# 线性回归模型的搭建（使用梯度下降法）

## 简介

（1）线性模型就是对输入特征加权求和，再加上一个偏置项（也称为截距项 ）的常数，以此进行预测。使用梯度下降法可以调整模型参数直至训练集上的成本函数调至最低。

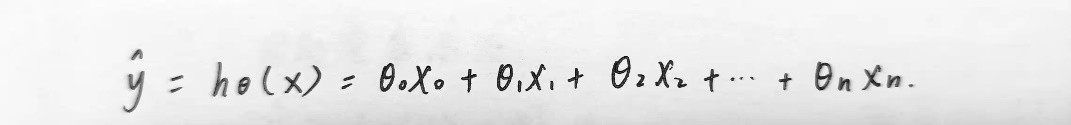
……

（2）多元线性回归：用多个特征量或变量来预测输出值

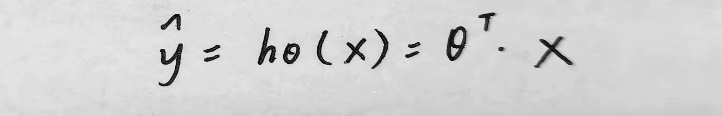
## 2.数学原理的理解

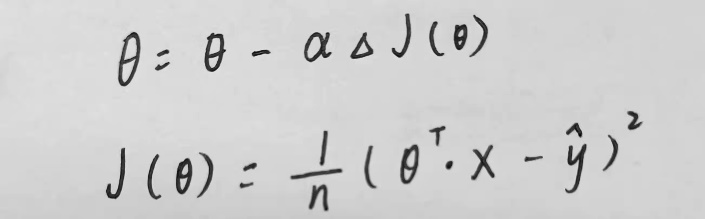
（1）模型的目的：给定一个数据集（称作训练集）D={(x1,y1),(x2,y2),(x3,y3)…,(xn,yn)},要学得一个线性模型，以尽可能准确地预测另一个数据集（称作测试集）D’={x1,x2,x3,…,xn}的输出。

（2）公式的理解：



用向量表示为：





·θT表示θ的转置向量（列向量）

·X为样本的特征向量

·y＾为预测值

·hθ为假设函数

·J(θ)为成本函数

·α为学习率

J(θ)为衡量预测结果与正确值的差距的函数，J(θ)越小，结果越准确，为了求出使J(θ)尽可能小的θ，要对最后两个式子进行迭代，更新θ的值，使J(θ)不断趋近于最小，使用θ预测的输出值逼近正确值。

## 3.模型具体实现过程

（1）对基础工具，基础库的认识：

第一周安装了Anaconda，开始使用Jupyter Notebook，了解了Numpy，pandas库的一些函数（这里有一些纸质笔记，后面编程使用一些函数的时候会翻开查看）

（2）数学知识的学习、补充：

第一周学习了矩阵，向量的有关知识：矩阵和向量的加减乘除、逆矩阵和矩阵的转置、向量的转置等等，在补充完这些知识后才对模型的公式有了更好的理解。

（3）代码实现：

第二周开始用代码实现模型，发现自己动手做很困难，于是先看了一些csdn的文章和别人的代码才动手写。

首先对数据进行预处理，对于训练集含有缺失值，采用以均值代替缺失值的方法，然后就直接开始写了梯度下降的算法，失败后才意识到忽略了数据预处理中一些重要的步骤——

再次补充学习：①归一化：归一化的目的就是使得预处理的数据被限定在一定的范围内（比如[0,1]或者[-1,1]），从而消除奇异样本数据导致的不良影响**。**②标准化：将数据变换为均值为0，标准差为1的分布。③特征缩放：将特征转化为相近似的范围（有待理解）。等等。对于数据预处理方法一些知识点还未学习完毕。