关于标准化和归一化，讲的很好的一篇文章：

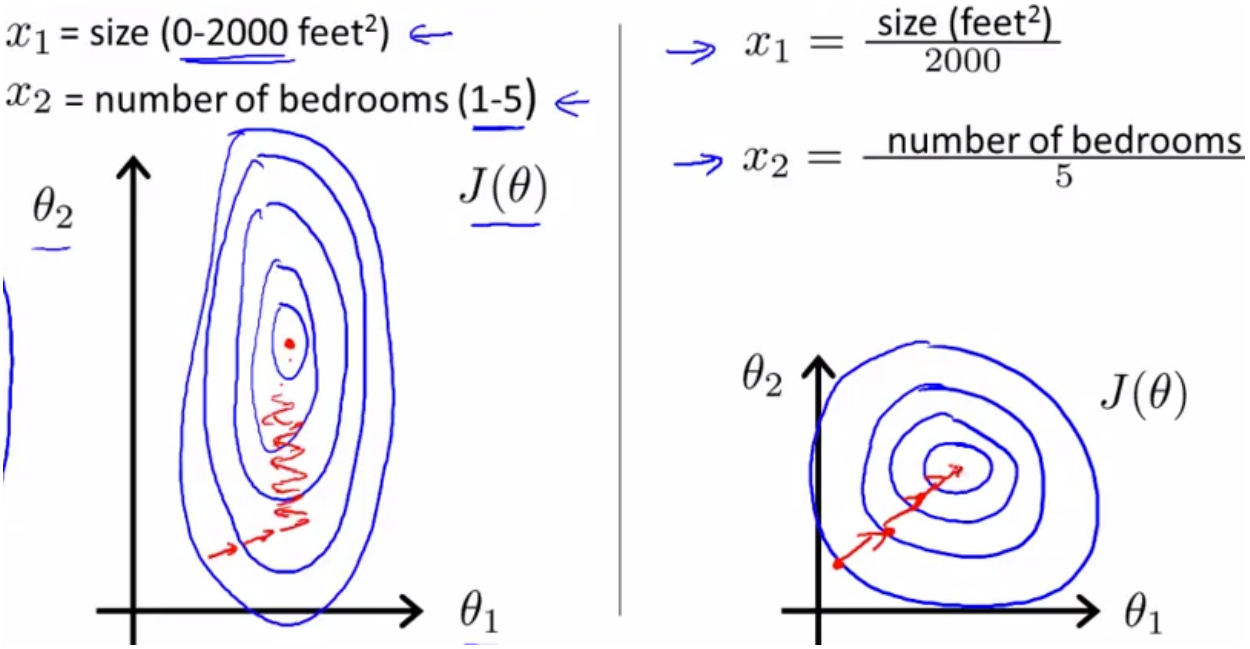
https://baijiahao.baidu.com/s?id=1609320767556598767&wfr=spider&for=pc

1.为何要经常对数据归一化

一般做机器学习应用的时候大部分时间花费在特征处理上，其中关键一步是对特征数据归一化。为什么要归一化，维基百科给出的解释：

（1）归一化后加快了梯度下降求最优解的速度；（2）归一化有可能提高精度

1.1 归一化为什么能提高梯度下降法求最优解的速度



蓝色的圈圈代表的是两个特征的损失等高线。其中左图两个特征x1和x2的区间相差非常大，所形成的等高线非常尖，当使用梯度下降法寻求最优解时，很有可能走“之字型”路线，从而导致需要迭代很多次才能收敛。而右图归一化后，等高线近似圆形，促使SGD往原点迭代，梯度下降能较快收敛。

因此如果机器学习模型使用梯度下降法求最优解时，归一化往往非常必要，否则何难收敛甚至不能收敛。

1.2 归一化有可能提高精度

一些分类器需要计算样本之间的距离（如欧式距离），例如KNN。如果一个特征值域范围非常大，那么距离计算就主要取决于这个特征，从而与实际情况相悖（比如这时实际情况是值域范围小的特征更重要）。

2.归一化的类型

2.1 线性归一化

通过对原始数据进行线性变换把数据映射到[0,1]之间，变换函数为：

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

这种归一化方法比较适用在数值比较集中的情况。缺点：如果max和min不稳定，很容易使得归一化结果不稳定，后续使用也不稳定，实际使用中可以用经验常量值来替代max和min

2.2 标准差标准化

经过处理的数据符合标准正太分布，即均值为0，标准差为1，其转化函数为：

$$x^* = \frac{x - \mu}{\sigma}$$

其中 μ 是样本的均值， σ 是样本的标准差。该种归一化方式要求原始数据的分布可以近似为高斯分布，否标准化的效果会变得很糟糕。它们可以通过现有样本进行估计。在已有样本足够多的情况下比较稳定，适合现代嘈杂大数据场景。

2.3 非线性归一化

经常用在数据分化比较大的场景，有些数值很大，有些很小。通过一些数学函数，将原始值进行映射。该方法包括 log、指数，正切等。需要根据数据分布的情况，决定非线性函数的曲线，比如log(V, 2)还是log(V, 10)等。

3.哪些需要归一化，哪些不需要

实际应用中，通过梯度下降法求解的模型一般需要归一化，比如线性回归、logistic回归、KNN、SVM、神经网络等

树形模型一般不需要归一化，因为他们不关心变量的值，而是关心变量的分布和变量之间的条件概率，如决策树、随机森林。数值缩放不影响分裂点位置，对树模型的结构不造成影响。