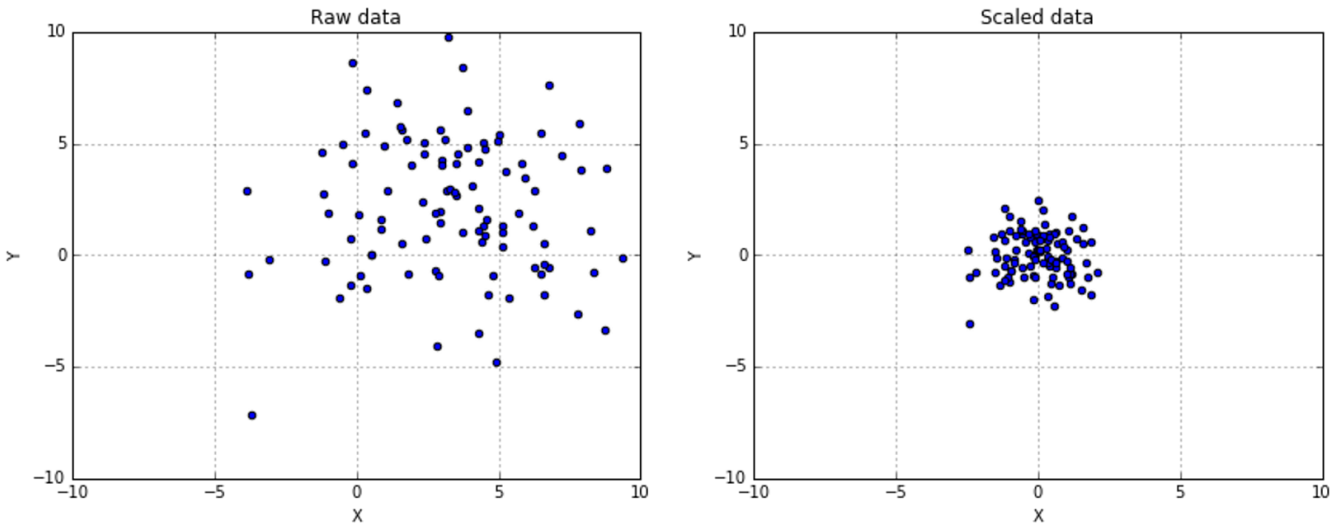
Data scaling and normalization

A generic dataset (we assume here that it is always numerical) is made up of different values which can be drawn from different distributions, having different scales and, sometimes, there are also outliers. A machine learning algorithm isn't naturally able to distinguish among these various situations, and therefore, it's always preferable to standardize datasets before processing them. A very common problem derives from having a non-zero mean and a variance greater than one. In the following figure, there's a comparison between a raw dataset and the same dataset scaled and centered:



This result can be achieved using the `StandardScaler` class:

Copy

```
from sklearn.preprocessing import StandardScaler

>>> ss = StandardScaler()
>>> scaled_data = ss.fit_transform(data)
```

It's possible to specify if the scaling process must include both mean and standard deviation using the parameters `with_mean=True/False` and `with_std=True/False` (by default they're both active...

‹ Previous Section (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1s

Next Section › (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1sec2