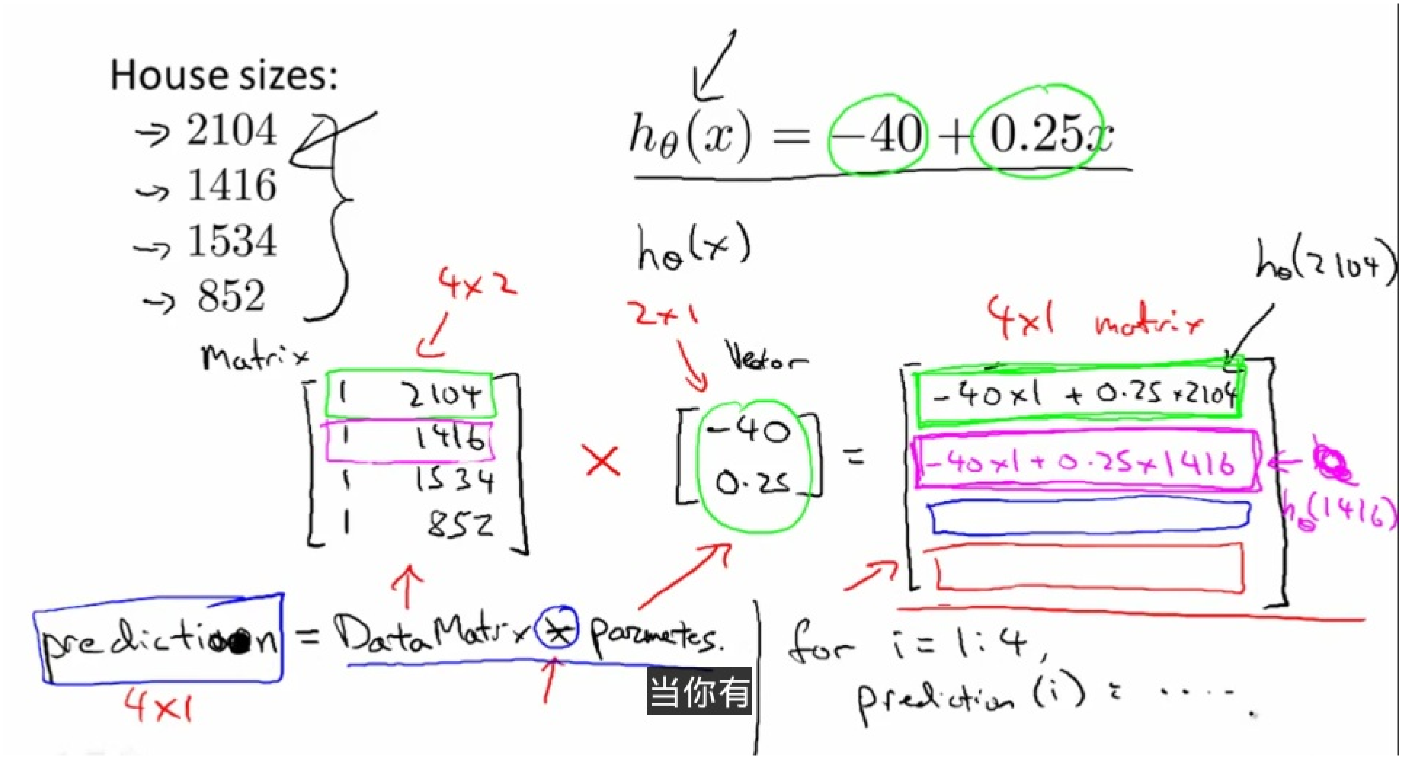
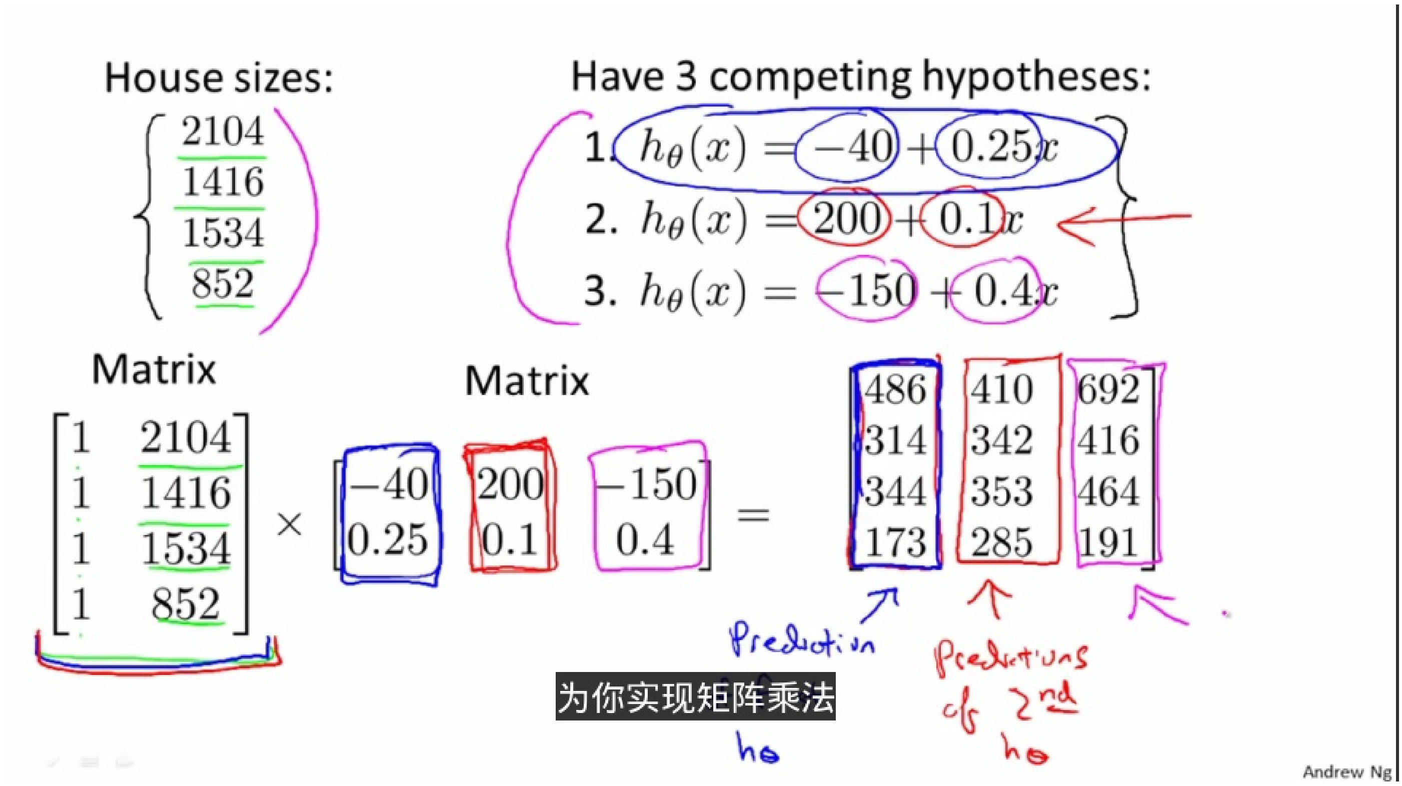
1. 矩阵和向量。矩阵加法、标量乘法、矩阵乘法。

用矩阵运算简化编码，提升运行速度。如：预测函数：

单个预测函数计算：



多个预测函数同时计算：



2. 矩阵相乘的性质：（1）不满足交换率（2）满足结合率 （3）单位矩阵 AI=IA=A （4）矩阵的逆 （5）矩阵转置

3. 多变量：n表示特征/变量数量。

Multiple features (variables).

