Scaling

Allen.Huang

sklearn.preprocessing

在scikit-learn中，有如下几种常用的数据预处理方法

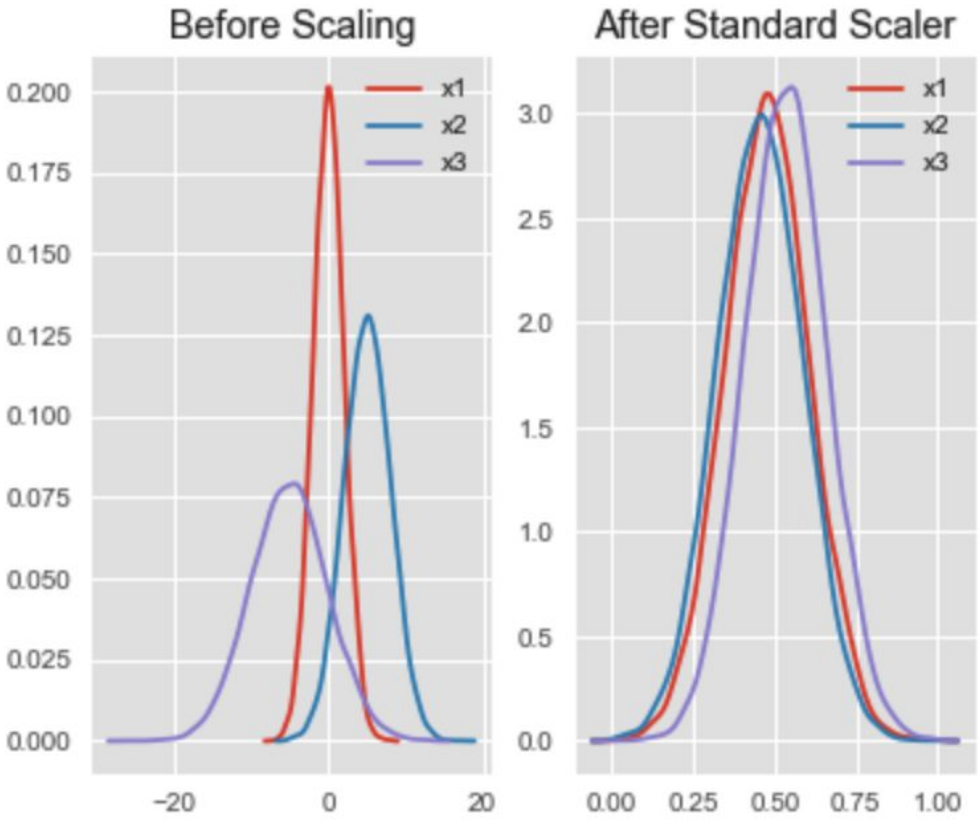
- StandardScaler
- MinMaxScaler
- RobustScaler
- Normalizer

StandardScaler

StandardScaler-标准化处理，目的是为了将数据的均值变成0，将方差变成1

$$\frac{x_i - mean(x)}{stdev(x)}$$

处理前后的对照


Before Scaling
After Standard Scaler
x1
x2
x3
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−20
0
20
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00

Min-Max Scaler

该变换器可以将数据变到(0,1)之间，或者是(-1,1)之间

优点：

1.在标准化处理效果不佳的时候可以考虑使用

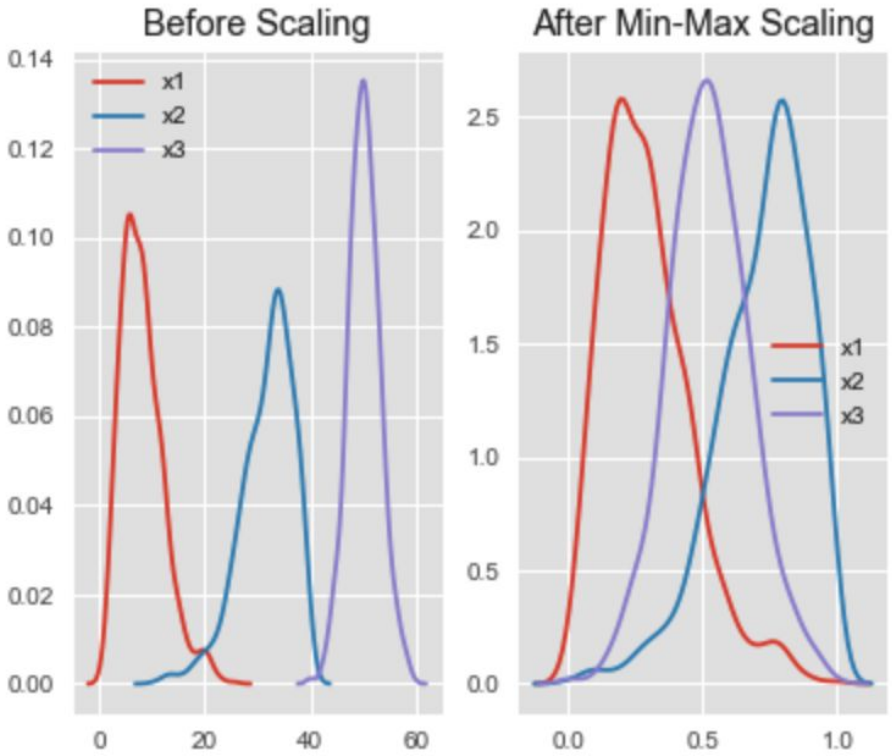
2.非高斯或者方差小的时候，效果比较好

$$\frac{x_i - min(x)}{max(x) - min(x)}$$

缺点：

对于离散点非常敏感，这时可以考虑使用RobustScaler

变换前后


Before Scaling
After Min-Max Scaling
x1
x2
x3
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0

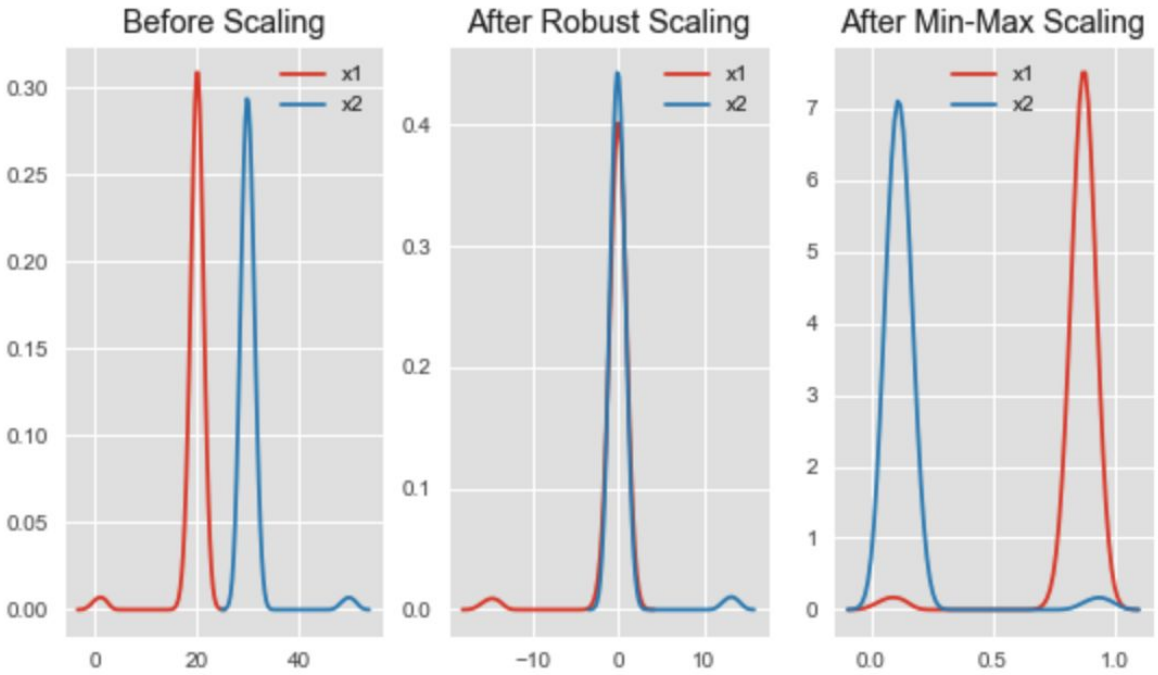
Robust Scaler

Robust类似于Min-Max，然而它没有使用最大和最小值，而是使用了一分位和三分位的点作为变换参数

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$$

该方法在有离群点的时候可以加以考虑

变换前后


Before Scaling
After Robust Scaling
After Min-Max Scaling
x1
x2
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
0.4
0.3
0.2
0.1
0.0
−10
0
10
7
6
5
4
3
2
1
0
0.0
0.5
1.0

Normalize

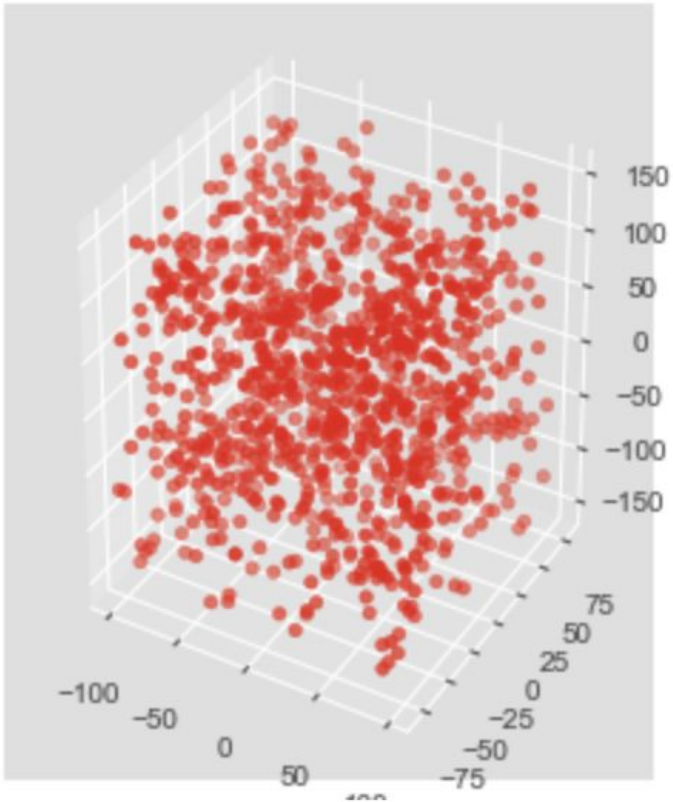
归一化操作，将所有的X变量拉入到一个球中

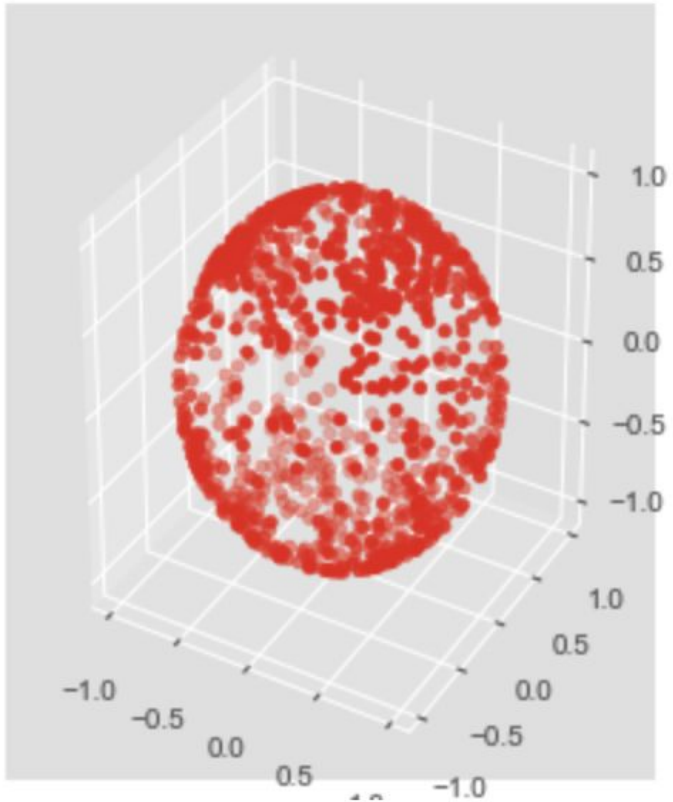
$$\frac{x_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}}$$

需要注意的是，在归一化的时候对于数值的范围很敏感

大范围的数值会将小范围的数值进行淡化

变换前后


150
100
50
0
−50
−100
−150
−100
−50
0
50
75
50
25
0
−25
−50
−75


1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
0.5
0.0
−0.5
−1.0