特征工程中的「归一化」有什么作用?

- 数据缩放的本质是什么
- 不同数据缩放的区别
- 如何选择适合的缩放方法

在这个回答下，我们对一维数据的缩放有如下定义：

- 归一化（normalization）：$\frac{X_i - X_{min}}{X_{max} - X_{min}}$
- 标准化（standardization）：$\frac{X_i - \mu}{\sigma}$

其中 μ 和 σ 代表样本的均值和标准差，X_{max} 为最大值，X_{min} 为最小值。

1. 归一化和标准化本质上都是一种线性变换

先看归一化，在数据给定的前提下，令常数 $\alpha = X_{max} - X_{min}$，常数 $\beta = X_{min}$，那么归一化的新的形式就是 $\frac{X_i - \beta}{\alpha}$。在这种改写后下，易发现和标准化形式 $\frac{X_i - \mu}{\sigma}$ 类似，因为在数据给定后 μ 和 σ 也可看做常数。

因此可以再稍微变形一下：$\frac{X_i - \beta}{\alpha} = \frac{X_i}{\alpha} - \frac{\beta}{\alpha} = \frac{X_i}{\alpha} - c$（*公式 1*）

就发现事实上就是对向量 X **按照比例压缩** α **再进行平移** c。所以**归一化和标准化的本质就是一种线性变换。**

举个简单的例子：

- 原始数据：$X = [1, 2, 5, 3, 4]$，其中 $\alpha = X_{max} - X_{min} = 4$，$\beta = X_{min} = 1$, $c = \frac{\beta}{\alpha} = \frac{1}{4}$
- 归一化：代入*公式 1*，将 X 压缩 4 倍并平移 $\frac{1}{4}$，得到 $\frac{1}{\alpha}X - c = [\frac{1}{4}, \frac{2}{4}, \frac{5}{4}, \frac{3}{4}, \frac{4}{4}] - \frac{1}{4}$，最终有 $[0, \frac{1}{4}, 1, \frac{2}{4}, \frac{3}{4}]$
- 标准化：与归一化类似，略

2. 线性变化的性质

线性变换有很多良好的性质，这些性质决定了为什么对数据进行改变后竟然不会造成 "失效"，反而还能提高数据的表现。拿其中很重要的一个性质为例，**线性变化不改变原始数据的数值排序。**

感兴趣的朋友可以试试下面的代码，就会发现这两种处理方法都不会改变数据的排序。对于很多模型来说，这个性质保证了数据依然有意义，顺序性不变，而不会造成了额外的影响。

```
原始顺序:[1.2.6.5.3.4.]
标准化顺序:[1.2.6.5.3.4.]
归一化顺序:[1.2.6.5.3.4.]
```

```python
from sklearn import preprocessing
from scipy.stats import rankdata

x = [[1], [3], [34], [21], [10], [12]]
std_x = preprocessing.StandardScaler().fit_transform(x)
print('原始顺序  : ', rankdata(x))
print('标准化顺序: ', rankdata(std_x))
print('归一化顺序: ', rankdata(norm_x))
```

说白了，只是因为线性变换保持线性组合与线性关系式不变，这保证了特定模型不会失效.

3. 归一化和标准化的区别

我们已经说明了它们的本质是缩放和平移，但区别是什么呢？在不涉及线性代数的前提下，我们给出一些直觉的解释：**归一化的缩放是“拍扁”统一到区间（仅由极值决定），而标准化的缩放是更加“弹性”和“动态”的，和整体样本的分布有很大的关系。**值得注意：

- 归一化：缩放仅仅跟最大、最小值的差别有关。
- 标准化：缩放和每个点都有关系，通过方差（variance）体现出来。与归一化对比，标准化中所有数据点都有贡献（通过均值和标准差造成影响）。

