二阶即为精确展开。一次牛顿法迭代就可得到最优解。

运用梯度下降法主要是方向和学习率的选取

如何改进和是提高收敛速度的关键。

一种常见的方法是针对不同维度进行自适应改进学习率，对不同维度采用不同的，比较常用的是RMSProp(root mean square propagation)。首先对每个维度进行移动平均 （平方是每个元素的平方）。一般选为0.9,0.99或0.999。参数的更新变为

（开根号是对每个元素进行，分子除以分母也是对每个元素进行）

这里可发现和经典的梯度下降法相比，多了一个类似权重的项用于不同的维度上。

另一类较常见的是Adam(adaptive moment estimation)

实际就是在梯度方向的选择上，也运用移动平均，既考虑当前的梯度方向，又考虑上一步的梯度方向，通过加权平均可减小锯齿效应。

注：其它改进的方法还有很多，主要也是针对方向和学习率。常用的这两个方法效果已经相当好。