

# 线性回归

- 假设你是一个餐饮连锁店的CEO，你打算在不同的城市开设不同的分店。你已经在一些城市开了分店而且你有这些城市人口与利润的数据（见data1a.txt），你希望通过这些数据来决定在哪些城市新开分店（也就是通过新城市的人口预测新城市的利润）。
- 数据格式如下表所示（每个城市包含人口数和利润两个数据）

人口数 (单位: $10^4$ 人)	利润 (单位: $10^4$ 美元)
6.1101	17.592

- 使用梯度下降法获得线性回归参数，并使用训练好的模型来预测以下两个城市的利润情况：  
城市A：人口数35000人  
城市B：人口数70000人

假设:  $h_{\theta}(x) = \theta_0 + \theta_1 x$

参数:  $\theta_0, \theta_1$

代价函数 (Cost function):

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

目标  $\underset{\theta_0, \theta_1}{\text{minimize}} J(\theta_0, \theta_1)$

# 批量梯度下降算法

需要注意：在更新  $\theta_0, \theta_1$  时，两者必须同步更新。

## Gradient descent algorithm

repeat until convergence {  
→  $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$  (for  $j = 0$  and  $j = 1$ )  
}

### Correct: Simultaneous update

temp0 :=  $\theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$   
temp1 :=  $\theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$   
 $\theta_0 :=$  temp0  
 $\theta_1 :=$  temp1

### Incorrect: ↙

→ temp0 :=  $\theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$   
→  $\theta_0 :=$  temp0  
→ temp1 :=  $\theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$   
→  $\theta_1 :=$  temp1

# 代价函数的导数

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

$$j = 0 \text{ 时: } \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

$$j = 1 \text{ 时: } \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m ((h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x^{(i)})$$

# $\theta$ 初始化

$\theta$ 需要初始化，可以初始化为全0，或者随机初始化。

假设 $\theta$ 初始化为全0，计算 $J(\theta)$  的值。

$\theta$ 初始化为全0 时， $J(\theta)$  应该等于32.07 ，你可以通过这个值验证你的程序是否有错误。

# 循环终止条件设定

循环终止条件可通过如下方式设定：

1. 设定一个比较大的迭代步数。
2. 画出 $J(\theta)$  随迭代步数变化的图。
3. 当两次迭代获的 $J(\theta)$  差异较小时终止迭代。

# Python相关

如果你使用python编程，那么你需要用到numpy和matplotlib库。  
如果你已经安装了Python，可以用pip安装上述包：

```
$pip install numpy matplotlib
```

Numpy: 科学计算基础包，用于数组运算。

```
import numpy as np
```

Matplotlib:科学绘图库，用于绘图

```
import matplotlib.pyplot as plt
```

# 绘制二维散点图

```
def plot_data(X,Y,title,xlabel,ylabel):  
    plt.plot(X,Y,'ro',markersize=6)  
    plt.title(title,fontsize=20)  
    plt.xlabel(xlabel,fontsize=10)  
    plt.ylabel(ylabel,fontsize=10)  
    plt.ioff( )
```

## 实验报告内容：

### 1、实验内容

实验要解决的问题、采用的模型或算法等

### 2、实验设置和实验结果

迭代终止条件的设置

梯度下降法获得线性回归参数

回归模型在所有训练数据 (train\_data.txt) 上最终的  $J(\theta)$  值。

城市A和城市B的预测利润。

循环过程中  $J(\theta)$  随迭代步数变化的图

### 3、其它（其它你觉得需要写在实验报告中的内容）

### 4、实验过程中遇到的问题

### 5、实验心得体会。

## 注意事项：

1、实验报告请使用老师提供的实验模板。

2、实验报告命名：ML\_project1\_完整学号\_姓名.doc

3、建立个人文件夹放实验报告和源代码(源代码需加注释说明)，文件夹名“ML\_project1\_完整学号\_姓名”

报告提交时间：第4周周四。