为了更加方便得处理数值型特征,往往需要

- 归一化,也就是做平移和缩放使得取值范围变成 `[0, 1]`
- 中心化,也就是做线性平移使得均值为0,方差不变
- 标准化,也就是
  1. 做平移和缩放,使得均值为0,方差为1,或者
  2. 做缩放使得向量的范数为1,

这样做的目的是为了使得数据更加贴合模型假设和减少设计矩阵(每行表示一个样例,每列一维特征)的条件数(进一步避免zigzag pa).

最大绝对值归一化

在已知特征最大最小值的情况下通过缩放除去量纲放到 `[-1, 1]` 区间

$$X' = \frac{X}{maxAbs(X)}$$

其中$maxAbs(X)$指的是取值的绝对值的最大值

min-max归一化

在已知特征最大最小值的情况下通过缩放除去量纲并缩放到 `[0, 1]` 区间.

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

Z-score

Z-score又叫做stardard score, normal score.借鉴了正态分布转化为标准正态分布的思想,使得特征的统计量变为均值为0,标准差为1.

$$X' = \frac{X - \mu}{\sigma}$$

其中μ为均值,σ为标准差

使用sklearn对特征进行归一化

sklearn中的 `sklearn.preprocessing` 提供了多个归一化函数用于对特征进行归一化,主要有:

- normalize 根据某种范数来标准化,范数可以为L-1, L-2和最大范数,支持稀疏矩阵
- minmax_scale(X, feature_range=(0, 1), axis=0, copy=True) min-max归一化,不支持稀疏矩阵
- maxabs_scale(X, axis=0, copy=True) 最大绝对值归一化,支持稀疏矩阵
- StandardScaler(copy=True, with_mean=True, with_std=True) Z-score标准化,不支持稀疏矩阵,因为这样稀疏矩阵就不稀疏了

In [ ]: