CODE PILLS

Data scaling

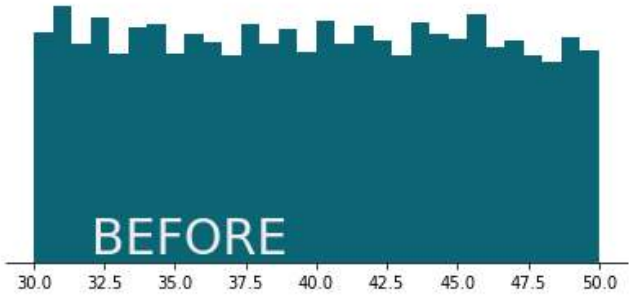
Jul 25, 2019

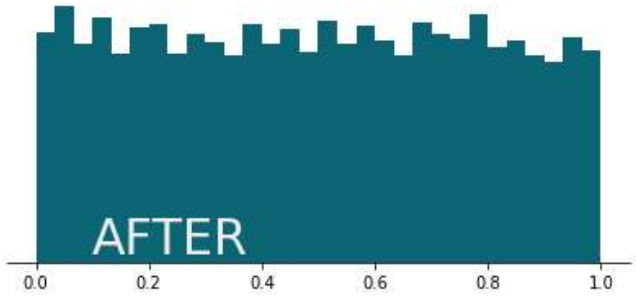
Normalization

Scale all the values between two extremes

```
X = np.random.uniform(low=30, high=50, size=[10000, 1])

from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
scaler.fit(X)
X_scaled = scaler.transform(X)
```



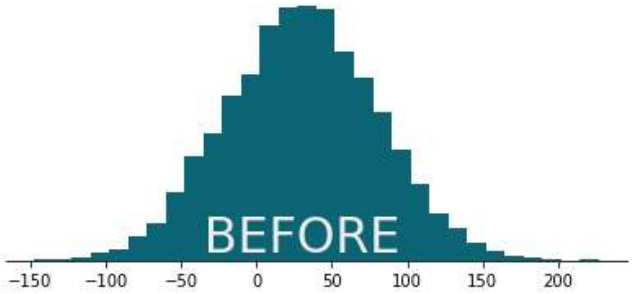


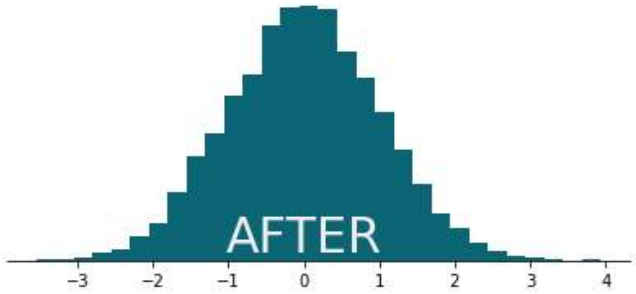
Standardization

Scale all the values to have zero means and unitary standard deviation

```
X = np.random.normal(loc=30, scale=50, size=[10000, 1])

from sklearn import preprocessing
scaler = skl.preprocessing.StandardScaler()
scaler.fit(X)
X_scaled = scaler.transform(X)
```



