西安電子科技大學

题　　目：________________________

学　　院：________________________

专　　业：________________________

姓　　名：________________________

学　　号：________________________

一．实验目的

熟练掌握利用梯度下降法求解一维和多维线性回归问题。

现有数据集 data1 和数据集 data2，data1 包含 97 个房间大小及其对应房间价格的样本数据，data2 包含 47 个房间大小、卧室数量及其对应房间价格的样本数据。请利用梯度下降法进行一维和多维线性回归，并完成以下问题：

（1）画出样本分布图。

（2）画出线性回归假设模型。

（3）画出成本函数收敛曲线。

二． 实验环境

MATLAB 程序语言设计。

三．实验内容和步骤

1 假设函数

一元线性回归的假设函数的一般形式是 $h_\theta(x)=\theta_0+\theta_1 x$，一般情况下，$h_\theta(x)$ 可以简写为 $h(x)$ 。

但假设函数的初始化是由样本集猜想而来，并不知道初始化的假设函数是否准确。而且假设函数中参数 θ_0 和 θ_1 的值会影响梯度下降的结果，因为梯度下降算法有可能收敛扎起局部的最小值。而函数局部最小值，而不一定是全局最小值。

2 成本函数

成本函数是用来衡量假设函数的准确性的函数。成本函数的值越小，说明假设函数越准确。而我们的目标就是要提高假设函数的精度，故成本函数又称为目标函数。

$$J(\theta_0, \theta_1)=1/2m\Sigma_{m1}(h_\theta(x_i)-y_i)_2$$

其中 $h_\theta(x_i)$ 是 x_i 根据假设函数计算的预测值，y_i 是样本的实际值。m 是样本的个数，成本函数亦被称之为平均均方误差，除以 2 是为了方便后面梯度下降算法的计算。

3 梯度下降算法

成本函数是来衡量假设函数的准确性，那提高假设函数的准确性则要靠梯度下降算法。改变假设函数，其实是改变θ_0和θ_1的值，通过改变θ的值，从而使成本函数J（θ_0，θ_1）最小。所以这里的自变量是θ_0和θ_1，因变量是J（θ_0，θ_1）。

实际上这是一个求极小值的过程，实际上若是目标函数是可以求导的情况下，参数较少的情况下，个人觉得是可以通过对每个自变量求偏导数，通过让偏导数为零来求得目标函数最小化时的自变量。

但在不可求导或是参数太多以至于求值很困难的情况下，使用梯度下降算法。该算法目的是找出梯度下降的方向，是自变量沿着梯度逐步减小，最终收敛到局部最小值。

repeat until convergence:

$\theta_j = \theta_j - \alpha\partial/(\partial\theta_j)J(\theta_0, \theta_1)$

$\theta_j = \theta_j - \alpha 1/m\Sigma_{m1}((h_\theta(x_i) - y_i) * x_{ij})$

for j = 0 and j = 1

注意：θ_0和θ_1的值是同步改变的

$\partial/(\partial\theta_j)J(\theta_0, \theta_1)$是自变量的偏导数，即自变量在该点的斜率，α是学习速度，为正。表明以多大的幅度来收敛，若幅度太小，收敛步数增加，若幅度太大有可能越过收敛点导致收敛困难。$\alpha\partial/(\partial\theta_j)J(\theta_0, \theta_1)$即为斜率乘以幅度得出下降的高度，一般情况下α不变，但是偏导数会随着收敛而逐步变小，故下降的高度会随着迭代次数而减小，而这也是该算法的优点之一。

4.3 多维线性回归（以本题为例阐述）

➢ **原理描述**

设 m 代表训练集中实例的数量，n 代表自变量的个数，x_i表示第 i 个输入变量（本题为房间大小和卧室数量），$i = 1 \ldots n$，y 表示输出变量（本题为房间价格），则$x_i^{(j)}$代表训练集中的实例，当x_i都与 y 线性相关时，这种回归分析称为多维线性回归分析。画出样本分布图如图 2 所示，由图可知x_1, x_2和 y 线性相关，二者的关系可用一个平面近似表示，其关系式如下

$$y = h_\theta(x) = \theta_0 - \theta_1 x_1 + \theta_2 x_2 = \theta^T X \qquad (12)$$

