吴恩达机器学习公开课

1. 监督式学习

部分数据集已经有正确答案，预测另一些数据集的答案。

回归问题：给出十年的房价，得到明年的房价。回归问题是监督的一种。回归的意思是预测一个连续值的输出。

分类问题：通过肿瘤的大小预测是否恶性肿瘤，给定一系列（大小，是否恶性）的数据集，通过这些数据，预测一个给定大小的肿瘤是否为恶性的概率。给定的特征可以上没有上限的，支持向量机(SVM)算法可以支持无穷多的特征。

2. 无监督学习

聚类算法是无监督学习中的一种，让算法自行对给出的数据集进行聚类。

应用场景：计算机集群中，哪些计算机趋于协同工作，可以把他们放得近一点来提高效率；好友系统中，哪些是很亲密的好友组，哪些仅是普通朋友的好友组；客户信息划分，自行将不同的客户划分到不同的市场部分。

另一种无监督学习是非聚类的（non-clustering？？），比如鸡尾酒算法可以把两个重叠的音源分别分离出来。

**初学机器学习，最好使用Octave作为开发环境。**

3. 线性回归算法

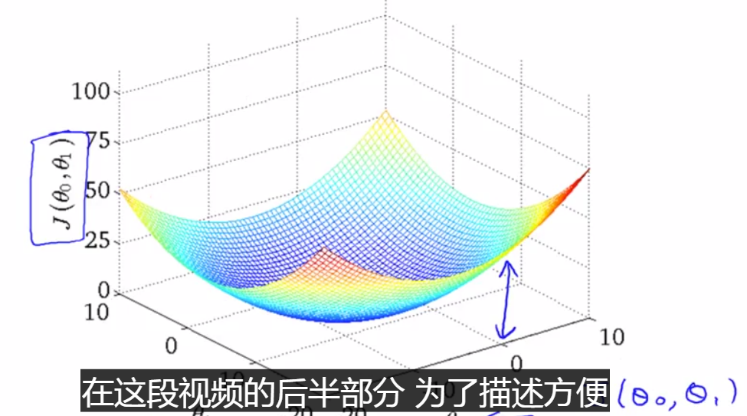
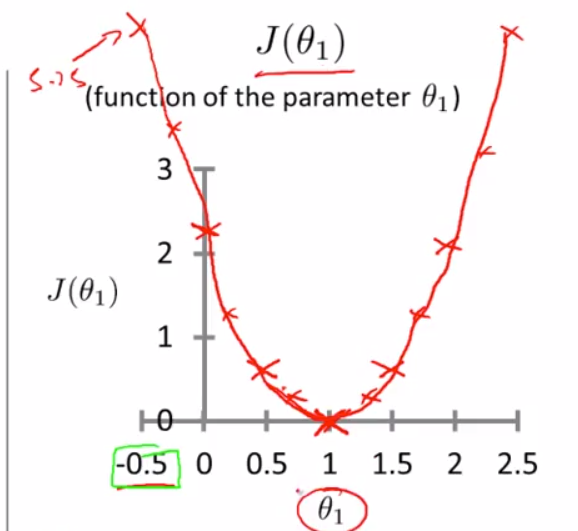
最简单的例子：通过历史数据预测特定面积的房子的价值，将历史数据输入到机器学习算法中，最终得到一个拟合出来的线性函数h(x) = kx + b

最后预测出来的直线，应该尽可能地让历史数据中的点落在或者靠近它，那么很自然的想法就是定义一个代价函数，可以量化地计算出一条直线与一群点之间的相似程度，对于这个问题，假设各个历史数据点为(xi, yi)。

那么单个点和直线的距离为；所有的点距离之和为 ；为了不受到历史数据数目的影响，因此最终的代价函数为。

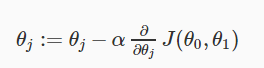
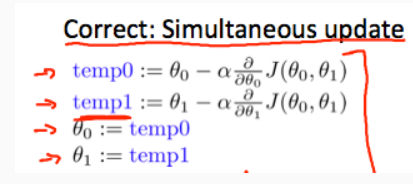
**所以问题转变为求出能够使得代价函数 最小的k,b，就可以确定预测结果h(x)**

如果假设b=0的话，h(x)=kx，那么代价函数就是关于k的二次曲线，可以很容易地找到它的最小值；如果b≠0的话，h(x)=kx+b，那么代价函数是关于k,b的曲面，找出其中的最小点就有点困难了。