当数据较为集中时，α 更小，于是数据在标准化后就会更加分散。如果数据本身分布很广，那么 α 较大，数据就会被集中到更小的范围内。

从输出范围角度来看，$\frac{X_i - X_{min}}{X_{max} - X_{min}}$ 必须在 0-1 间。对比来看，显然 $\sigma \leq X_{max} - X_{min}$，甚至在极端情况下 $\sigma = 0$，所以标准化的输出范围一定比归一化更广。

- 归一化： 输出范围在 0-1 之间
- 标准化：输出范围是负无穷到正无穷

4. 什么时候用归一化？什么时候用标准化？

我们已经从第三部分得到了一些性质，因此可以得到以下结论：

- 如果对**输出结果范围有要求**，用归一化
- 如果**数据较为稳定，不存在极端的最大最小值**，用归一化
- 如果数据**存在异常值和较多噪音**，用标准化，可以间接通过中心化避免异常值和极端值的影响

一般来说，我个人建议优先使用标准化。在对输出有要求时再尝试别的方法，如归一化或者更加复杂的方法。很多方法都可以将输出调整到 0-1，如果我们对于数据的分布有假设的话，更加有效方法是使用相对应的概率密度函数来转换。让我们以高斯分布为例，我们可以首先计算高斯误差函数（Gaussian Error Function），此处定为 $erfc(\cdot)$，那么可用下式进行转化：