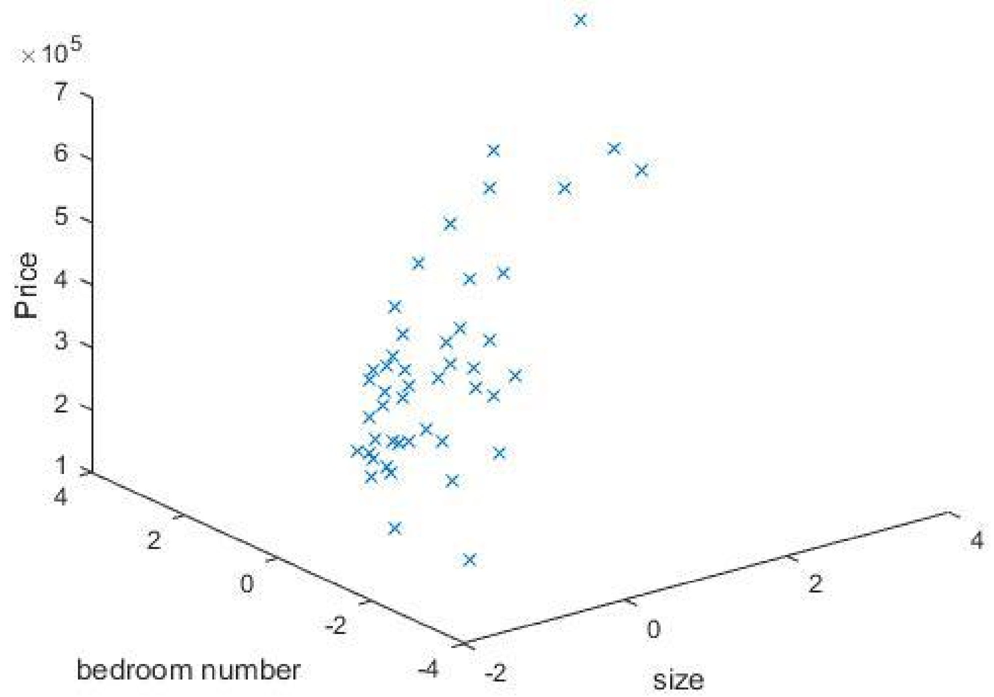

图 2 data2 样本分布图

因此问题转换成如何确定 $\theta_0, \theta_1, \theta_2$ 的值使得拟合出的曲线更加接近实际的增长情况，即模型误差的平方和能够最小。因此定义平方误差函数为

$$J(\theta_0, \theta_1, \theta_2) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

- **特征缩放**

在面对多维问题时，要保证自变量 X 都具有相近的尺度，这将帮助算法更快的收敛。因此对自变量进行如下处理

$$x_n = \frac{x_n - \mu_n}{s_n}$$

其中 μ_n 是平均值， s_n 是标准差。

- **多维线性回归中的梯度下降法**

在此问题中，将更新规则改为公式（1）（2）（3）即可，其他与一维线性回归中的梯度下降法相同。

$$\theta_0 = \theta_0 - t \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)} \qquad (1)$$

$$\theta_1 = \theta_1 - t\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_1^{(i)} \qquad (2)$$

$$\theta_2 = \theta_2 - t\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_2^{(i)} \qquad (3)$$

