**Feature Scaling**

* **Feature scaling** refers to the methods used to normalize the range of values of independent variables.
* In other words, the ways to set the feature value range within a similar scale.

**Why Feature Scaling Matters**

Feature magnitude matters for several reasons:

* The scale of the variable directly influences the regression coefficient.
* Variables with a more significant magnitude dominate over the ones with a smaller magnitude range.
* Gradient descent converges faster when features are on similar scales.
* Feature scaling helps decrease the time to find support vectors for SVMs.
* Euclidean distances are sensitive to feature magnitude.

## Algorithms sensitive to feature magnitude

* Linear and logistic regression
* Neural networks
* Support vector machines
* KNN
* K-means clustering
* Linear discriminant analysis (LDA)
* Principal component analysis (PCA)

## Algorithms intensive to feature magnitude

* Classification and regression trees
* Random forests
* Gradient boosted trees

**Scaling methods**

Next, let’s take a closer look at the following feature scaling methods:

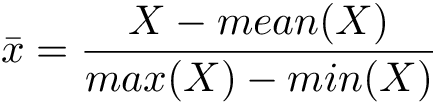
* Mean normalization
* Standardization
* Robust scaling (scaling to median and IQR)
* Scale to maximum and minimum
* Scale to the absolute maximum
* Scale to unit norm

**Mean Normalization**

Mean normalization suggests centering the variable at 0 and re-scaling the variable’s **value range** tothe range of -1 to 1.

This technique will not **normalize the variable distribution.**

The method includes subtracting the mean from each variable observation and then dividing by the difference between the minimum and the maximum value of that variable:



Here’s what you need to remember about mean normalization:

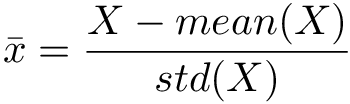
* It centers the mean at 0.
* The resulting variance will be different.
* It may modify the shape of the original distribution.
* It “normalizes” the minimum and maximum values within the range[-1, 1].
* It preserves outliers if they exist.

# Standardization

This technique will not **normalize the variable distribution.**

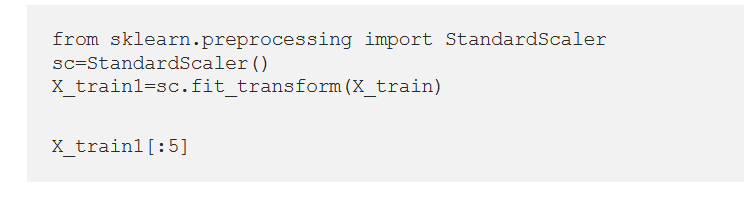
It is also called **Z-score normalization**.

Standardization suggests centering the variable at 0 and standardizing the variance to 1. The procedure includes subtracting the mean from each variable observation and then dividing by the standard deviation:



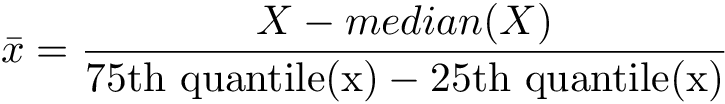
Here’s what you need to remember about standardization:

* It scales the variance at 1.
* It centers the mean at 0.
* It preserves the shape of the original distribution.
* It preserves outliers if they exist.
* Minimum and maximum values vary.



# Robust Scaling (scaling to median and IQR)

In this method, the median is used instead of the mean. We remove the median from the variable observations, and then we scale to the inter-quantile range (IQR).



The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).

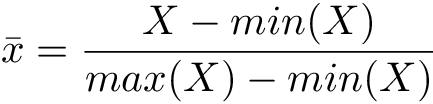
Here’s what you need to remember about robust scaling:

* It centers the medianat 0 .
* The resulted variance varies across variables.
* It may not preserve the shape of the original distribution.
* The minimum and maximum values vary.
* It is robust to outliers.

