特征归一化（数值型变量）：

1）.归一化方法有哪些？

2）.为什么要归一化

3）.那些算法必须要做归一化

1.归一化是将有量纲的表达式，经过线性变换，化为无量纲的表达式

2.目的是为了消除指标之间量纲的影响，以解决数据指标之间的可比性

3.归一化能提高梯度下降最优求解速度，还有可能提高精度

4.归一化常用方法有线性归一化和标准归一化

5.用距离度量相似性和PCA算法中标准归一化更适合，因为进行了方差归一化

1. 定义与目的

归一化是特征工程中常用的数据处理的方法，它的定义是：将有量纲的表达式，经过线性变换，化为无量纲的表达式，成为标量。这样做目的是为了消除指标之间量纲的影响，以解决数据指标之间的可比性。

量纲：有单位的变量就是有量纲，无单位而表示纯数的变量就是无量纲。

举个例子：

比如身高，体重，年龄等，这些都是有单位的变量，它们有个特点，就是两两之间不具可比性，比如8岁和9千克，在数字上9大于8，但是加上单位后就不具可比性 了。这里就产生了一个问题，就是在建立模型的过程往往涉及变量间的度量，比如Kmeans、KNN等需要用距离去度量距离的时候，有量纲的变量显然会影响它们之间的可比性。

因此，归一化的本质就是解决变量间不具可比性的问题，原始数据归一化后各指标处于同一数量级，适合进行综合对比评价。

二.归一化的优点

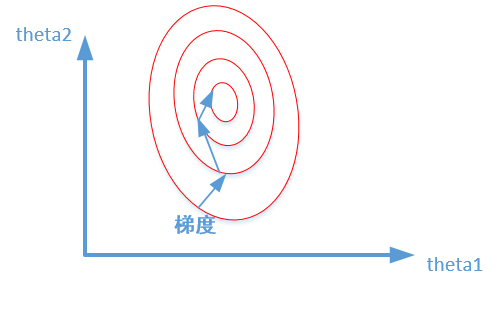
归一化有两个优势：

（1）归一化后加快了梯度下降法求最优解的速度。

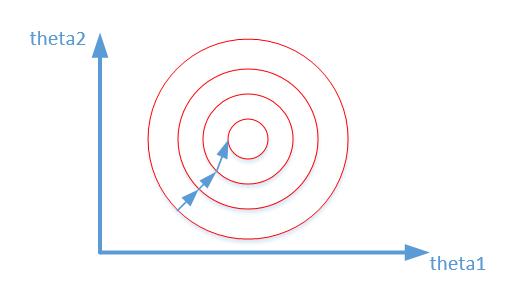
（2）归一化有可能提高精度

如果要更新的两个变量x1,x2(比如x1为房子数量，x2是房间面积）量纲不一样，它们对应的参数是theta1和theta2，那么归一化前后更新过程如下图所示：

未归一化梯度迭代过程



归一化梯度迭代过程



数据归一化后，最优解的寻优过程明显会变得平缓，更容易正确的收敛到最优解。

对于第二点，主要是用在距离度量的时候可以提高精度，比如KNN算法。

三．归一化的常用方法

（1）线性归一化（2）标准归一化

线性归一化就是min-max标准化，它是对数据进行线性变换，映射到固定区间范围。

将数据映射到[0, 1]，转化公式为：(x-min)/(max-min)

将数据映射到[-1,1]，转化公式为：(x-mean)/(max-min)

线性归一化有两个缺陷：

a. 当有新数据加入时，可导致max和min的变化，需要重新定义。

b. 若数据出现较大或较小的异常值，则该标准化方法存在较大的误差（所以先处理异常值再归一化）

标准归一化也称0均值标准化或Z-score标准化，它是通过原始数据的均值和标准差进行标准化，经过处理的数据符合标准正态分布，即均值为0，标准差为1。

（X-Mean）/(Standard deviation)

其中，Mean为所有样本数据的均值。Standard deviation为所有样本数据的标准差。

4.归一化的适用场景

(1)在分类、聚类算法中，需要使用距离来度量相似性的时候或者使用PCA技术进行降维的时候，第二种方法(Z-score)表现更好。

第二种归一化方法对新的数据进行了方差归一化，这时候每个维度的量纲其实已经等价了，每个维度都服从均值为0、方差1的正态分布，在计算距离的时候，每个维度都是去量纲化的，避免了不同量纲的选取对距离计算产生的巨大影响。

(2)在不涉及距离度量、协方差计算、数据不符合正太分布的时候，可以使用第一种方法或其他归一化方法。

需要归一化的算法：

线性回归、逻辑回归、支持向量机、神经网络

不需要归一化的算法：决策树