Size ($feet^2$) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

$m = 47$

Notation:

- n = number of features $\quad n = 4$
- $x^{(i)}$ = input (features) of i^{th} training example.
- $x_j^{(i)}$ = value of feature j in i^{th} training example.

$$x^{(2)} = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \end{bmatrix}$$

$$x_3^{(2)} = 2$$

假设的形式改变：

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

为了记号方便起见，定义x0 = 1. 有：

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

For convenience of notation, define $x_0 = 1$. $\quad (x_0^{(i)} = 1)$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$h_\theta(x) = \theta_0 \overset{=1}{x_0} + \theta_1 x_1 + \cdots + \theta_n x_n$$

$$= \theta^T x.$$

$$\underbrace{\begin{bmatrix} \theta_0 & \theta_1 & \cdots & \theta_n \end{bmatrix}}_{\theta^T} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$(n+1) \times 1$ matrix $\qquad x$

$\theta^T x$

4. 多元模型中的梯度下降方法： 可以将theta写成向量形式。

Hypothesis: $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$ $\quad x_0 = 1$

Parameters: $\cancel{\theta_0, \theta_1, \ldots, \theta_n}$ $\quad \theta \quad$ n+1 - dimensional vector

Cost function:

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

$J(\theta)$

Gradient descent:

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} \cancel{J(\theta_0, \ldots, \theta_n)} \; J(\theta)$$

}

就是代价函数的对参数 θj 的偏导数 every $j = 0, \ldots, n$)

对thetaj求偏导数，会剩余xj，于是有：

Gradient Descent

Previously (n=1):

Repeat {

$$\theta_0 := \theta_0 - \alpha \underbrace{\boxed{\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)}}_{\boxed{\frac{\partial}{\partial \theta_0} J(\theta)}}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)}$$

$x_1^{(i)}$

(simultaneously update θ_0, θ_1)

}

New algorithm $(n \geq 1)$:

Repeat {

$\frac{\partial}{\partial \theta_j} J(\theta)$

$$\theta_j := \theta_j - \alpha \boxed{\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}}$$

(simultaneously update θ_j for $j = 0, \ldots, n$)

}

$$\boxed{\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_1^{(i)}$$

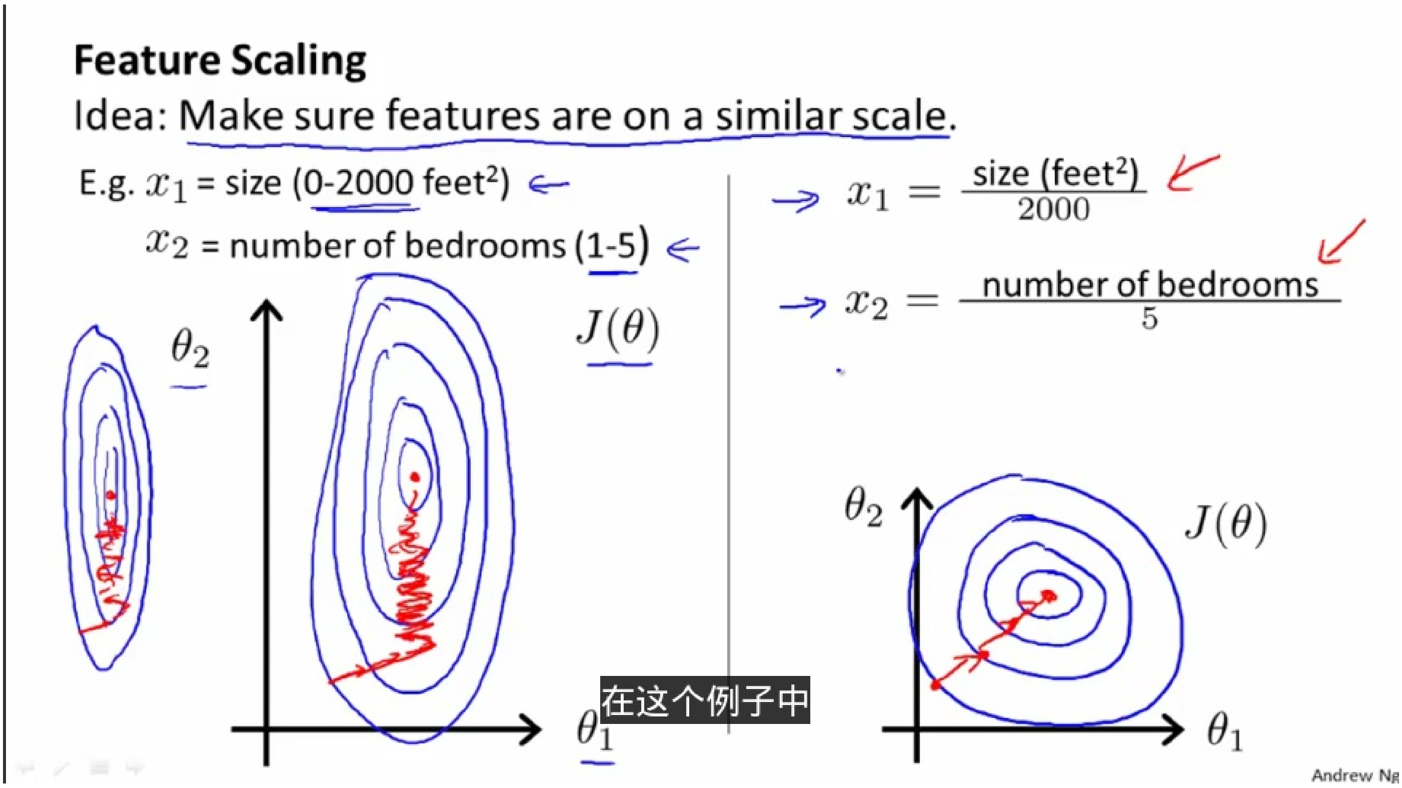
$$\theta_2 := \theta_2 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_2^{(i)}$$

