**Standardization (Standard Normal Distribution:)**

Converting a standard distribution in to standard normal distribution is called Standardization. The standard normal distribution is also called the z-distribution; **any normal distribution can be standardized by converting its values into z-scores formula**.

When you standardize a normal distribution, then the mean becomes 0 and the standard deviation becomes 1. This allows you to easily calculate the probability of certain values occurring in your distribution, or to compare data sets with different means and standard deviations. We can convert normal distribution to Standard Normal Distribution using the Z-score Formula

**Steps of Standardizing a normal distribution:**

1. Calculate the Standard Deviation:
2. Calculate Standard Normal Distribution using the Zscore Formula.
3. **Calculate the Standard Deviation:**

X [1,2,3,4,5]

|  |  |  |
| --- | --- | --- |
| Standard Deviation  (σ)= | standard deviation formula for population and sample |  |
|  |
|  |
|  |
|  | |  |
|  |
| 1.414 | | |
|

1. **Calculate Standard Normal Distribution using the Zscore Formula.**

X {1,2,3,4,5}

|  |  |  |
| --- | --- | --- |
| Formula | Value | Y |
|  | 1-3/1.414 | -1.414 |
| 2-3/1.414 | -0.707 |
| 3-3/1.414 | - |
| 4-3/1.414 | 0.707 |
| 5-3/1.414 | 1.414 |

Using the standard normal distribution, or more specifically, standardizing data to have a mean of 0 and a standard deviation of 1, offers several advantages in statistics and data analysis.

Here are some of the main reasons why the standard normal deviation is commonly used:

* **Comparability:**
* **Simplification:**
* **Outlier Detection:**
* **Feature Scaling:**

**Comparability:**

Allows for direct comparability between different datasets

**Simplification:**

simplifies calculations and interpretations

**Outlier Detection:**

Identifying outliers becomes easier with the standard normal distribution. Data points with Z-scores far from zero (e.g., Z-scores greater than 3 or smaller than -3) are often considered outliers, making outlier detection more straightforward.

**Feature Scaling:**

When features have vastly different scales, some algorithms (e.g., gradient descent-based optimization) may converge faster and provide better results when the features are standardized.

**Normalization (Min Max Scaler)**

Normalization is the process of organizing the data in the database. · Normalization is used to minimize the redundancy from a relation or set of relations.

**Min Max Scaler for Normalization**

Many machine learning algorithms perform better when numerical input variables are scaled to a standard range.

MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Formula | | Values | Y |
| 1 |  | X – Xmin | 1-1/10-1 | 0.000 |
| 2 |  | 2-1/10-1 | 0.111 |
| 3 |  | 3-1/10-1 | 0.222 |
| 4 | Xscaler = | 4-1/10-1 | 0.333 |
| 5 | Xmax - Xmin | 5-1/10-1 | 0.444 |
| 6 |  | 6-1/10-1 | 0.556 |
| 7 |  | 7-1/10-1 | 0.667 |
| 8 |  | 8-1/10-1 | 0.778 |
| 9 |  | 9-1/10-1 | 0.889 |
| 10 |  | 10-1/10-1 | 1.000 |

**Normalization Vs Standardization**

|  |  |
| --- | --- |
| Normalization | Standardization |
| The minimum and maximum values of features are used for scaling | Mean and the standard deviation is used for scaling. |
| It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| It is really affected by outliers. | It is much less affected by outliers. |
| Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | It translates the data to the mean vector of the original data to the origin and squishes or expands. |
| It is useful when we don’t know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |
| It is often called Scaling Normalization | It is often called Z-Score Normalization. |

**When to use Normalization or Standardization?**

Which is suitable for our machine learning model, Normalization or Standardization? This is probably a big confusion among all data scientists as well as machine learning engineers. Although both terms have almost the same meaning, choosing normalization or standardization will depend on your problem and the algorithm you use in the models.

**Normalization** is a transformation technique that helps to improve the performance as well as the accuracy of your model better. Normalization is useful when the feature distribution of data does not follow a **Gaussian** (bell curve) distribution. Normalization must have an abounding range, so if you have outliers in data, they will be affected by Normalization.

**Standardization:** in the machine learning model is useful when you are exactly aware of the feature distribution of data or, in other words, your data **follows a Gaussian distribution**. However, this does not have to be necessarily true. Unlike Normalization, Standardization does not necessarily have a bounding range, so if you have outliers in your data, Standardization will not affect them.