$$max\left\{0, erfc\left(\frac{x-\mu}{\sigma \cdot \sqrt{2}}\right)\right\}$$

具体讨论可参考我的文章 机器学习「输出概率化」：一种无监督的方法。

为什么要进行归一化处理，下面从**寻找最优解**这个角度给出自己的看法。

例子

假定为预测房价的例子，自变量为面积，房间数两个，因变量为房价。

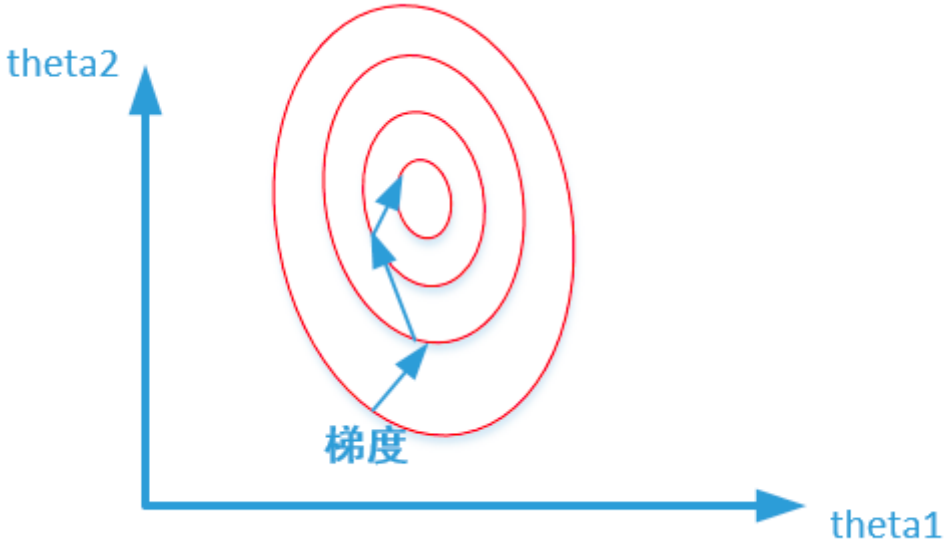
那么可以得到的公式为：

$$y = \theta_1 x_1 + \theta_2 x_2$$

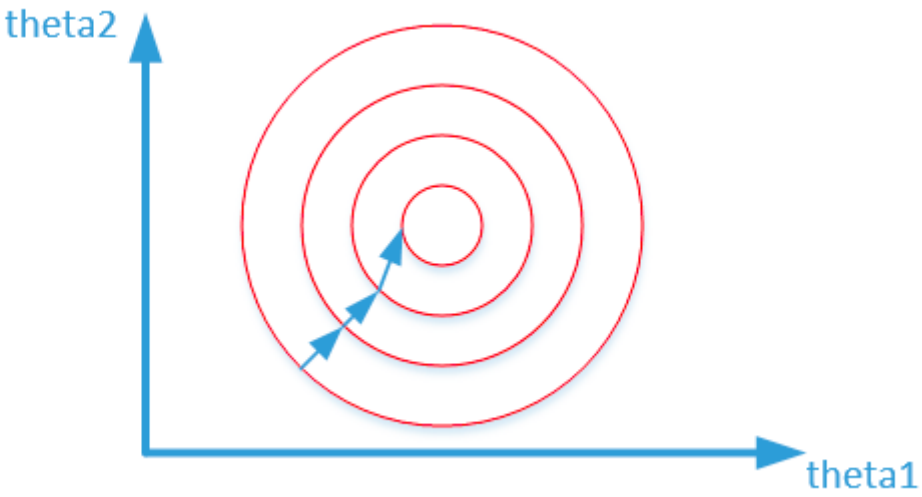
其中x1代表房间数,$\theta 1$代表x1变量前面的系数。
其中x2代表面积,$\theta 2$代表x2变量前面的系数。

