Gaussian distribution: Why is it important in data science and machine learning?

Rohit Sharma Follow
Jun 13, 2019 · 4 min read

And how to transform your dataset to Gaussian?

Gaussian Distribution (credit: Physion)

Distributions can be either tricky or simple things depending on your background and mistakes you have made in the past. One of my dear friend recently asked me”

> *if I have a data distribution that is not strictly gaussian in nature, taking a log of the data distribution will make it gaussian. Conceptually why is that so?*

Function *log(x)* is simply used as another transform that was suitable for his example.

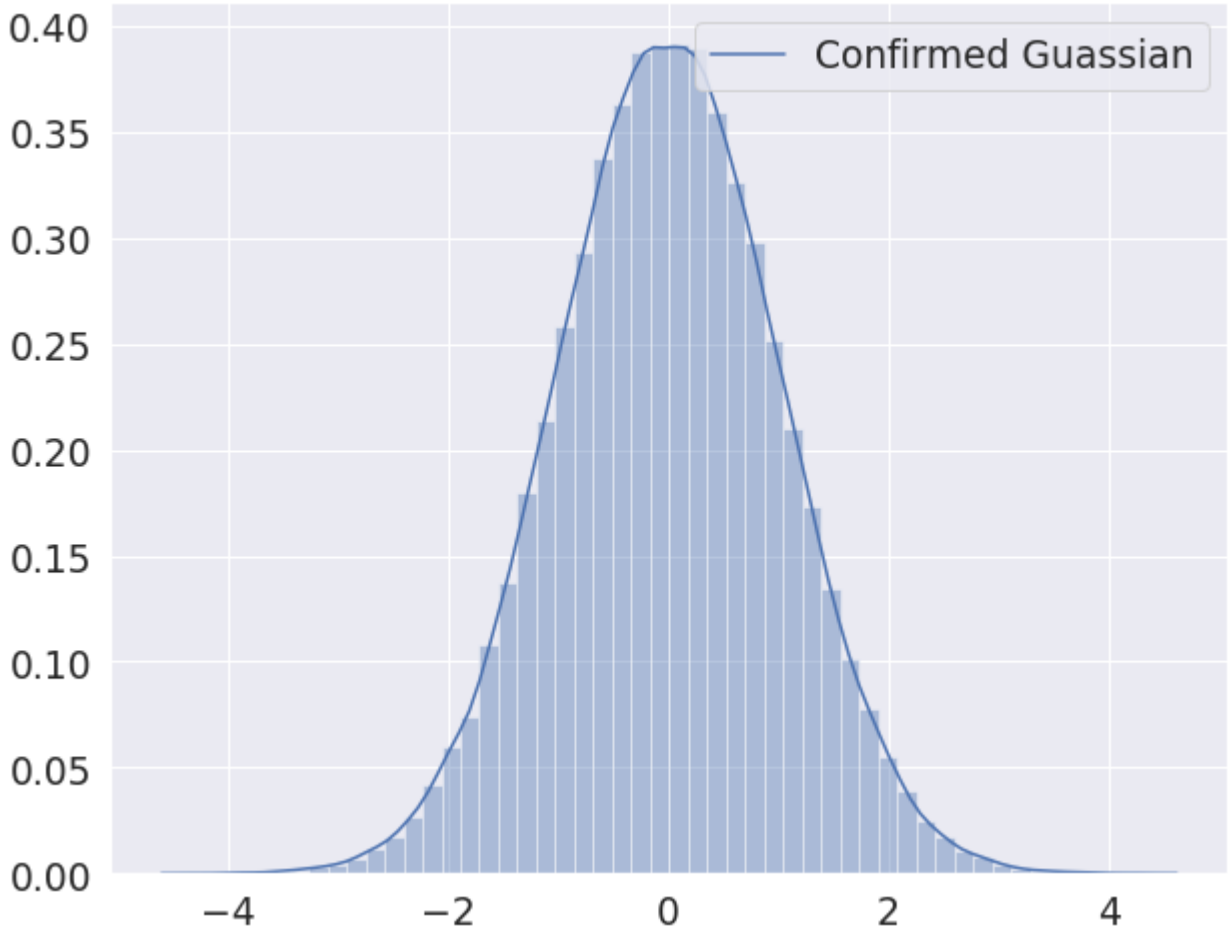
> *And if that is true, is it true for any distribution or are there limits to “likeness to Gaussian distribution” that will limit the virtue of making then distribution normalized.*

So let's define Gaussian and benefits thereof in layman's terms:

What is a Gaussian Distribution?

A distribution is simply a collection of value and frequency of a given observation, like age of a population.

Samples of a ideal Gaussian distribution (aka normal distribution or bell curve) follow bell curve distribution meaning values are more likely around mean over extremes. Figure below shows 10k floating values generated with Gaussian distribution.



Gaussian Distribution Dataset with 10,000 samples

Other commonly used distributions are binomial distribution poisson distribution.

Why is Gaussian Distribution Important?

