数据变换——标准化变换

标准化，又称规范化，目的是将原来的度量值转换为无量纲的值，使得不同量纲的指标可以在同一水平线上进行比较，而且除了概率模型（树模型）之外，其他模型如神经网络、最邻近分类和聚类算法等，都需要先对数据进行标准化，以消除量纲，缩放数据，加快算法的收敛速度。

MATLAB 提供了 normalize 函数对数据进行标准化，语法格式如下：

N = normalize(A, dim, method, methodtype)

A: 输入数据，指定为标量、向量、矩阵、多维数组、表或时间表。

dim: 运算维度，默认为 1，按列进行标准化； 设为 2 则按行进行标准化。

method: 字符型参数，默认为‘zscore’法，具体选项见下表：

方法	说明
‘zscore’	z-score法标准化，将数据转化为均值为0，方差为1的z值
‘norm’	将每个元素除以所在列的2-范数
‘scale’	按标准差缩放，每个元素除以所在列的标准差
‘range’	min-max标准化，将数据范围缩放到[0,1]
‘center’	中心化，每个元素减去所在列的均值

连续属性离散化：也即离差标准化，公式如下：

$$x^* = \frac{x - min}{max - min}$$

其中，m a x maxmax 和 m i n minmin 为样本数据的最大值和最小值。该方法保留了原始数据中存在的关系，是消除量纲和数据取值范围影响的最简单的方法，缺点是若数据值集中或某个数值很大，规范化后各值会接近 0 且相差不大。

零-均值规范化： 即经过处理后均值为 0，标准差为 1，公式如下：

$$x^* = \frac{x - \overline{x}}{\delta}$$

该方法目前使用广泛，不过均值和标准差受离群点影响比较大，因此通常需要修改上述变换，比如用中位数M取代均值，用绝对标准差$\delta^* = \sum_{i=1}^{n} |x_i - W|$，其中$W$为平均数或中位数。

小数定标规范化：通过移动属性值的小数位数，将属性值映射到［-1，1］之间，移动的小数位数取决于属性值绝对值的最大值。转化公式为：

$$x^* = \frac{x}{10^k}$$

methodtype：字符型参数，方法类型，为上一个参数“method”指定更加具体的方法类型，具体选项见下表：

方法	方法类型或选项	说明
‘zscore’	‘std’（默认值）	中心化并缩放，使之均值为0，标准差为1
	‘robust’	中心化并缩放，使之中位数为0，中位数绝对偏差为1
‘norm’	正整数（默认为2）	p-范数
	Inf	无穷范数
‘scale’	‘std’（默认值）	按标准差缩放
	‘mad’	按中位数绝对偏差缩放
	‘first’	按数据的第一个元素进行缩放
	数值标量	按数值缩放数据
‘range’	二元素行向量（默认为[0 1]）	[a b]形式的区间，其中a<< span=""> b
‘center’	‘mean’	中心化以使其均值为0
	‘median’	中心化以使其中位数为0
	数值标量	按数值平移中心