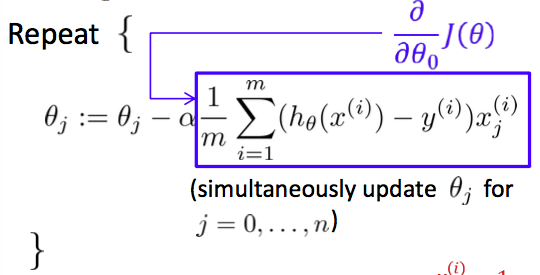
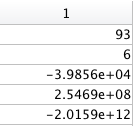
数据挖掘第一次作业

**1、线性回归**

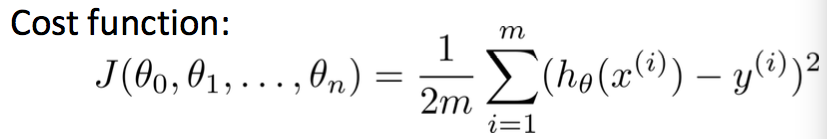
(1)这里我们按照ppt中所给的计算公式：



在matlab中写程序进行相应实现可以算出我们梯度下降第一次迭代之后的参数向量为：



(2)这里我们按照ppt中求代价函数的公式：



在matlab中用代码实现即可计算出两个参数向量与分别对应的代价函数值cost与cost1为：



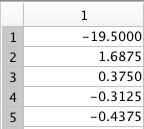
所以经过比较我们可以看出经过第一次迭代得到的参数向量并不会使得代价函数下降。

(3)我们通过(2)中的计算结果可以看出由于我们选取的学习率太大，从而导致经过迭代之后代价函数反而变大了，所以我们可以选取一个小一些的比较合适的学习率，例如0.001，即可使得经过第一次迭代后代价函数下降。

(4)同上，我们按照ppt中的标准方程公式：



在matlab中计算可得相应的参数向量为



所以我们求得最优的多元线性回归方程(精确到0.01)为

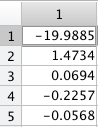
m=-19.50+1.69\*p+0.38\*c-0.31\*e-0.44\*ch。

所以代入数据可得所需预测数学分数为m\_pred=-19.50+1.69\*88+0.38\*73-0.31\*87-0.44\*92=89.51。

(5)在matlab中我们可进行如下计算;



可以求得符合题意的参数向量如下：



所以最优的L2正则化多元线性回归方程(参数精确到0.01)为：

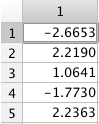
m=-19.99+1.47\*p+0.07\*c-0.23\*e-0.06\*ch，在本问中并未给定相应数据，所以我们无法判定两个线性回归方程那个好。但是加入了正则化参数可以减轻过拟合的问题，但是可能会造成欠拟合的问题，这主要取决于具体数据值。

**2、逻辑回归**

(1)我们这里选择使用逻辑回归函数来求解，所以我们在matlab中先按照上述表格中的数据来输入对应X与Y，然后再使用：



可以计算得到相应的参数向量为



(2)通过我们上述所求得的参数向量我们可以看出是否使用过非雌激素是对影响子宫内膜癌发病的最直接的因素。

**3、支持向量机**

(1)画图如下：

这里我们将支持向量设为w = (a, b);

则可得方程组：

a + = -1;

b + = -1;

2\*a + = 1;

a + b + = 1；

所以解上述方程可以得到

a = 2; b = 2; = -3

所以超平面方程为g(x) = 2\* + 2\* - 3。

(2)由于新增的训练样本点是能够被正确分类并且是远离最优超平面的，所以它们并不会影响决定最优平面的支持向量，所以并不会影响最优超平面。但是线性回归是根据每一个训练样本点而得到的，所以加入远离边界的训练样本点之后会造成边界线偏移。

(3)支持向量为(0, 1), (1, 0), (1, 1), (2, 0);

两个异类支持向量到最优超平面的距离之和为2 \* / 4 = / 2。