四．实验结果与分析

5.1 一维线性回归

➢ 线性回归假设模型（θ_0=-3.2414, θ_1=1.1273）

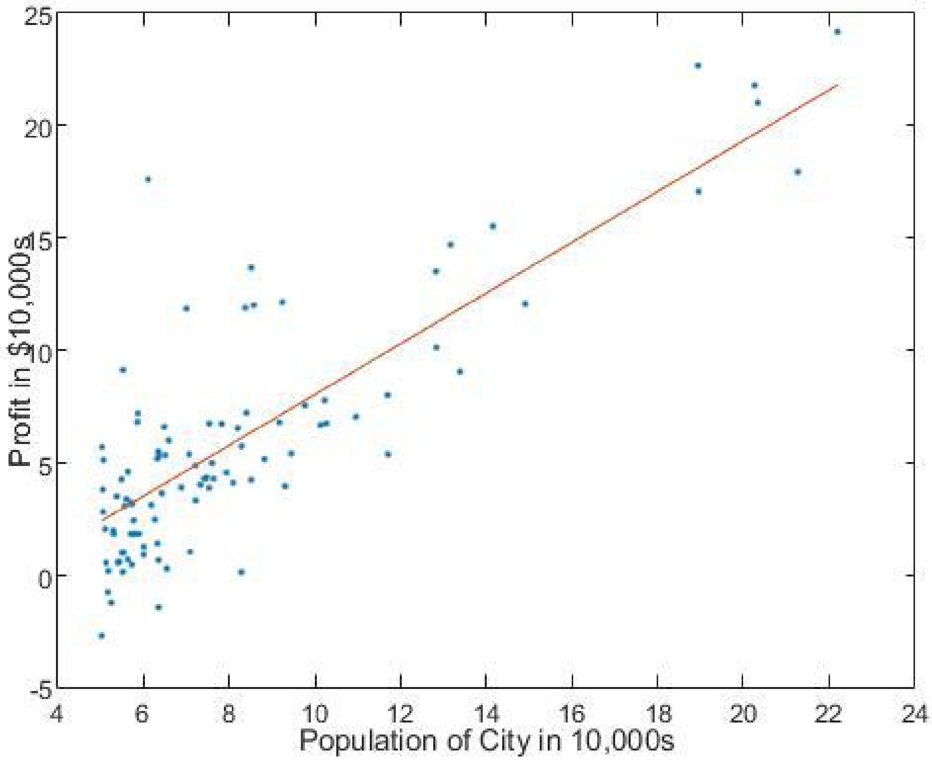

图 3 线性回归假设模型图

由图可知，曲线较好的拟合 profit 和 population 的关系。

➢ 成本函数收敛曲线

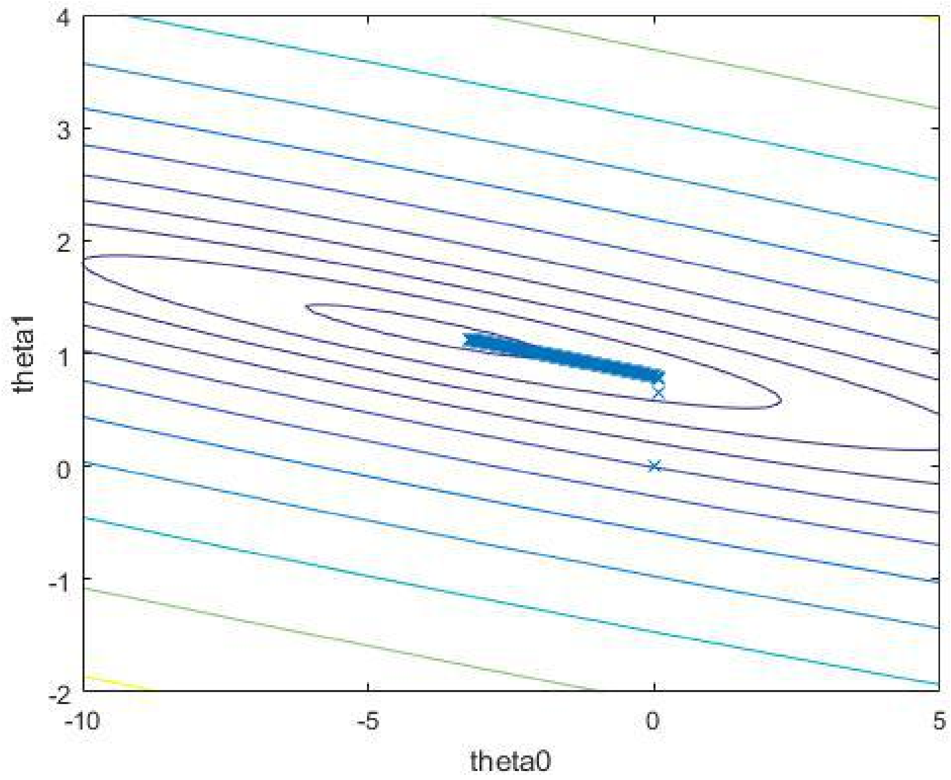

图 4 成本函数收敛等值线曲线图

5.2 多维线性回归

➢ 线性回归假设模型(θ_0=334302.06, θ_1=100087.11, θ_2=3673.55)

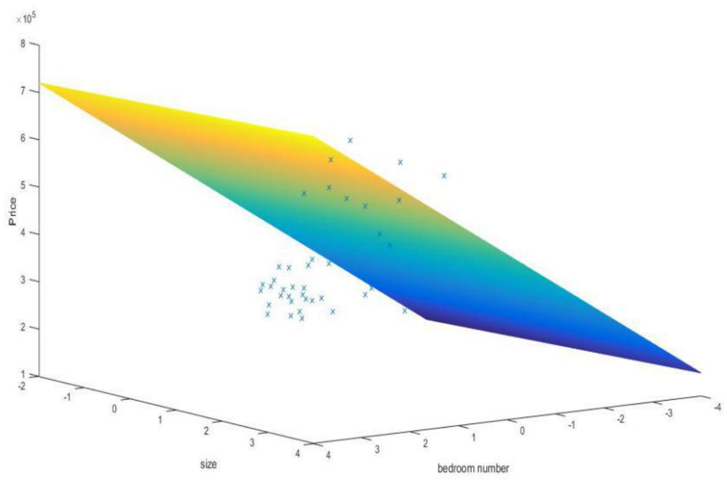

图 5 多维线性回归假设模型图

由图可知，曲线较好的拟合了数据使其均匀的分布在平面两侧。

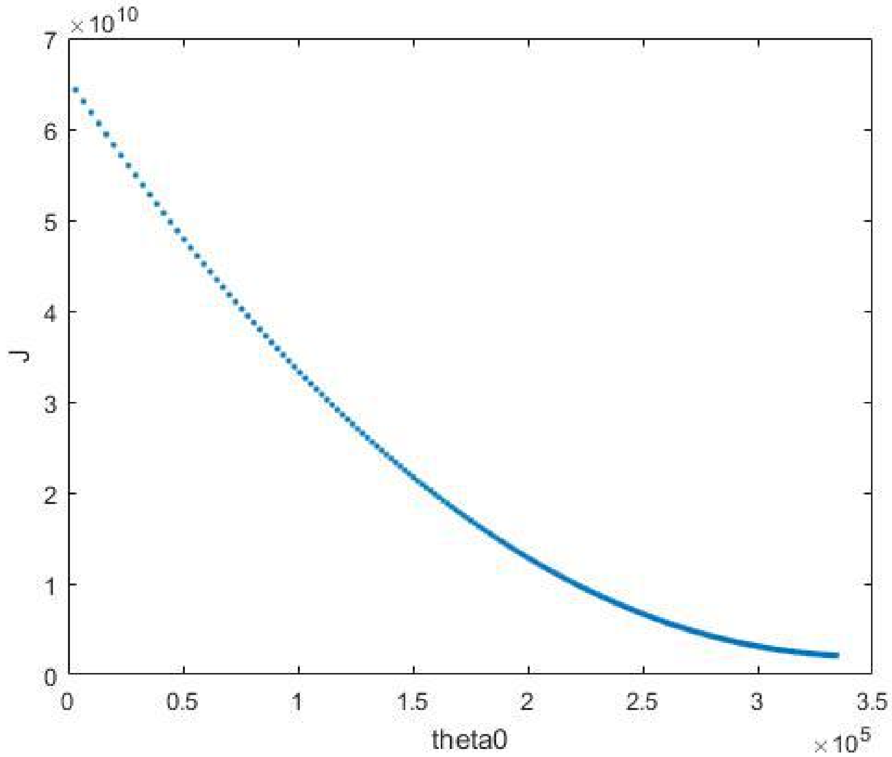

图 6 成本函数随 θ_0 收敛曲线图

