特征缩放：

对于方程$y = \theta_0 \ + \ \theta_1 * x_1 \ + \ \theta_2 * x_2$，如果$x_1$的规模远大于$x_2$，那么$x_1$对应的参数$\theta_1$的变化就会比较小；而$x_2$对应的参数$\theta_2$的变化就会比较大。这是因为在更新参数的时候，公式为：

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

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

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

可以看到参数的更新幅度与对应数据的大小呈正比，但是对应的参数本身应该很小才对，这就导致了需要很长时间才可以对这个参数达到收敛的目的（当我们尽可能地将学习率降低）。对应于x_2而言，参数本身比较大，而收敛时候对应的x_2比较小，为使收敛速度增快，应当尽可能增大学习率。此处产生了矛盾，当选择某一个学习率的时候，往往可能出现两难问题，其中一个在等待另外一个收敛。因此进行特征缩放将问题规模在每个尺度上保持一致，使得收敛同步，减少迭代次数。

特征缩放的几种方法：

Rescaling (min-max normalization) [edit]

[illegible]

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

[illegible]

Mean normalization [edit]

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

where x is an original value, x' is the normalized value.

Standardization [edit]

[illegible]

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, $\bar{x}$ is the mean of that feature vector, and σ is its standard deviation.

Scaling to unit length [edit]

[illegible]

$$x' = \frac{x}{||x||}$$

[illegible]