Getting to Know Your Data-II

IT- 3031 (Cr-3)

Dr. Amiya Ranjan Panda
Assistant Professor [II]
School of Computer Engineering,
Kalinga Institute of Industrial Technology (KIIT),
Deemed to be University,Odisha

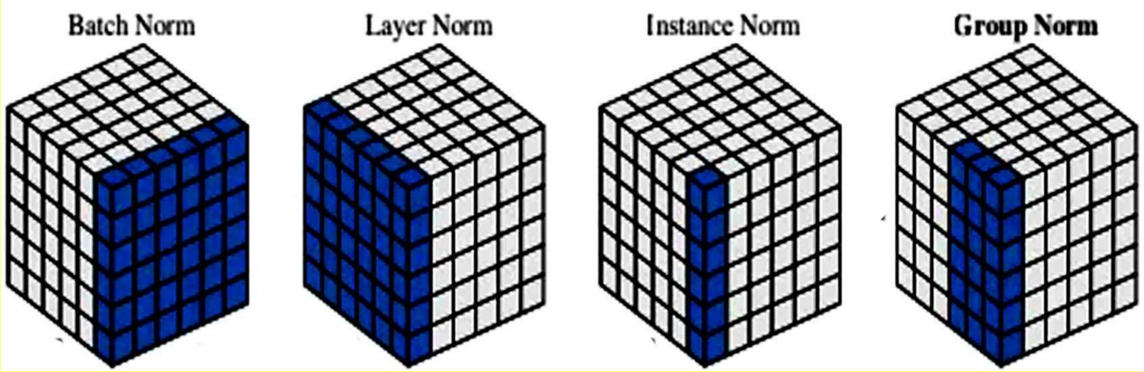
Data Normalization

- Normalization is generally required when multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations.
- Otherwise, it may lead to a dilution in effectiveness of an important equally important attribute(on lower scale) because of other attribute having values on larger scale.
- Heterogenous data with different units usually needs to be normalized. Otherwise, data has the same unit and same order of magnitude it might not be necessary with normalization.
- Unless normalized at pre-processing, variables with disparate ranges or varying precision acquire different driving values.

Methods of Data Normalization

- Normalization is normally done, when there is a distance computation involved in our algorithm.
- Methods of Data Normalization:
  - **Decimal Scaling**
  - **Min-Max Normalization**
  - **z-Score Normalization(zero-mean Normalization)**

- **There are several approaches in normalisation which can be used in deep learning**
  - **Batch Normalization**
  - **Layer Normalization**
  - **Group Normalization**
  - **Instance Normalization**
  - **Weight Normalization**



Methods of Data Normalization...

- **Decimal Scaling Method For Normalization**
  - It normalizes by moving the decimal point of values of the data.
  - To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data.
  - The data value, v_i, of data is normalized to v'_i by using the formula

$$\boxed{v_i' = \frac{v_i}{10^j}}$$

where j is the smallest integer such that max($|v'_i|$)<1.

In this technique, the computation is generally scaled in terms of decimals. It means that the result is generally scaled by multiplying or dividing it with pow(10,k).

Example:

- Let the input data is:-15,121,201,421,561,601,850
- To normalize the above data,
- **Step 1:** Maximum absolute value in given data(m):850
- **Step 2:** Divide the given data by 1000 (i.e j=3)
- Result: The normalized data is:-0.015,0.121,0.201,0.421,0.561,0.601,0.85

Methods of Data Normalization...

- ## Min-Max Normalization

  - In this technique of data normalization, linear transformation is performed on the original data.
  - Minimum and maximum value from data is fetched and each value is replaced according to the following formula.

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}(\text{new\_max}(A) - \text{new\_min}(A)) + \text{new\_min}(A)$$

- Where A is the attribute data,
- ***min***(A), m***ax***(A) are the minimum and maximum absolute value of A respectively.
- v' is the new value of each entry in data.
- v is the old value of each entry in data.
- new_max(A), new_min(A) is the max and min value of the range(i.e boundary value of range required) respectively.

Roll No	Marks
1	10
2	15
3	50
4	60

Example

If we were to normalize it between the ranges of 0 to 1 we would get the following

Roll No	Marks
1	0
2	0.1
3	0.8
4	1

Methods of Data Normalization…

- **z-Score Normalization (zero-mean Normalization)**
  - In this technique, values are normalized based on mean and standard deviation of the data A.
  - It is also called **Standard Deviation method.**
  - So, the unstructured data can be normalized using z-score parameter, the formula for z-score is as below;

$$v' = \frac{v - \bar{x}}{s}$$

where, $\bar{x}$ is the mean and S is the standard deviation.
v is the old value of each entry in data.
v' is the Z-score normalized of each entry in data.

Example

Roll No	Marks
1	10
2	15
3	50
4	60

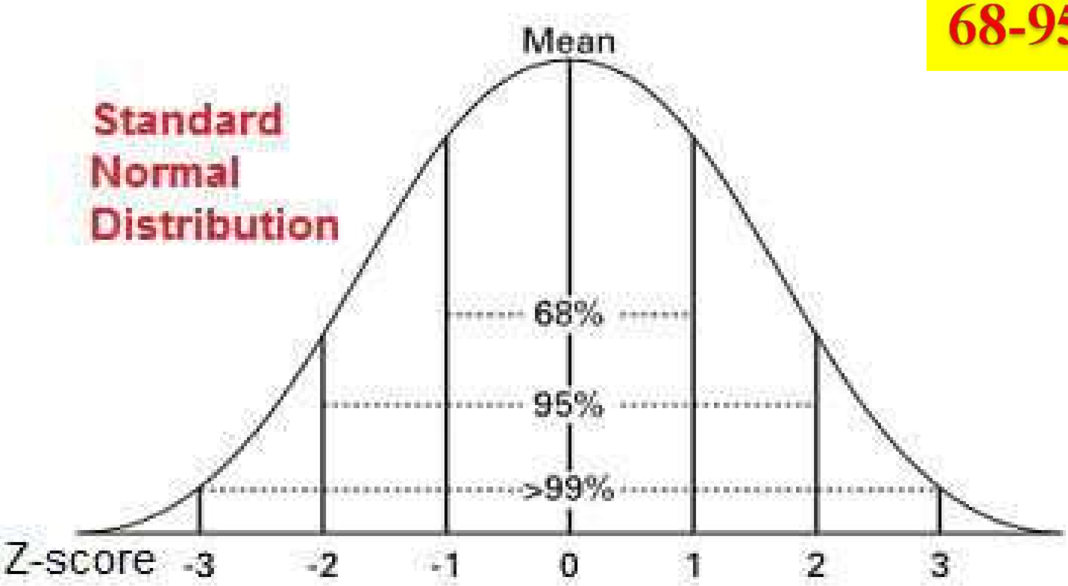
Mean is 33.75 and Standard Deviation is 24.95

Roll No	Marks
1	-0.951587303
2	-0.751253134
3	0.651086049
4	1.051754387

Methods of Data Normalization...

The normal distribution is a probability function that describes how the values of a variable are distributed.

No matter what μ and σ are,
the area between $\mu-\sigma$ and $\mu+\sigma$ is about 68%;
the area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%; and the area between $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%.
Almost all values fall within 3 standard deviations.



68-95-99.7 Rule, in Math terms…

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .997$$

THANK YOU!