首先我们祭出两张图代表数据是否均一化的最优解寻解过程。

未归一化：



归一化之后



为什么会出现上述两个图，并且它们分别代表什么意思。

我们在寻找最优解的过程也就是在使得损失函数值最小的 theta1,theta2。

上述两幅图代码的是损失函数的等高线。

我们很容易看出，当数据没有归一化的时候，面积数的范围可以从 0~1000，房间数的范围一般为 0~10，可以看出面积数的取值范围远大于房间数。

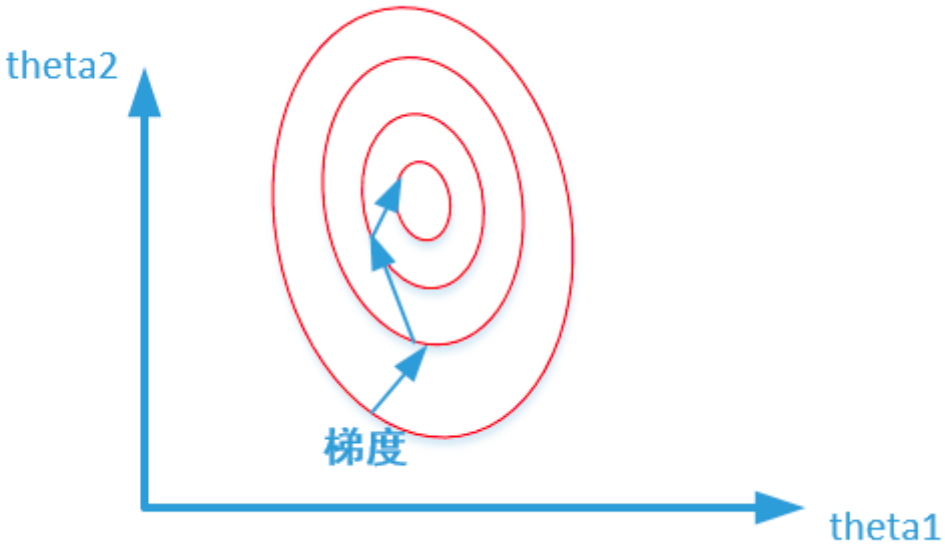
影响

这样造成的影响就是在画损失函数的时候，

数据没有归一化的表达式，可以为：

$$J = (3\theta_1 + 600\theta_2 - y_{correct})^2$$

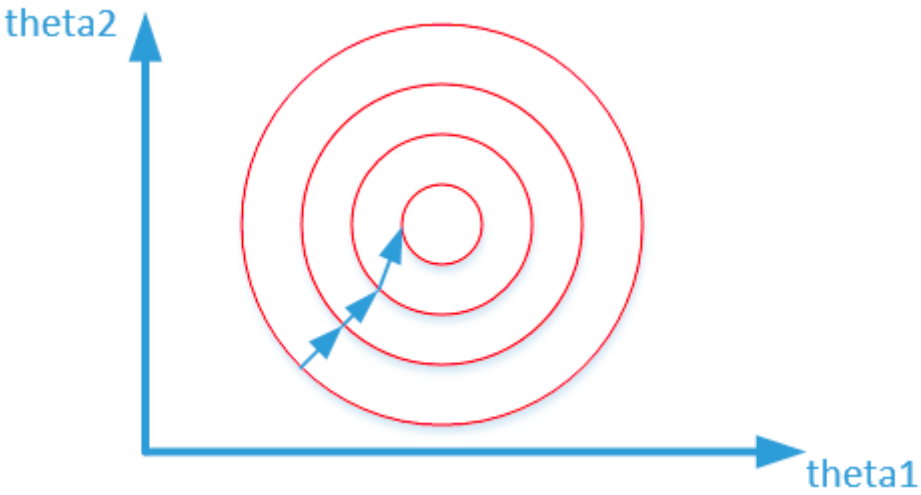
造成图像的等高线为类似椭圆形状，最优解的寻优过程就是像下图所示：



而数据归一化之后，损失函数的表达式可以表示为：

$$J = (0.5\theta_1 + 0.55\theta_2 - y_{correct})^2$$

其中变量的前面系数几乎一样，则图像的等高线为类似圆形形状，最优解的寻优过程像下图所示：



从上可以看出，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。**

这也是数据为什么要归一化的一个原因。

在进行数据分析的时候，什么情况下需要对数据进行标准化处理？

主要看模型是否具有伸缩不变性。

有些模型在各个维度进行不均匀伸缩后，最优解与原来不等价，例如SVM。对于这样的模型，除非本来各维数据的分布范围就比较接近，否则**必须**进行标准化，以免模型参数被分布范围较大或较小的数据dominate。

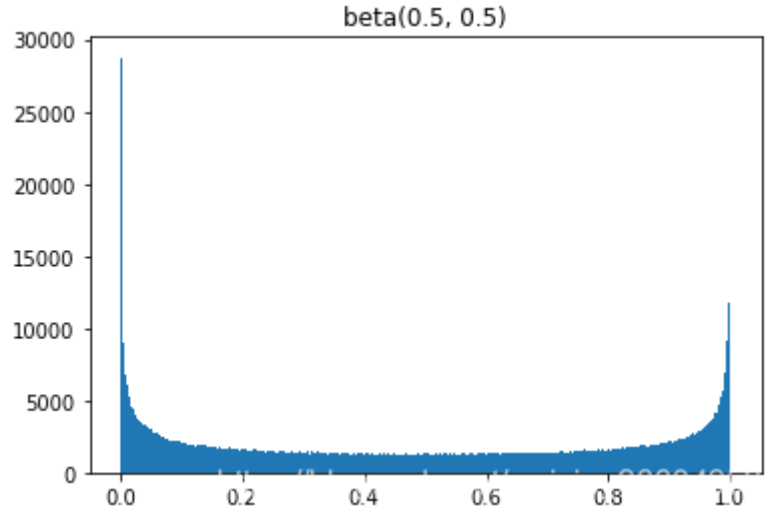
有些模型在各个维度进行不均匀伸缩后，最优解与原来等价，例如logistic regression。对于这样的模型，是否标准化理论上不会改变最优解。但是，由于实际求解往往使用迭代算法，如果目标函数的形状太"扁"，迭代算法可能收敛得很慢甚至不收敛。所以对于具有伸缩不变性的模型，**最好**也进行数据标准化。

