

ASSIGNMENT NO 02

DATASCIENCE

SUBMITTED TO: DR. AAMIR ARSALAN

SUBMITTED BY: UBAIDA WAHEED

2019\_BSE\_034

1. **Ordinal Encoding and Label Encoding** [**https://www.youtube.com/watch?v=w2GglmYHfmM**](https://www.youtube.com/watch?v=w2GglmYHfmM)

Ordinal Encoding and Label Encoding –

In this video, the contrast between label encoding and ordinal encoding are discussed along with their applications in machine learning.

When converting categorical data to numerical values, ordinal encoding is used, and the categorical values' are in specific order (such as "low", "medium", and "high").

On the other hand, label encoding gives each category a different numerical value without taking the order into account.

The video also demonstrates how it's crucial to make sure there is no connection between the numbers assigned to each category and their real values.

1. **One Hot Encoding -** [**https://www.youtube.com/watch?v=U5oCv3JKWKA**](https://www.youtube.com/watch?v=U5oCv3JKWKA)

One Hot Encoding –

The idea of one hot encoding, which enables categorical variables to be represented as binary vectors, is explained in this video.

Each unique category is given a binary vector, in which every value is set to zero except for the point where the category belongs.

One hot encoding might be helpful when dealing with categorical variables that have several categories and when the categories lack any sort of intrinsic order.

However, it can make the data more dimensional, which would make the model more complicated and challenging to train.

1. **Log Transform**

[**https://www.youtube.com/watch?v=cTjj3LE8E90**](https://www.youtube.com/watch?v=cTjj3LE8E90)

Log Transform –

The video demonstrates how skewed data with a longer tail on one side can be normalised using the log transform.

The log transform can be used to lessen the impact of outliers and enhance the visibility of data patterns. Additionally, it demonstrates that data with a broad range of values and when we want to lessen the influence of outliers can benefit from log transformation.

The data are transformed using the log function, and the log function's base can be changed to fit the particular data set being analyzed.

Additionally, the film demonstrates the application of log transformation to a number of situations, including feature scaling, non-linear problem solving, and visualization of data.

Additionally, it demonstrates the limitations of log transformation, such as the inability to apply to negative and zero values.

1. **Box Cox Transform**

[**https://www.youtube.com/watch?v=lV\_Z4HbNAx0**](https://www.youtube.com/watch?v=lV_Z4HbNAx0)

Box Cox Transform –

The Box-Cox transformation, a technique for giving non-normal dependent variables a normal shape, is described in this video.

This transformation is applied to reduce the variance of a variable and improve the normality of the data distribution. When studying skewed data with the intention of transforming it into a more normal distribution, it is also helpful.

Only positive data can be transformed using the Box-Cox method, and the user must specify the transformation parameter lambda.

The video also demonstrates how to use the maximum likelihood estimate approach to determine the ideal lambda value and how to read box-cox transformation results. Additionally, it highlights the restrictions placed on the box-Cox transformation, including the requirement for normalcy and the inability to apply to negative numbers.

1. **Yeo Johnson Transform -** [**https://www.youtube.com/watch?v=lV\_Z4HbNAx0**](https://www.youtube.com/watch?v=lV_Z4HbNAx0)

Yeo-Johnson Transform –

The Yeo-Johnson transformation, an expansion of the Box-Cox transformation that may be used with both positive and negative data, as explained in the video.

It is employed to reduce the variance of a variable and improve the normality of the data distribution. Similar to the Box-Cox transformation, the user must specify the lambda transformation parameter.

Unlike the Box-Cox transformation, which is only applicable to positive data, the Yeo-Johnson transformation can be used to both positive and negative data.

The video also demonstrates how to use the maximum likelihood estimate approach to determine the ideal lambda value and how to understand the outcomes of the Yeo-Johnson transformation.

1. **Discretization**

[**https://www.youtube.com/watch?v=kKWsJGKcMvo**](https://www.youtube.com/watch?v=kKWsJGKcMvo)

The discretization technique, which transforms continuous values into categorical variables, is explained in the video.

This can be achieved by creating a collection of intervals from a variable's range and allocating each data point to a certain interval.

Data visualisation and the development of new features for machine learning models can both benefit from discretization. Different discretization techniques, including equal width, equal frequency, k-means, decision trees, and chi-merge, are explained in the video.