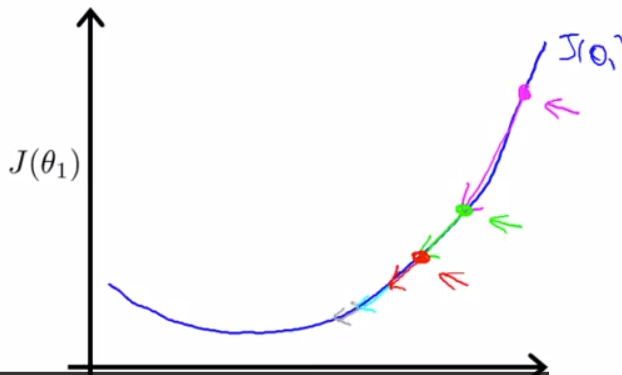
4. 使用梯度下降算法找出最优的代价函数

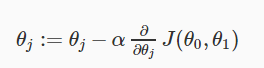
对于代价函数H(k,b)，给出一组特定的(ki,bi)，使用梯度下降算法，可以得到在(ki,bi,H(ki,bi))点，使得H(k,b)的值下降最快的方向，所以持续的进行梯度下降，可以最终得到一个局部的最小值，但不一定是绝对的最小值。

梯度下降公式：，其中α代表学习率，也可以说是每次下降幅度，而因为是以360°的范围来选择下降的方向，所以(k,b)一定是同时更新的，因此有这样的公式：，循环进行直到收敛。

5. 为什么梯度下降算法会起作用

首先考虑b=0的情况，那么代价函数就是关于k的二次曲线，那么整个梯度下降的过程如下图，会慢慢地逼近局部的最优点。

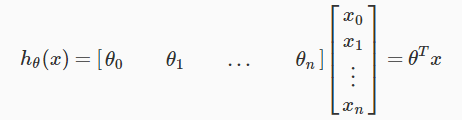


接着考虑b≠0的情况，那么代价函数是关于k,b的二元函数，，公式中分别对k和b求偏导代表着k，b各自的变化率，同时更新k,b的值，可以接近局部最小值。又因为线性回归方法的代价函数总是一个弓形的凸函数，所以使用梯度下降的方法，得到的总是全局最优解。

我觉得梯度下降算法更像是一种启发式算法，通过计算各个特征的偏导可以保证代价函数的值一定是下降的，然后就是设置一个学习率，尝试往降低代价函数值的趋势上走，最后走到局部的最低处。

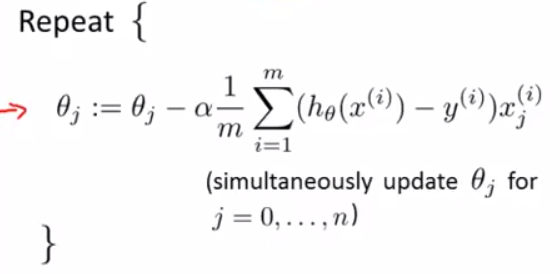
6. 多元线性回归

现在考虑特征不止一个的情况，h(x) = θ0+θ1x1+θ2x2 + …….

这样的公式可以写成矩阵的形式：

其中θ = [θ0，θ1，θ2，θ3……，θn]；x = [x0，x1，x2，x3……，xn]

仍然使用梯度下降的方法，可以简单地得到：



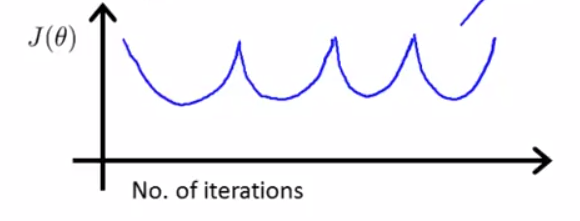
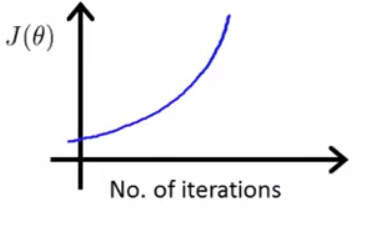
7. 梯度下降方法的使用技巧

a. 特征值范围的选取：

当多个特征之间的取值范围比例差距非常大的时候，比如x1∈[0, 5]，x2∈[0, 2000]时，通过代价函数画出的轮廓图会非常狭长，在进行梯度下降算法的时候，也会需要大量的时间才能收敛。

为了让梯度下降算法能够更快地工作，各个特征值最好是在同一个数量级范围内的，对于x2∈[0, 2000]，我们可以简单的把x2的值都除以2000，也可以使用其他的数学上的标准化方法。

b. 学习率α：没有什么好的方法可以提前确定最合适的学习率，一般来说先给定一个值，然后进行梯度下降算法，根据代价函数的值与迭代的次数形成的二维图来判断之前设定的学习率是大了，还是小了：