1、标准化 (Standardization) 和归一化 (Normalization) 概念

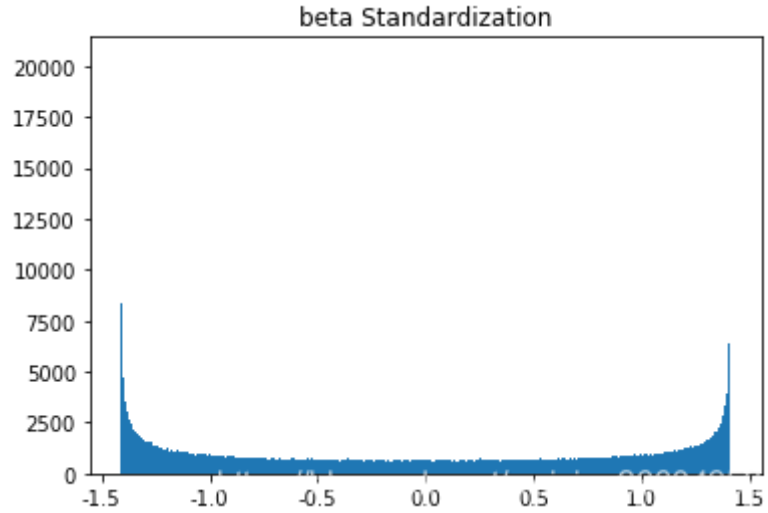
1.1、定义

归一化和标准化都是对数据做变换的方式，将原始的一列数据转换到某个范围，或者某种形态，具体的：

很多博客甚至书中说， Standardization 是改变数据分布，将其变换为服从 N(0,1)的标准正态分布，这点是错的，Standardization 会改变数据的均值、标准差都变了 (当然，严格的说，均值和标准差变了，分布也是变了，但分布种类依然没变，原来是啥类型，现在就是啥类型)，但本质上的分布并不一定是标准正态，完全取决于原始数据是什么分布。我个举例子，我生成了 100 万个服从 $beta(0.5, 0.5)$的样本点 (你可以替换成任意非正态分布，比如卡方等等，beta(1,1) 是一个服从 U(0,1) 的均匀分布，所以我选了$beta(0.5, 0.5)$ ，称这个原始数据为 b_0，分布如下图所示：

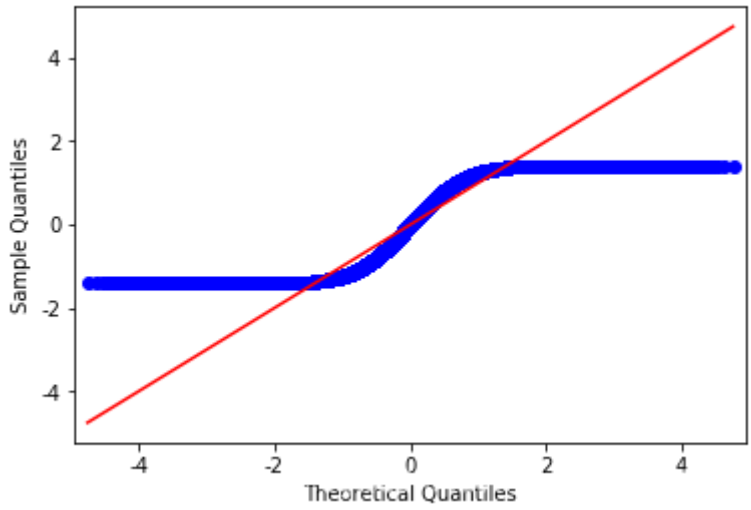


通过计算机计算，样本b_0的均值和方差分别为 0.49982 和 0.12497(约为 0.5 和 0.125)
对这个数据做 Standardization，称这个标准化后的数据为b_1，分布如下：



可以看到数据形态完全不是正态分布，但是数学期望和方差已经变了。beta 分布的数学期望为$\frac{a}{a+b}$，方

差为 $\frac{ab}{(a+b)^2(a+b+1)}$，所以$E(b_0)=\frac{0.5}{0.5+0.5}=\frac{1}{2}$，$Var(b_0)=\frac{1}{8}$，这也和我们上文所计算的样本均值和方差一致，而 b_1的均值和方差分别为：-1.184190523417783e-1 和 1，均值和方差已经不再是 0.5 和 0.125，分布改变，但绝不是一个正态分布，你不信的话，觉得看分布图不实锤，通过 qq 图和检验得到的结果如下：




```
KstestResult(statistic=0.09699696253991646, pvalue=0.0)
```