# Min-Max Scaling

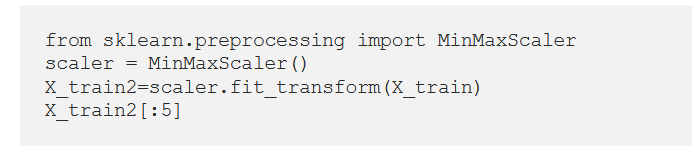
Minimum and maximum scaling both compress the values between 0 and 1. It subtracts the minimum value from all the variable observations, and then divides it by the variable’s value range:

This technique will not **normalize the variable distribution.**



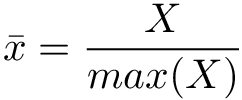
Here’s what you need to remember about Min-Max Scaling:

* It does not center the mean at 0.
* It makes the variance vary across variables.
* It may not maintain the shape of the original distribution.
* The minimum and maximum values are in the range of [0,1].
* This method is very sensitive to outliers.



# Maximum Absolute Scaling

Maximum absolute scaling scales the variable values between -1 and 1 by dividing the data by its maximum value:

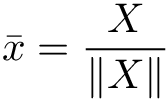


Here is what you need to remember about Maximum Absolute Scaling:

* The resulting mean is not centered.
* It doesn't scale the variance.
* It’s sensitive to outliers.

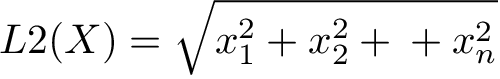
# Scaling to vector unit norm

In scale to the vector unit norm, we divide each feature’s vector by the distance of the vector, as shown below:



For the distance measure, you can use either:

* **The** [**Euclidean distance**](https://en.wikipedia.org/wiki/Euclidean_distance) ( — or **L2** norm) with the formula:



* **The** [**Manhattan distance**](https://en.wikipedia.org/wiki/Taxicab_geometry) ( — or **L1** norm) with the formula:

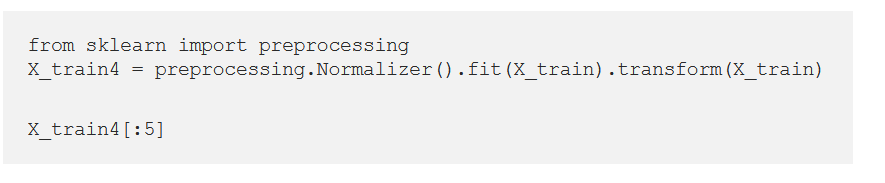


Here’s what you need to remember about scaling to the vector unit norm:

* The length of the resulting vector is 1.
* It normalizes the **feature** vector and not the **observation** vector.
* It’s sensitive to outliers.
* Recommended for text classification and clustering.

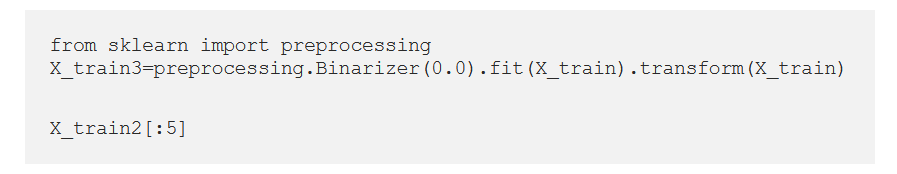
**Normalizing**

It is used to rescale each sample. **Each sample** (i.e. each row of the data matrix) with at least one non zero component is rescaled independently of other samples so that its norm (l1 or l2) equals one.



**Binarizing**

It is used for binary thresholding of an array like matrix.