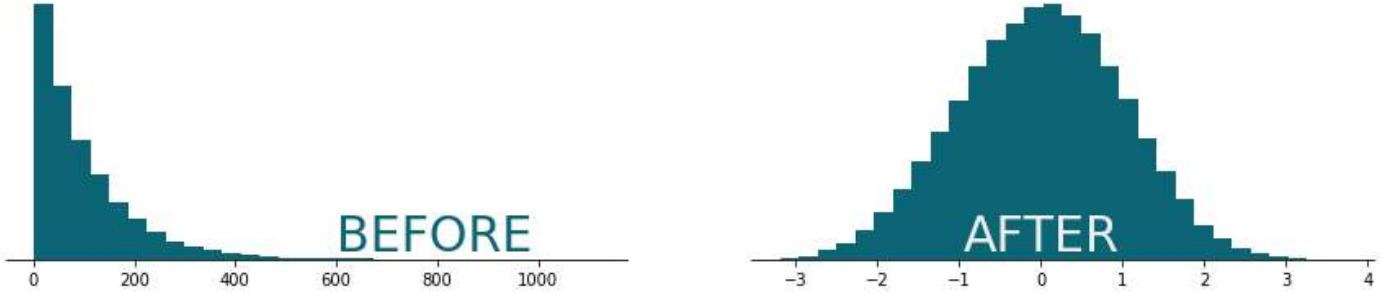


Box-cox transformation

Gaussianizes a distribution and standardizes it. It works only with positive values.

```
X = np.random.exponential(scale=100, size=[10000, 1])

from sklearn import preprocessing
scaler_box = preprocessing.PowerTransformer(method='box-cox', standardize=True)
scaler_box.fit(X)
X_scaled_box = scaler_box.transform(X)
```

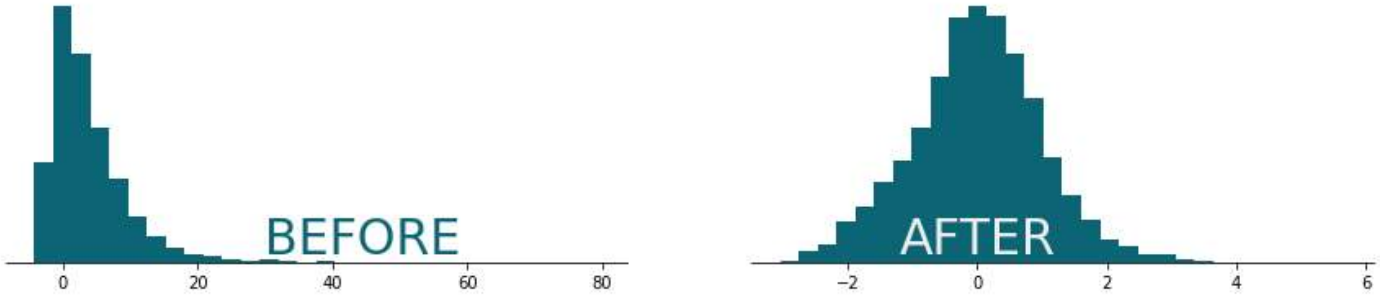


Yeo-Johnson transformation

Similar effect to the *Box-cox transformation* but can be used for negative values as well.

```
SHIFT = 5
X = np.random.lognormal(mean=2, sigma=0.6, size=[10000, 1]) - SHIFT

from sklearn import preprocessing
scaler = preprocessing.PowerTransformer(method='yeo-johnson', standardize=True)
scaler.fit(X)
X_scaled = scaler.transform(X)
```



Code for plots

```
import matplotlib.pyplot as plt
from sklearn import preprocessing

X = np.random.normal(loc=30, scale=50, size=[10000, 1])
scaler = skl.preprocessing.StandardScaler()
scaler.fit(X)
X_scaled = scaler.transform(X)

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15,3))
for ax in (ax1, ax2):
    ax.set_yticks([])
    ax.set_yticks([])
    ax.spines['left'].set_visible(False)
    ax.spines['top'].set_visible(False)
    ax.spines['right'].set_visible(False)

ax1.text(30, 0, 'BEFORE', size=30, color=colors[8], ha='center', va='bottom')
ax2.text(0, 0, 'AFTER', size=30, color=colors[8], ha='center', va='bottom')
ax1.hist(X, bins=30, density=True, color=colors[0])
```

```
ax2.hist(X_scaled, bins=30, density=True, color=colors[0]);
# fig.savefig('./plots/standardization.png',  bbox_inches='tight')
```

Code pills

Code pills
amico.andrea.90@gmail.com

AndreaAmico

I'm just too lazy to use dropbox or GitHub to copy-paste code for everyday science scripts. Need to fit some cute data points? Need to analyze your favorite images? Need 20 lines of code to write a trash GUI you will use for a couple of weeks? If you want to steal something, feel free to copy-paste at your own risk, you will probably find a forest of bugs.