1.2、联系和差异

一、联系

Standardization 和 Normalization 本质上都是对数据的线性变换，广义的说，你甚至可以认为他们是同一个母亲生下的双胞胎，为何而言，因为二者都是不会改变原始数据排列顺序的线性变换：
假设原始数据为 X X X，令 $\alpha=X_{max}-X_{min}$，令$\beta=X_{min}$(很明显，数据给定后 α、β就是常数)，则 $X_{Normalization}=\frac{X_i-\beta}{\alpha}=\frac{X_i}{\alpha}-\frac{\beta}{\alpha}=\frac{X_i}{\alpha}-c$，可见， Normalization 是一个线性变换，按 α进行缩放，然后平移 c 个单位。其实 $\frac{X_i-\beta}{\alpha}$中的β和α就像是 Standardization 中的μ和 σ(数据给定后，μ和 σ也是常数)。线性变换，必不改变原始的排位顺序。

二、差异

1. 第一点：显而易见， Normalization 会严格的限定变换后数据的范围，比如按之前最大最小值处理的 Normalization，它的范围严格在[0,1] 之间；
   而 Standardization 就没有严格的区间，变换后的数据没有范围，只是其均值是 0，标准差为 1。
2. 第二点：归一化 (Normalization) 对数据的缩放比例仅仅和极值有关，就是说比如 100 个数，你除去极大值和极小值其他数据都更换掉，缩放比例 $\alpha=X_{max}-X_{min}$是不变的；反观，对于标准化 (Standardization) 而言，它的$\alpha=\sigma$， $\beta=\mu$，如果除去极大值和极小值其他数据都更换掉，那么均值和标准差大概率会改变，这时候，缩放比例自然也改变了。

2、标准化、归一化的原因、用途

为何统计模型、机器学习和深度学习任务中经常涉及到数据 (特征) 的标准化和归一化呢，我个人总结主要有以下几点，当然可能还有一些其他的作用，大家见解不同，我说的这些是通常情况下的原因和用途。

1. 统计建模中，如回归模型，自变量 X X X的量纲不一致导致了回归系数无法直接解读或者错误解读；需要将 X X X 都处理到统一量纲下，这样才可比；
2. 机器学习任务和统计学任务中有很多地方要用到 “距离” 的计算，比如 PCA，比如 KNN，比如 kmeans 等等，假使算欧式距离，不同维度量纲不同可能会导致距离的计算依赖于量纲较大的那些特征而得到不合理的结果；
3. 参数估计时使用梯度下降，在使用梯度下降的方法求解最优化问题时， 归一化 / 标准化后可以加快梯度下降的求解速度，即提升模型的收敛速度。

3、什么时候 Standardization，什么时候 Normalization

我个人理解：如果你对处理后的数据范围有严格要求，那肯定是归一化，个人经验，标准化是 ML 中更通用的手段，如果你无从下手，可以直接使用标准化；如果数据不为稳定，存在极端的最大最小值，不要用归一化。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，标准化表现更好；在不涉及距离度量、协方差计算的时候，可以使用归一化方法。

4、所有情况都应当 Standardization 或 Normalization 么

当原始数据不同维度特征的尺度 (量纲) 不一致时，需要标准化步骤对数据进行标准化或归一化处理，反之则不需要进行数据标准化。也不是所有的模型都需要做归一的，比如模型算法里面有没关于对距离的衡量，没有关于对变量间标准差的衡量。比如决策树，他采用算法里面没有涉及到任何和距离等有关的，所以在做决策树模型时，通常是不需要将变量做标准化的；另外，概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率。