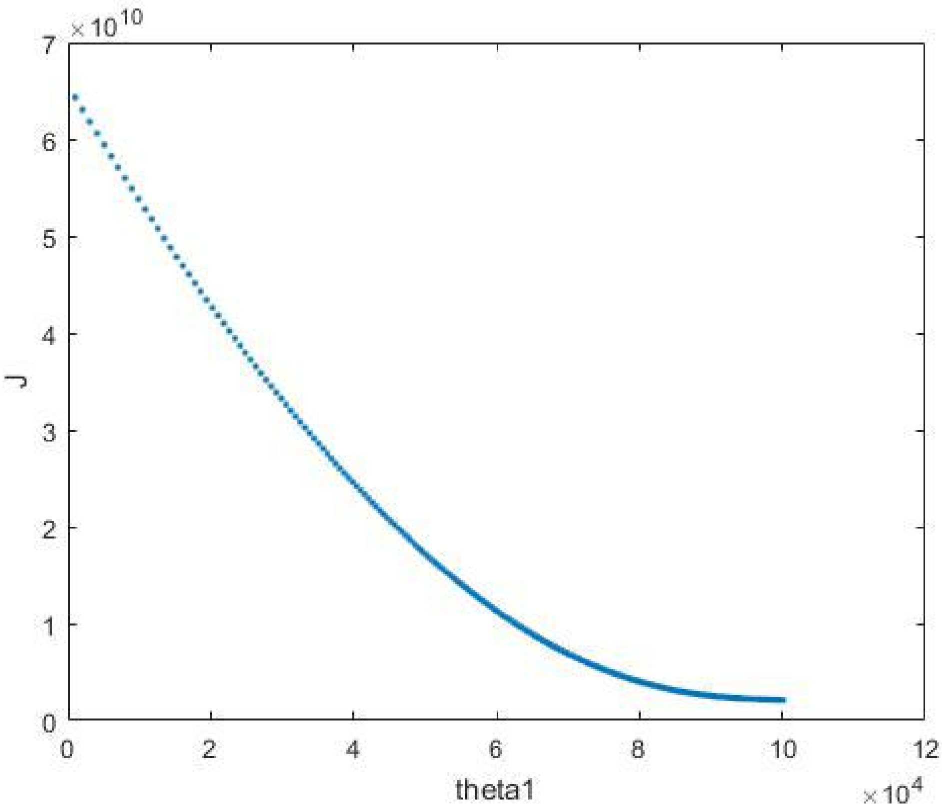

图 7 成本函数随 θ_1 收敛曲线图

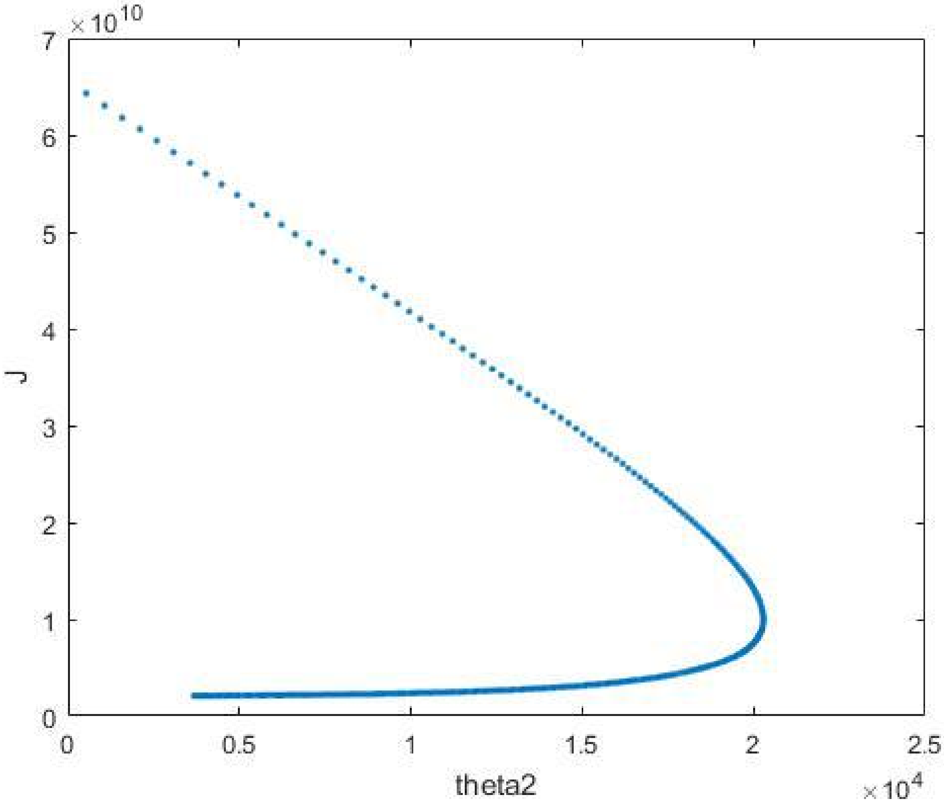

图 8 成本函数随 θ_2 变化曲线图