1. Gaussian distribution is ubiquitous because a dataset with finite variance turns into Gaussian as long as dataset with independent feature-probabilities is allowed to

grow in size. Gaussian distribution is the most important probability distribution in statistics because it fits many natural phenomena like age, height, test-scores, IQ scores, sum of the rolls of two dices and so on.

2. Datasets with Gaussian distributions makes applicable to a variety of methods that fall under parametric statistics. The methods such as propagation of uncertainty and least squares parameter fitting that make a data-scientist life easy are applicable only to datasets with normal or normal-like distributions.

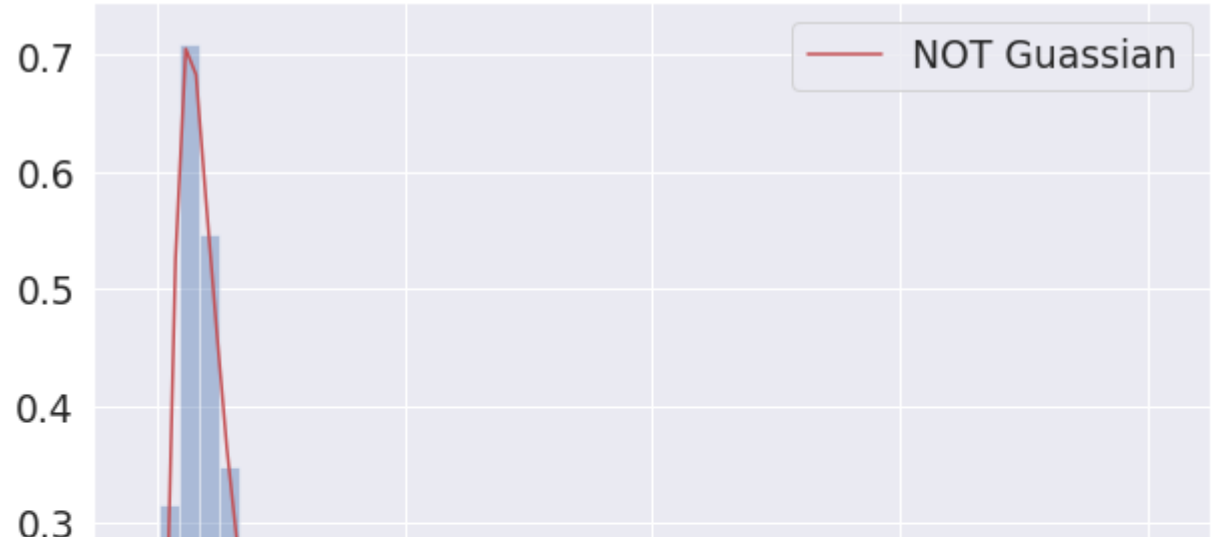
3. Conclusions and summaries derived from such analysis are intuitive and easy to explain to audiences with basic knowledge of statistics.

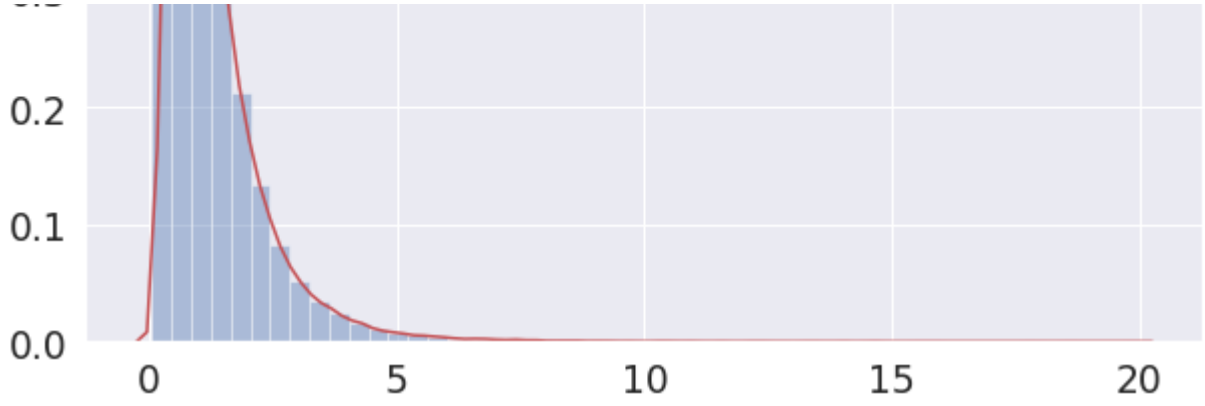
Why are They Important to Machine Learning?

In machine learning, cost function or a neuron potential values are the quantities that are expected to be the sum of many independent processes (such as input features or activation potential of last layer) often have distributions that are nearly normal. One can continue to use parametric statistics knowing gaussian nature of dataset.

An Example of How to Transform any Distribution to a Gaussian-like Distribution.