n=1的情况相同

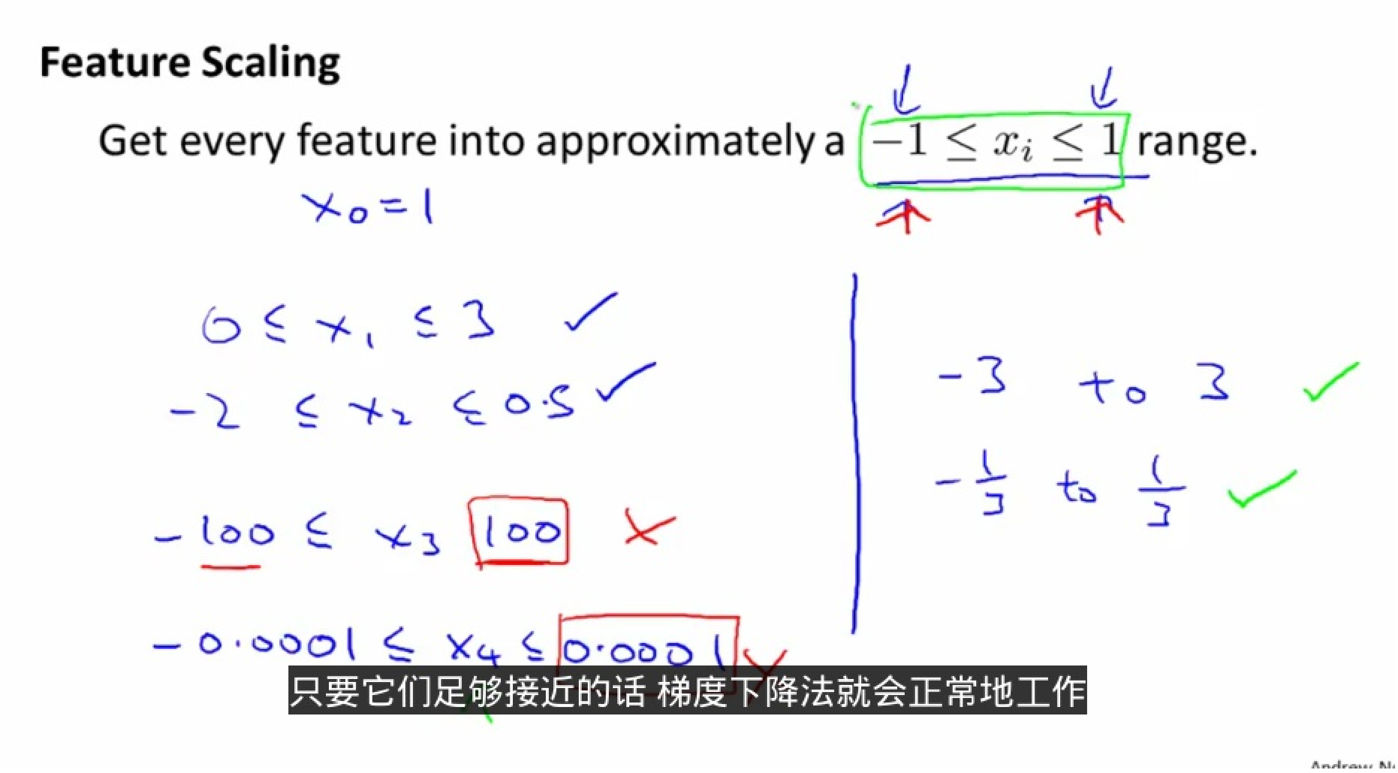
...

Andrew Ng

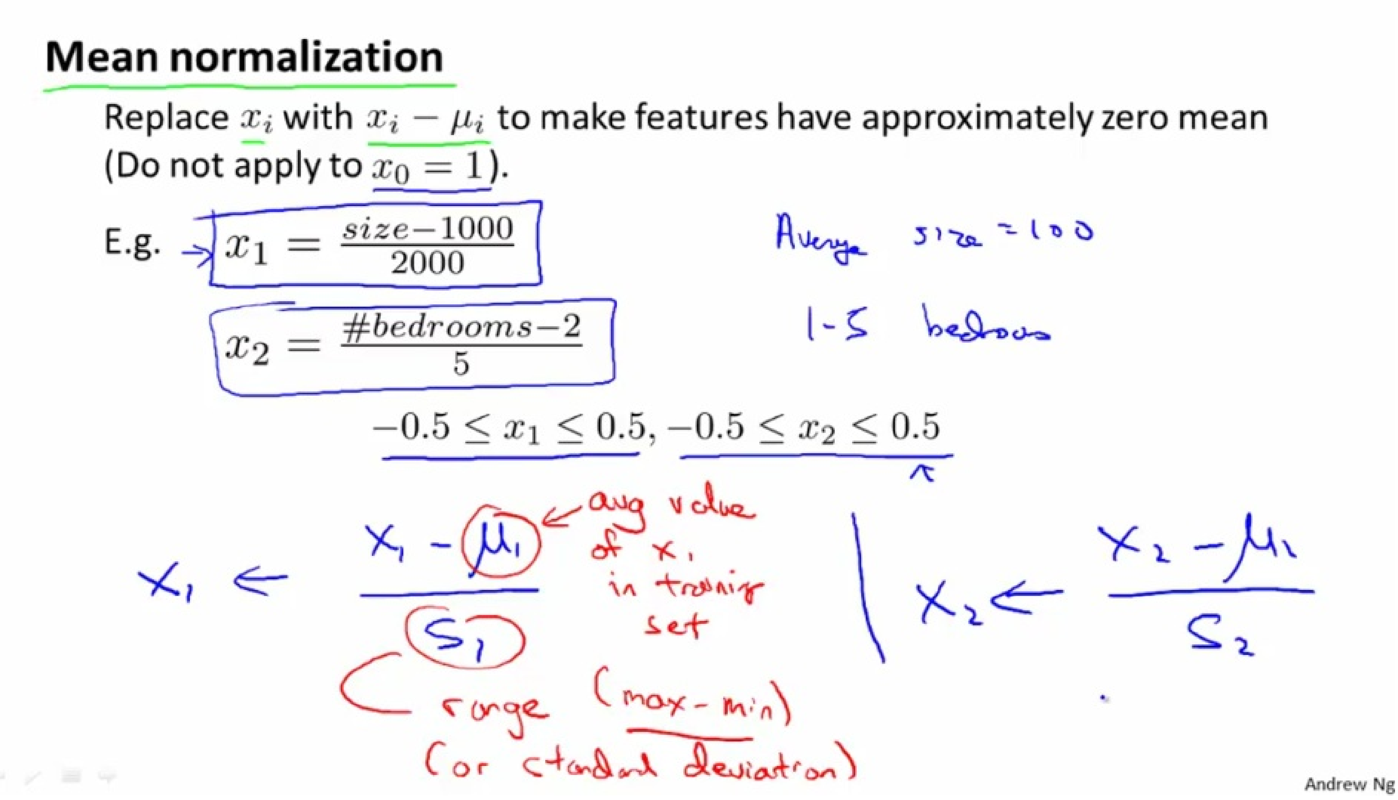
5. 特征缩放（Features Scaling）：确保不同特征在相似的范围之内，从而加快代价函数收敛速度：