The data is divided into a given number of intervals using the equal width approach, and each interval is the same width. The data is divided into intervals using the equal frequency method such that each interval has the same number of data points. The k-means approach divides the data into a predetermined number of intervals using the k-means clustering algorithm. The decision tree approach divides the data into intervals based on the values of the feature using a decision tree algorithm. The chi-merge approach reduces the number of intervals by merging intervals using the statistical test chi-square.

It also illustrates how discretization can be helpful for selecting features, addressing missing information, and lowering dimensionality. But it also notes that Discretization can result in loss of information, as well as the problem of arbitrary binning.

1. **What are Outliers?**

[**https://www.youtube.com/watch?v=Lln1PKgGr\_M**](https://www.youtube.com/watch?v=Lln1PKgGr_M)

A data point that differs significantly from the other data points in a dataset is an outlier.

Inaccuracies in data collection, measurement, or data input might result in outliers, or they can be the result of real differences in the group under study.

The analysis of a dataset can be significantly impacted by outliers, and they can also affect the outcomes of statistical tests. They can be viewed as noise, which has a detrimental effect on the model's capacity to generalise to new data, and they can also influence how well machine learning models perform. The movie describes the various kinds of outliers, including Univariate and Multivariate outliers, as well as their causes, including measurement inaccuracy, data input problems, data manipulation, and natural outliers.

Additionally, it demonstrates how to locate outliers using several techniques, including box plot, scatter plot, z-score, and Mahalanobis distance. The movie also demonstrates how to deal with outliers by eliminating, changing, or imputing them.

It highlights how crucial it is to understand the context and data domain in order to distinguish between an outlier and a true result.

1. **Outlier detection and removal using Z-score method**

[**https://www.youtube.com/watch?v=OnPE-Z8jtqM**](https://www.youtube.com/watch?v=OnPE-Z8jtqM)

The Z-score approach for locating and eliminating outliers from a dataset is described in the video. A data point's Z-score indicates how far away from the dataset's mean it is from the mean. For every data point, the Z-score can be determined by subtracting the mean from the value and dividing by the standard deviation. Outliers are data points that have a Z-score over a predetermined limit (often 3 or -3), and they can be eliminated from the dataset.

The Z-score method in Python is also demonstrated in the video using the numpy package. It demonstrates how to calculate the Z-score for each data point, choose an outlier detection threshold, and exclude outliers from the dataset.

It also describes the benefits and drawbacks of the Z-score approach. Its simplicity, foundation in the conventional normal distribution, and reduced sensitivity to the presence of extreme values are its main benefits. The drawbacks are that it is unsuitable for datasets with non-normal distributions, it is sensitive to the existence of outliers, and it is not resistant against the presence of extreme values.

1. **Percentile method**

[**https://www.youtube.com/watch?v=bcXA4CqRXvM**](https://www.youtube.com/watch?v=bcXA4CqRXvM)

Another method for locating and eliminating outliers in a dataset is the percentile method for outlier detection.

Based on the idea that most of the data in a dataset should fall inside a specific range of percentiles, the percentile method was developed. The 25th and 75th percentiles, for instance, can be used to find outliers that are outside the first and third quartile range.

The Percentile Method involves calculating the appropriate percentiles first, and then determining the lower and upper bounds for the data by either adding or removing a specific number of standard deviations from the percentiles.

Outliers are any data points that deviate from these bounds and can be eliminated from the dataset. Similar to the IQR approach, it's critical to keep in mind that the Percentile Method may not be the ideal one for every dataset, therefore before using it, it's vital to take into account the situation and the data's underlying assumptions.

**Outlier detection and removal using IQR method**

[**https://www.youtube.com/watch?v=Ccv1-W5ilak**](https://www.youtube.com/watch?v=Ccv1-W5ilak)

Outliers in a dataset can be found and eliminated using the IQR method for outlier detection.

The approach is founded on the idea that the majority of data in a dataset should fall between the first and third quartiles (Q1 and Q3). Data points that are outside of this range are known as outliers.

To apply the interquartile range (IQR) approach, first get the IQR by deducting Q1 from Q3. Then, by multiplying the IQR by 1.5 and then adding or removing it from Q1 and Q3, determine the lower and upper bounds for the data. Outliers are any data points that deviate from these bounds and can be eliminated from the dataset.It's important to note that while the IQR method can be an effective way to identify and remove outliers, it is not always the best method for every dataset, and it's important to consider the context and the underlying assumptions of the data before applying it.