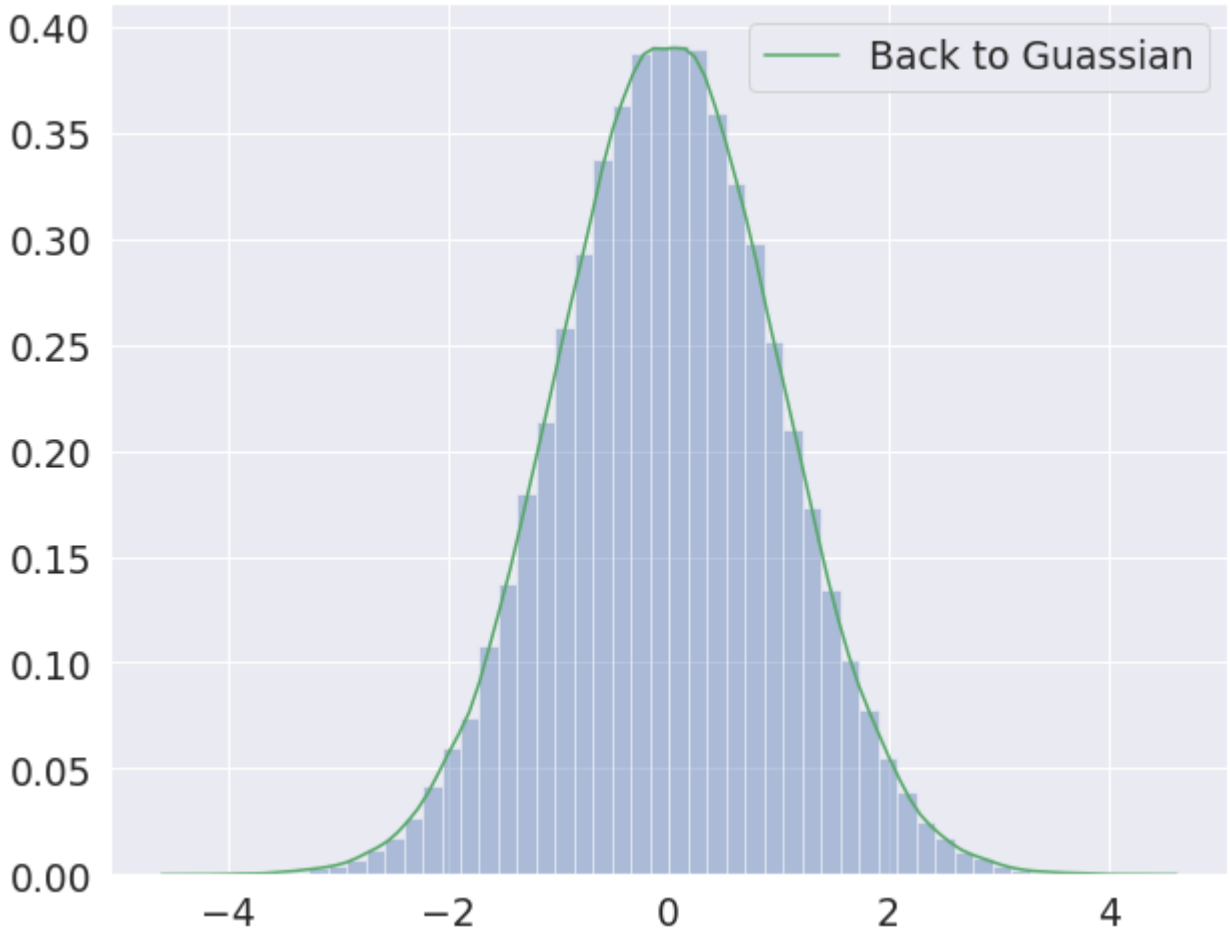
It is possible to transform any distribution to a Gaussian-like distribution using an appropriate transform. The picture below shows a dataset with sample frequencies multiplied by power of 2, so the distribution becomes skewed to the left with a long tail on the right.





Non-Gaussian Distribution.

We can come up with an appropriate reverse-transform to make this dataset Gaussian. With the knowledge that sample frequencies were multiplied by a power of 2, we can use reverse-transform as logarithm of 2 to change the distribution. Picture below shows the distribution after applying log2(x) transformation.



Transformed Log-Gaussian Distribution

This was simply a contrived example for showing a good application of power transform. This approach was generalized by George Box and Sir David Cox in a paper titled “An Analysis of Transformations” published in 1964.

Box Cox transformation is a way to **transform** non-normal dependent variables into a normal shape. Other family of power transformation have been proposed since.

Summary

In summary, it is possible to turn most datasets to Gaussian. For more complex ones with several peaks and long tails, you may have to consider advanced methods like power transform or other data transformation methods.

References:

1. Python notebook for converting a dataset from non-gaussian to gaussian with contrived example.
2. Box, George E. P.; Cox, D. R. (1964). “An analysis of transformations”. *Journal of the Royal Statistical Society, Series B*. **26** (2): 211–252. JSTOR 2984418. MR 0192611.
3. Yeo, In-Kwon; Johnson, Richard A. (2000). “A New Family of Power Transformations to Improve Normality or Symmetry”. *Biometrika*. **87** (4): 954–959. doi:10.1093/biomet/87.4.954. JSTOR 2673623.