通常情况下，将特征都缩放在大约[-1, 1]范围之内。这里不管太大或者太小都不合适，绝对值在接近1的适当范围是可以接受的。



均值归一化（Mean normalization）：用xi-ui来代替xi从而保证特征均值大约为0.（ui是数据集i的均值，注意不要运用到x0=1上）。

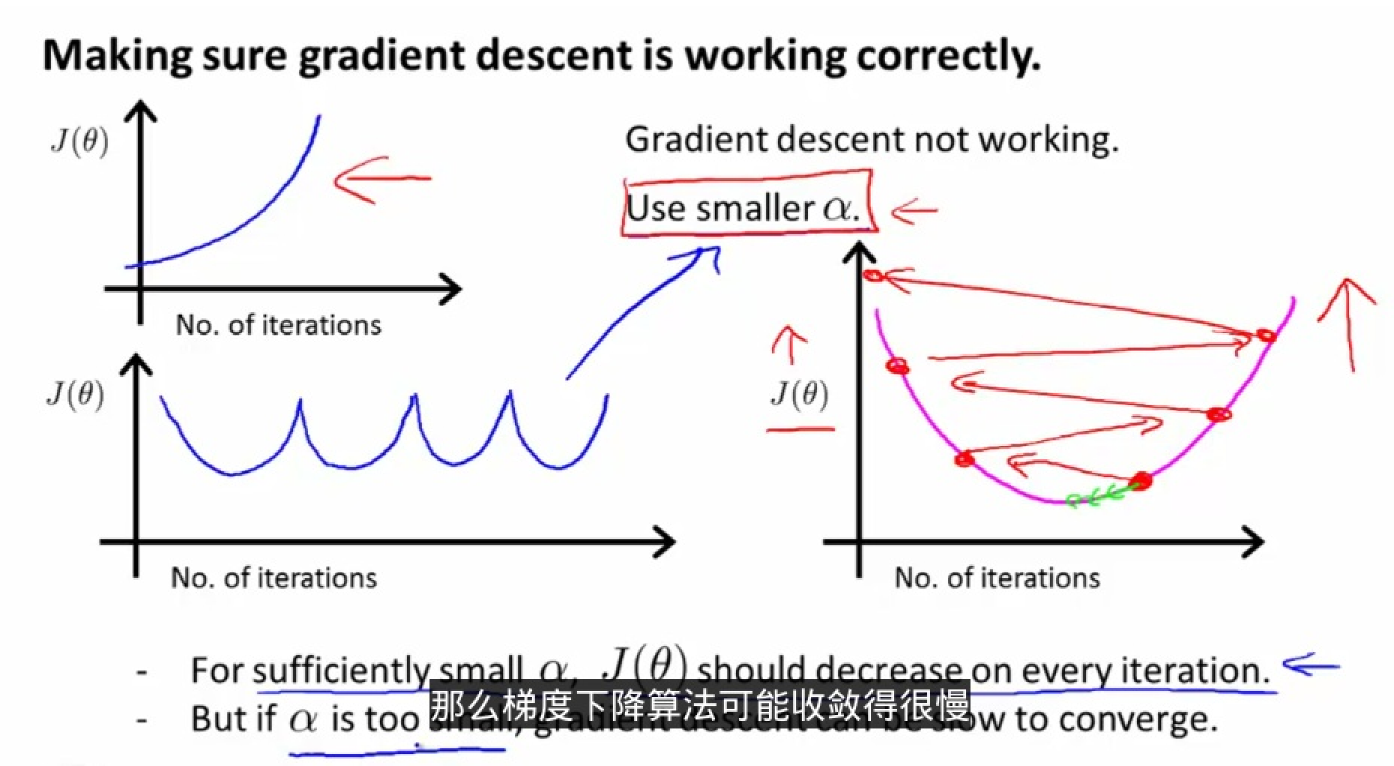


其中S不必是标准差，范围即可（即最大-最小）。

6. 学习速率。

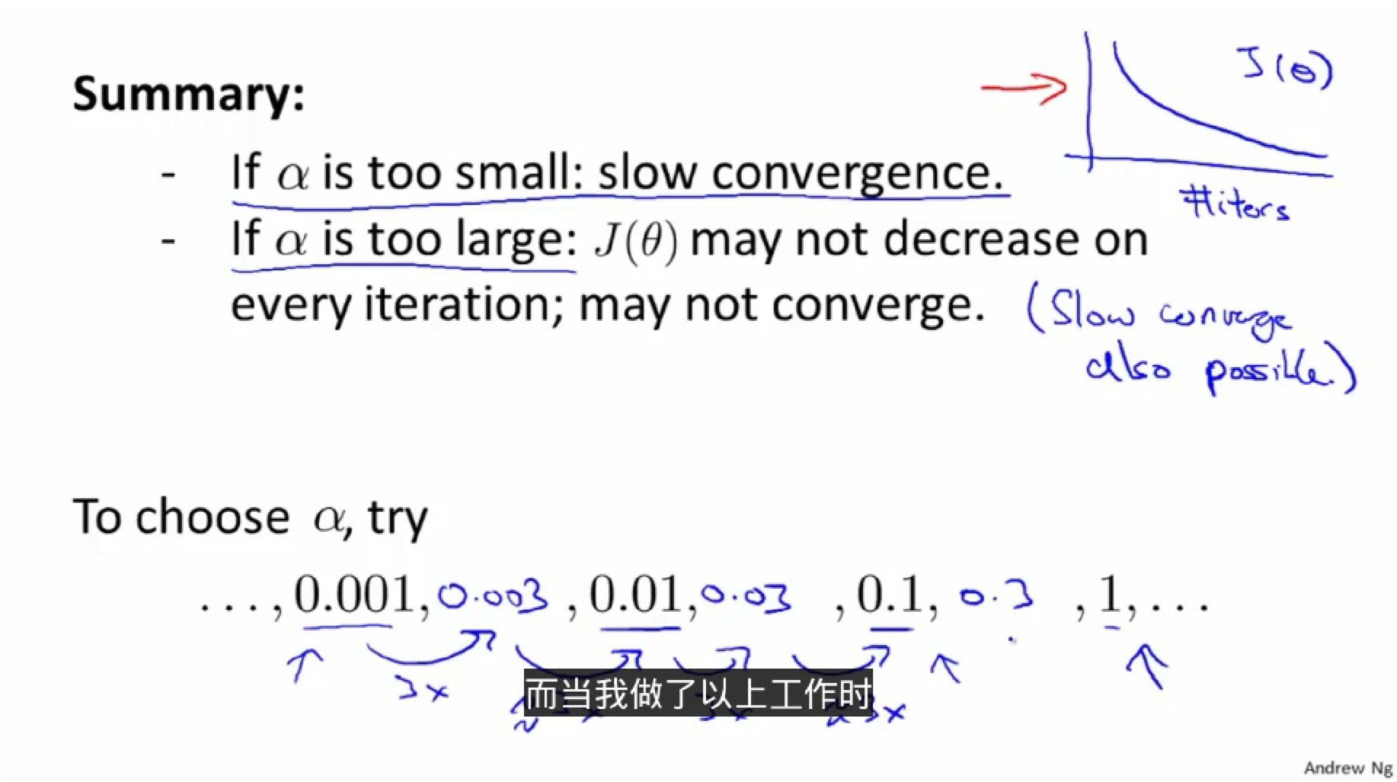
首先要确保梯度下降算法正确工作，也就是说在每一次迭代中代价函数J都必须下降，在达到某一步代价函数不再变化，已经接近收敛。