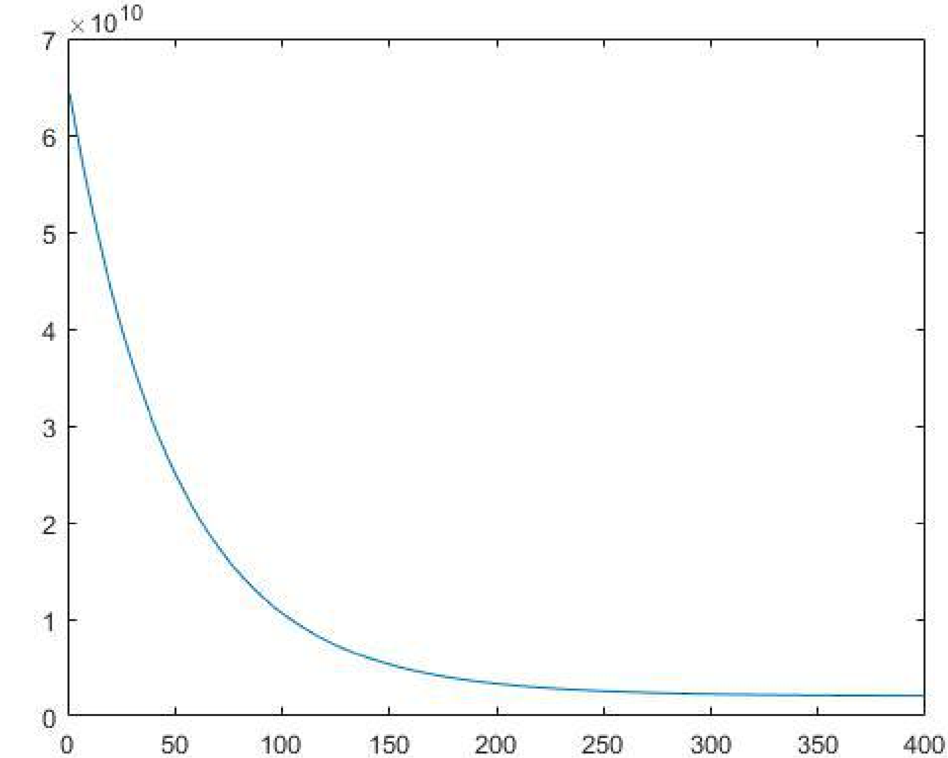

图 9 成本函数随迭代次数收敛曲线图

五．总结

通过此次上机实验，对梯度下降法有了更深的理解，知道了归一化，学习速度，代价函数这些原来不了解的东西，在查询资料中对机器学习有了一点点理解。很感谢老师此处布置的上机任务，收获很大。