比如说在上面两张图的情况下，显然是学习率设置的太大了导致无法收敛，应该降低学习率α。

8. 多项式回归

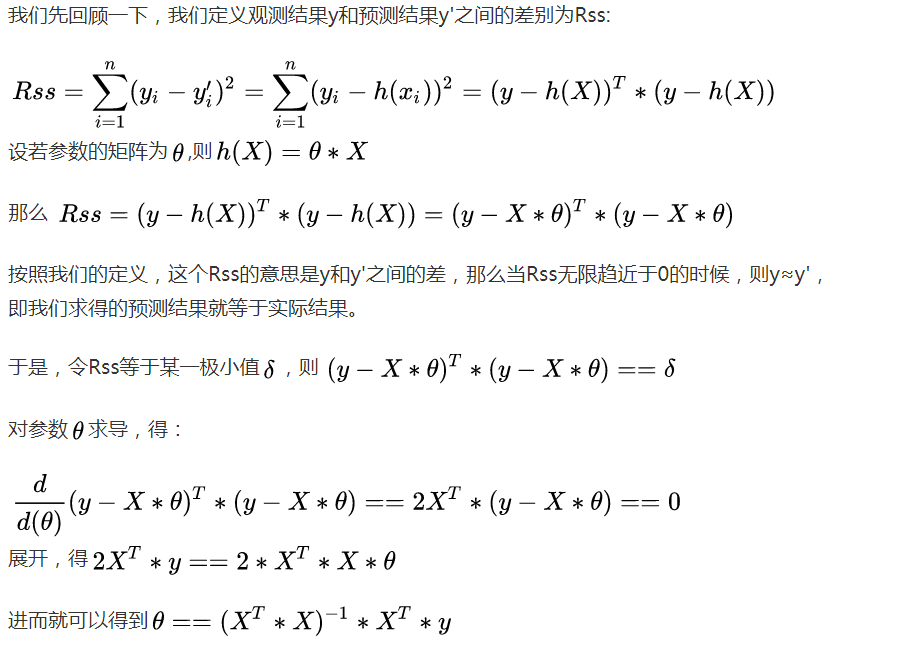
假设现在只有一个特征，如果想用二次曲线去拟合数据，那么就会有x的平方项，比如：

，这种情况下就将x^2作为第二个特征，转化成多元线性回归的方法来进行计算，同时要注意特征值范围的问题，最好统一在一个数量级里，以保证梯度下降算法正常工作。

9. 正规方程

梯度下降的算法可以让我们逐步接近最优解，而正规方程能让我们通过计算一步就得到最优解。

下图为正规方程的推导：



使用正规方程，通过矩阵的计算，一步就能得到最优解的值。

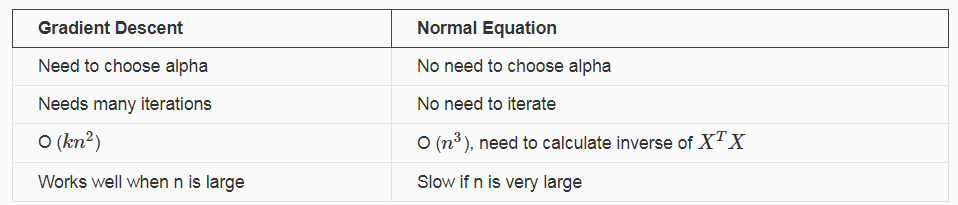
如果XTX不可逆怎么办？？有待仔细看？？？

一般来说，当XTX不可逆的时候，要么是特征的维度远大于训练的样本数目；要么是特征矩阵中含有多余的特征，比如说两个线性相关的值，被分成了两个特征。(经验之谈)

10. 线性回归中梯度下降与正规方程的选择

梯度下降：学习率α需要自行取值，在学习过程中还需要人工地调整；但是，在特征矩阵的维度特别大的情况下，只要学习率合适，梯度下降算法仍然能保证他的效率，总共要进行n\*m\*r次计算，n是特征维度，m是学习样本数目，r是迭代次数。

正规方程：由于不需要人工调整学习率，只需要进行矩阵运算就能得到最优解，因此在矩阵的维度比较小的情况下，正规方程无疑比梯度下降更有优势。但是，求逆矩阵的计算复杂度大约是O(n^3)，所以在特征向量维度特别大的情况下(n>10000)，正规方程的性能非常低下，这时梯度下降的方法就能更快的得到最优解。



11.