如果出现问题，需要调整学习速率：



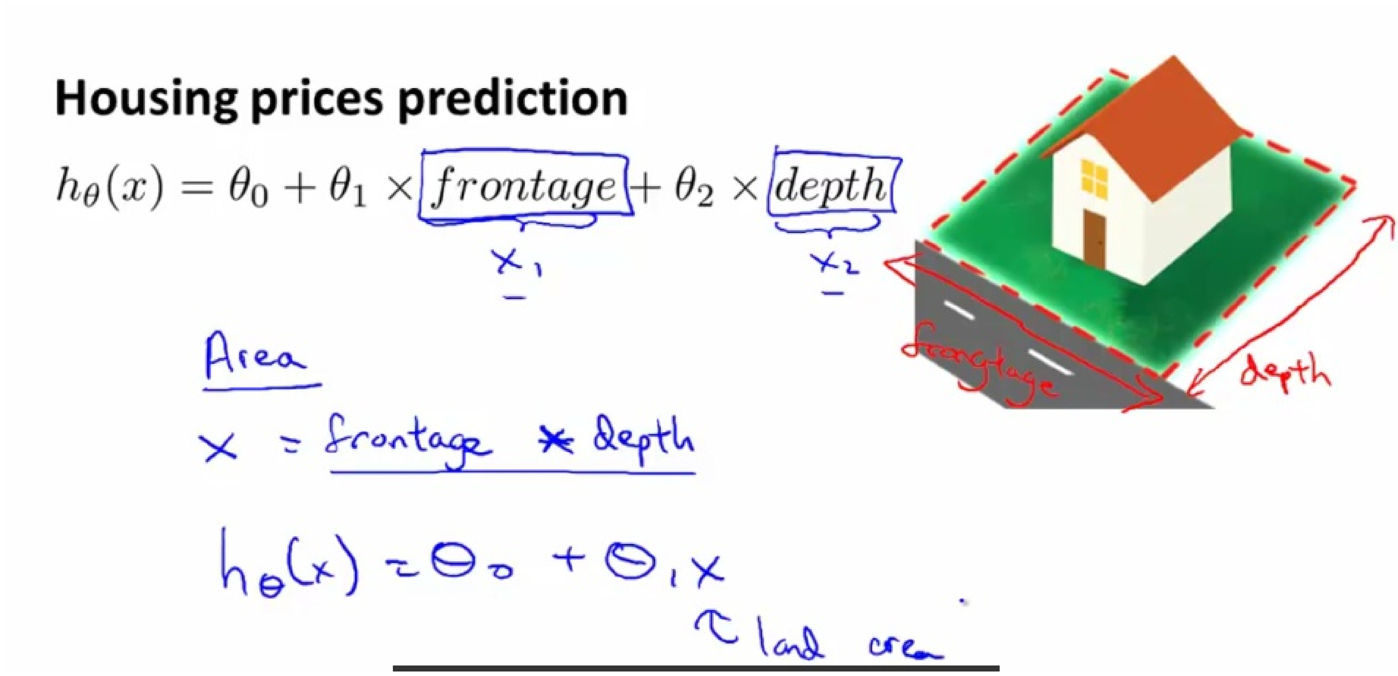
总结：

选择学习速率，... 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, ...