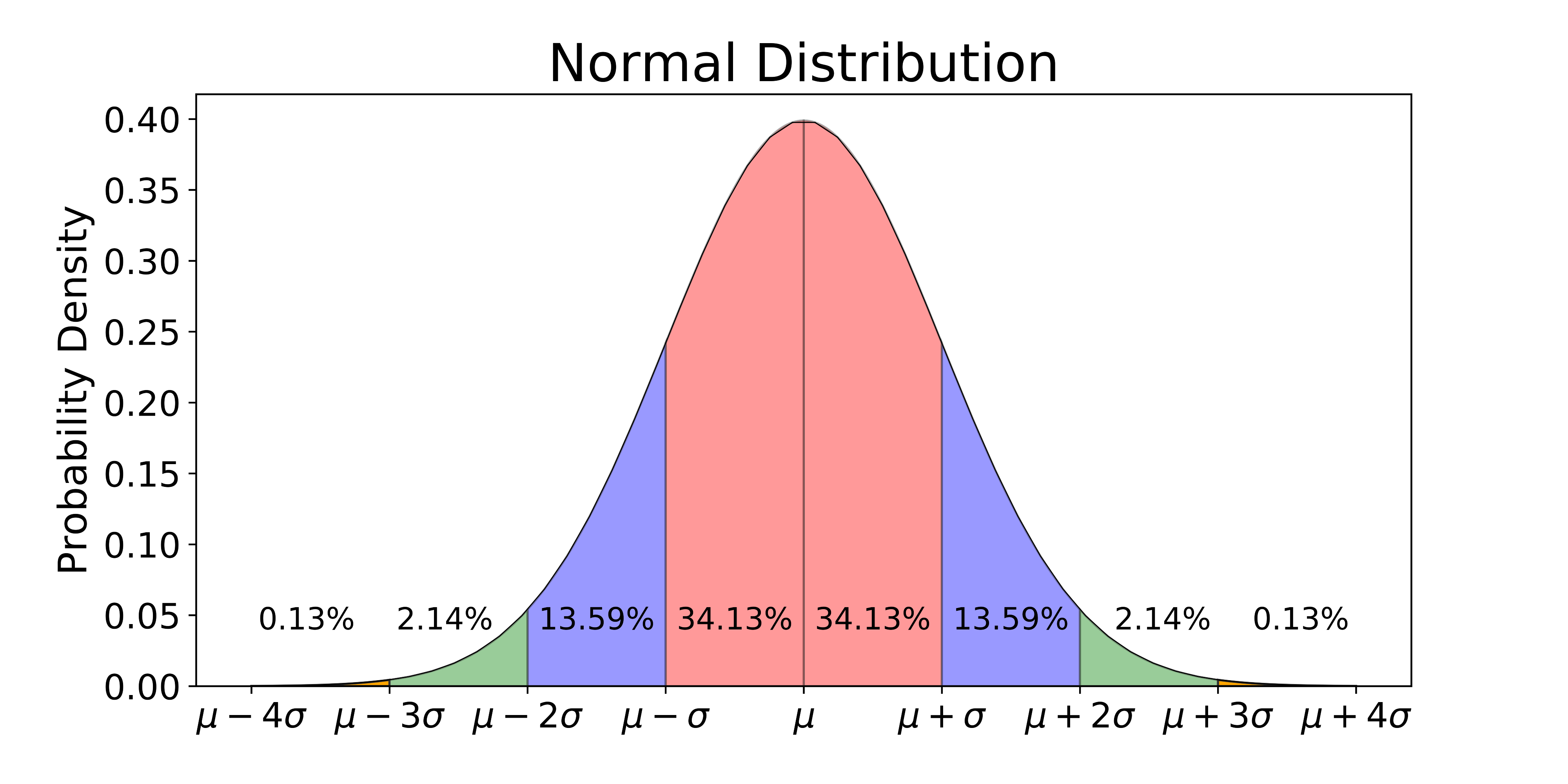
## Normalize or Standardize?

**Normalization** :

* Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution.
* This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

**Standardization :**

* Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true.
* Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.



**Normalizer** :

* [Normalizer](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html#sklearn.preprocessing.Normalizer) is also a normalization technique.
* The only difference is the way it computes the normalized values. By default, it is calculating the l2 norm of the row values i.e. each element of a row is normalized by the square root of the sum of squared values of all elements in that row.
* It is useful in text classification where the dot product of two [Tf-IDF](https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/) vectors gives a cosine similarity between the different sentences/documents in the dataset.
* From sklearn.preprocessing.Normalizer