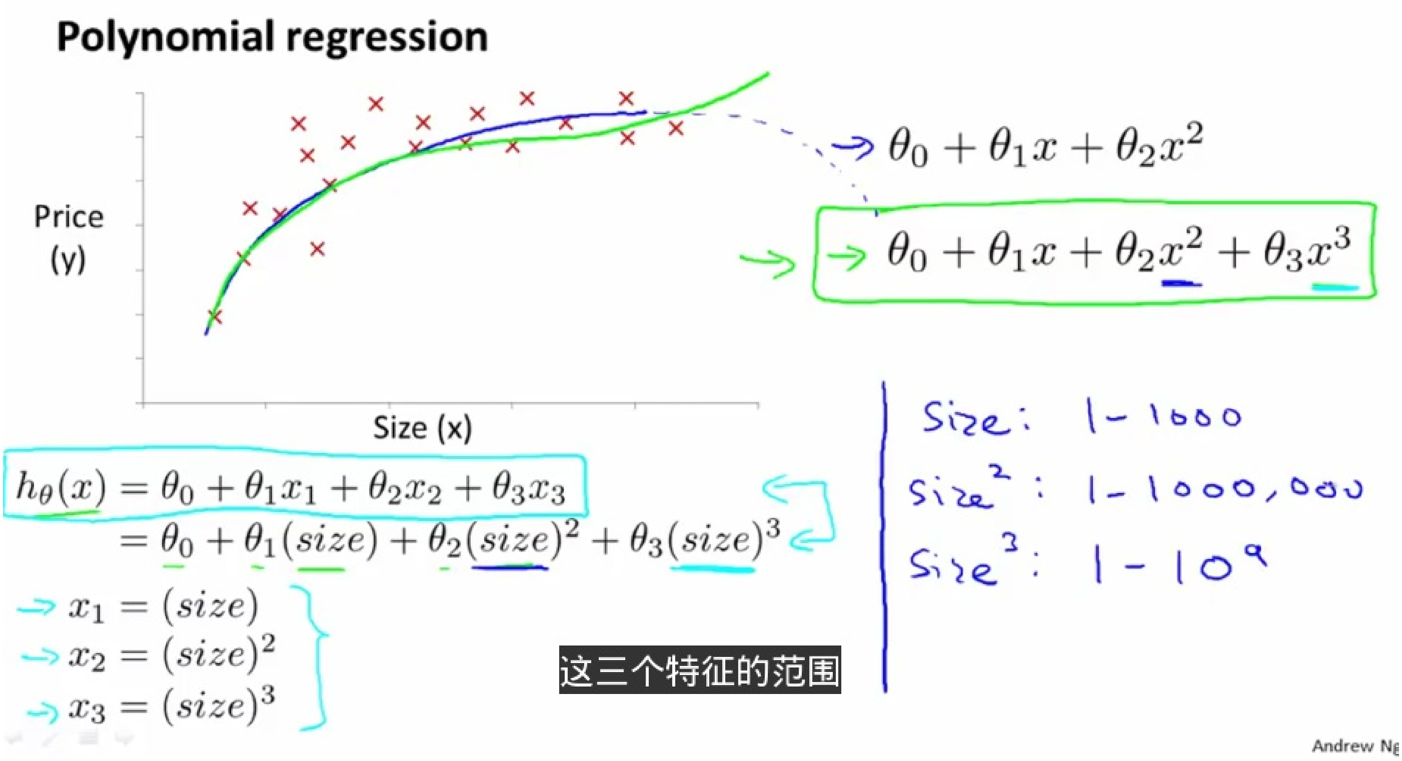


7. 特征和多项式回归。

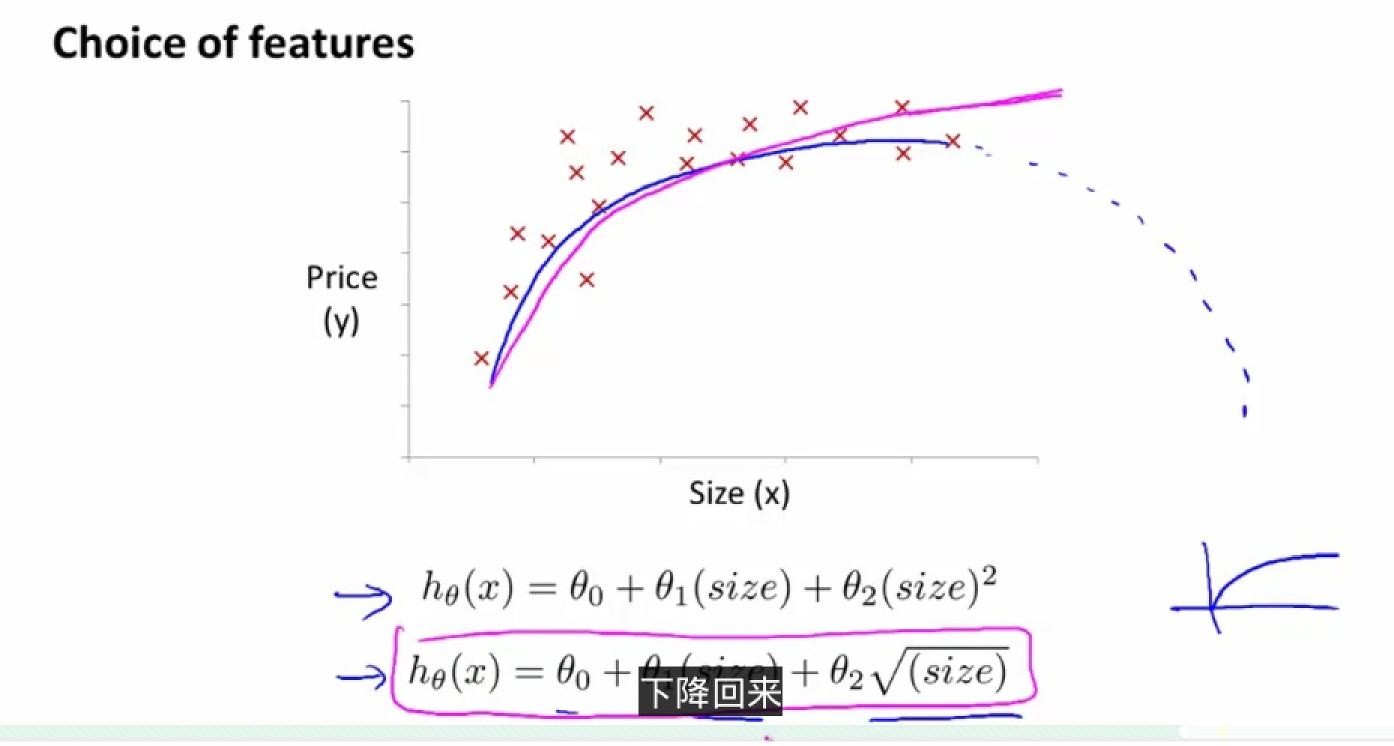
选择合适的特征，比如：有两个特征，长度和宽度，有时候我们可以选择他们的乘积，即面积。有时候定义新的特征可能会得到更好的模型。



多项式回归：（Polynomial Regression）数据关系不是线性。但是可以通过线性回归的简单变换拟合，变换x。但是要注意特征缩放，缩放到相似的范围。



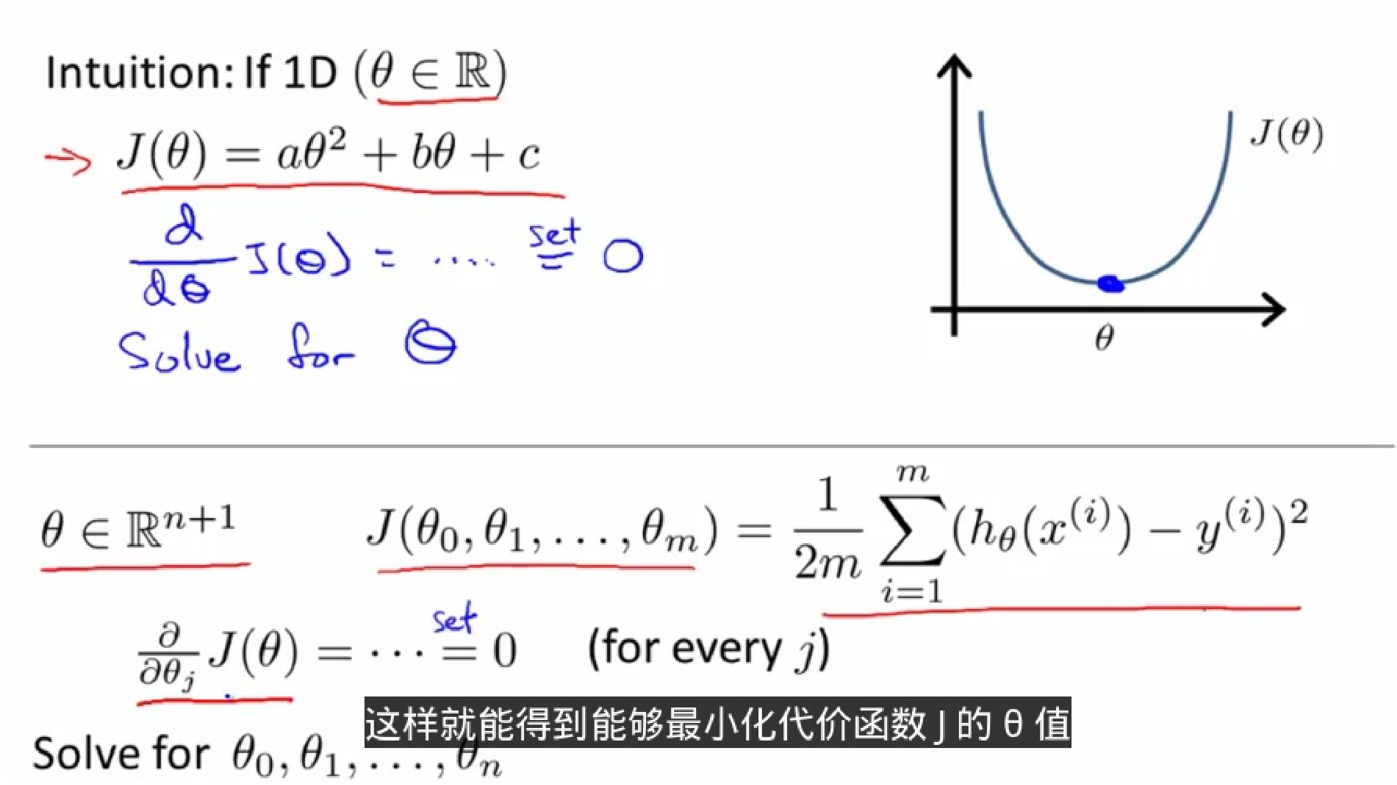
最后，注意模型的选择，以便对数据更好的拟合。（为防止二次函数下降，不用三次函数，用根号）



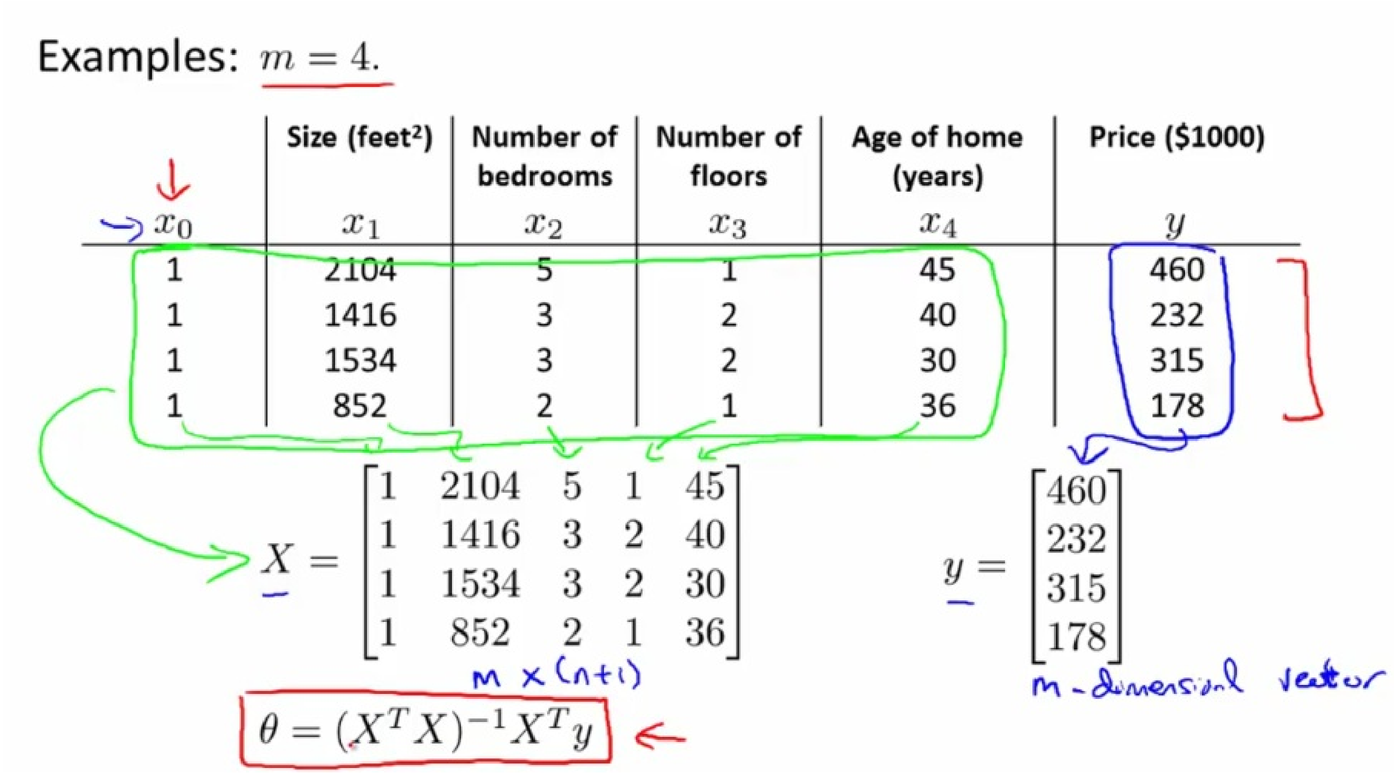
但是，这么多模型要怎么选择呢？

9. 正规方程（Normal Equation）:对于某些情况求解theta更快。不是用迭代，而是用解析方法直接求解出theta的值。

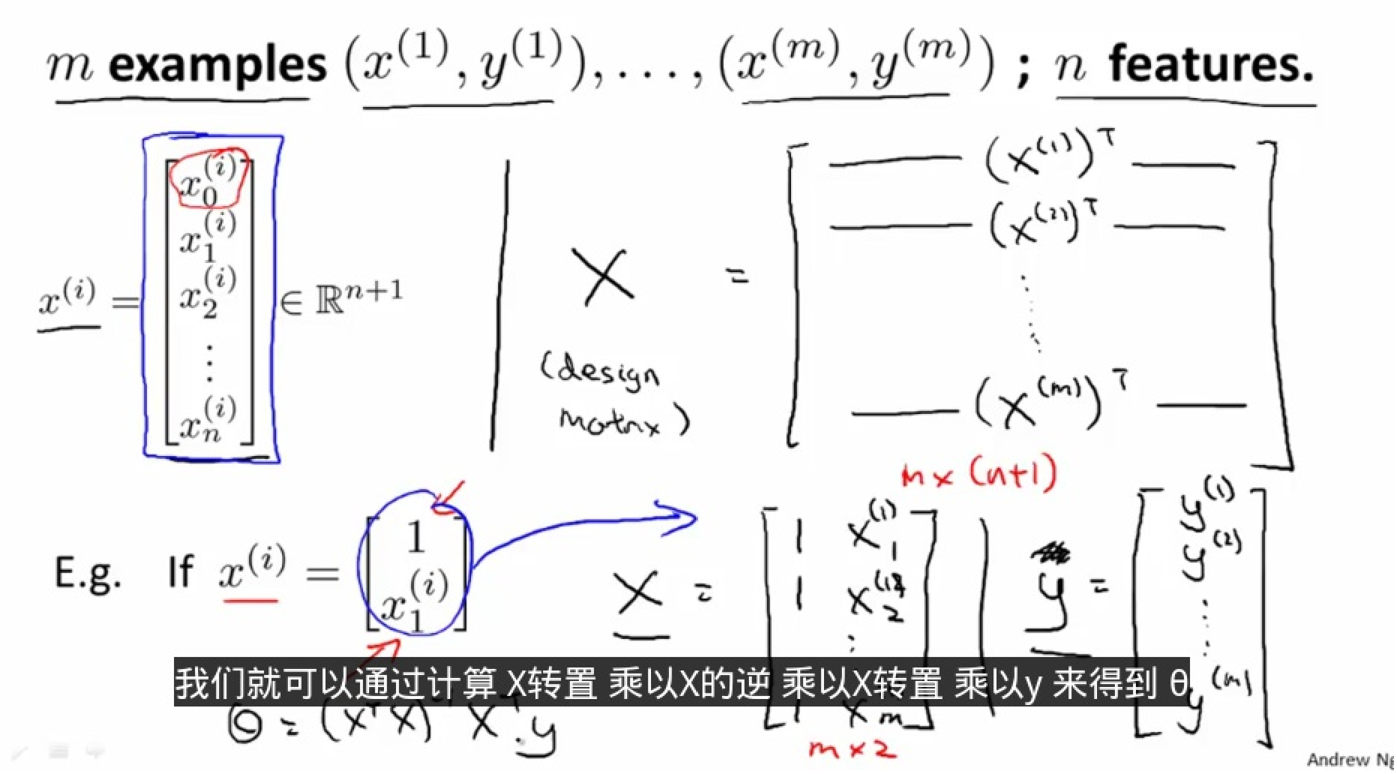
代价函数对每个theta求偏导数，令偏导数为0,然后解出所有的theta值，便得到最小代价函数的theta。



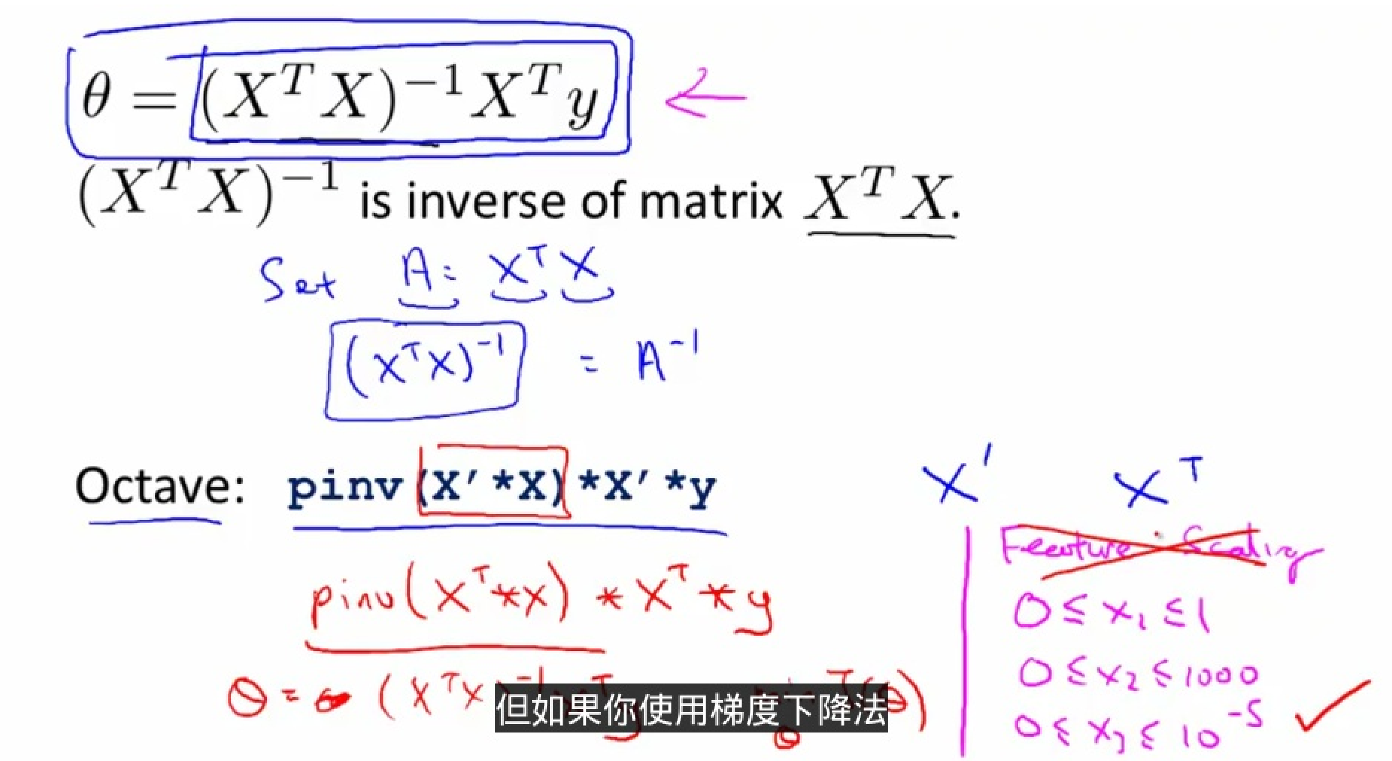
向量方法：



X和y的构建如下：

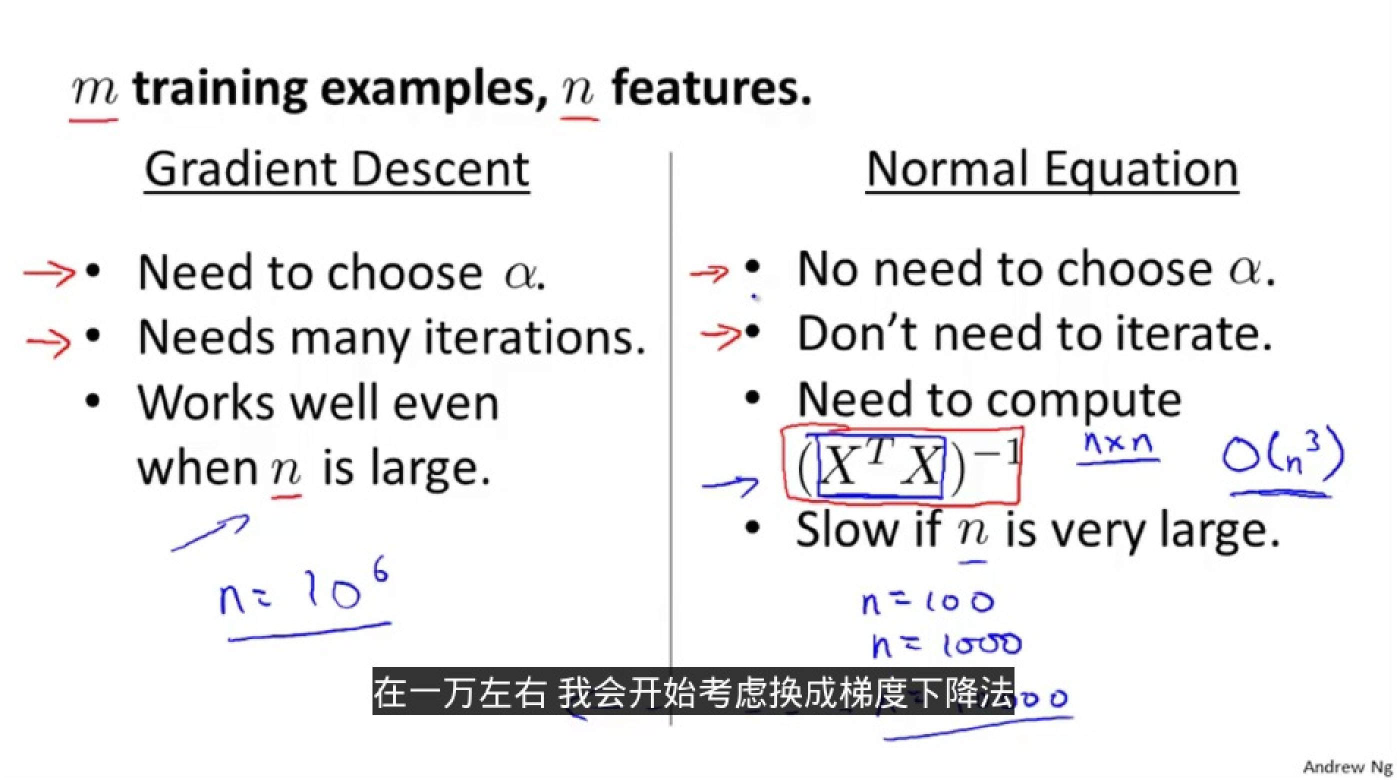


从而得到了最小二乘矩阵矩阵形式：也即正规方程：



同时注意，正规方程方法不需要进行特征缩放。

10. 梯度下降和特征方程方法的比较：



正规方程法公式：theta = inv(X'*X)*X'*y, 即最小二乘法。