DATA TRANSFORMATION AND DISCRETIZATION

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran
Department of Mathematical Sciences
Universiti Kebangsaan Malaysia

INTRODUCTION:

- In the data transformation, the data is modified into a more appropriate form before the data mining analysis is carried out

- Some data transformation techniques::

1. Normalization:

- Involves the process of re-scaling attribute values.

Example: The original data attribute range (0 - 100) is scaled in a smaller range, (0 - 1). Especially for a dataset with a multiple attributes that are not the same in term of unit of measurements.

- This process also being used to transform the distribution of the original data to a Normal distribution form.
- Most methods in statistics and data mining require assumptions of data normality

Example: Regresion model.

INTRODUCTION:

2. Discretization :

- The process of converting an attribute value (example: age) to a value in a particular form of intervals (example: 0–10, 11–20, 20-40)
- Or in a conceptual form (example: children, adolescents, adults, seniors).

3. Attribute Formation:

- New attributes are formed from a combination or tranformation of existing attributes in the data.

4. Smoothing and etc.

NORMALIZATION:

- Intended to re-scale attribute values in some specific range.
- Transform the data distribution to approximate the Normal distribution.

i. Min-Max Normalization:

- Involves a linear transformation of the data.
- Suppose min_X and max_X are the minimum and maximum values of the attribute X.
- We want to re-scale the range of X [min_X, max_X] to a new range given as [new_min_X, new_max_X].
- This normalization will convert all X values to a V value correspond to the interval [new_min_X, new_max_X].
- This can be done through the following formula:

$$V = \frac{\left[X - \min(X)\right] \times \left[baru\_\max(X) - baru\_\min(X)\right]}{\max(X) - \min(X)} + baru\_\min(X)$$

ii. Z-score Normalization:

- This method is also known as zero-mean normalization.
- The value of the X attribute will be converted to Z-score using the following formula:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

- where μ_X and σ_X are the mean and standard deviation for attribute X.
- If the values of μ_X and σ_X are unknown, it will be estimated from the sample.
- The value of the Z-score will have a mean of 0 and a standard deviation of 1.
- This method is more suitable than min-max normalization if the outliers are present in the dataset.

iii. Normalization based on decimal scaling:

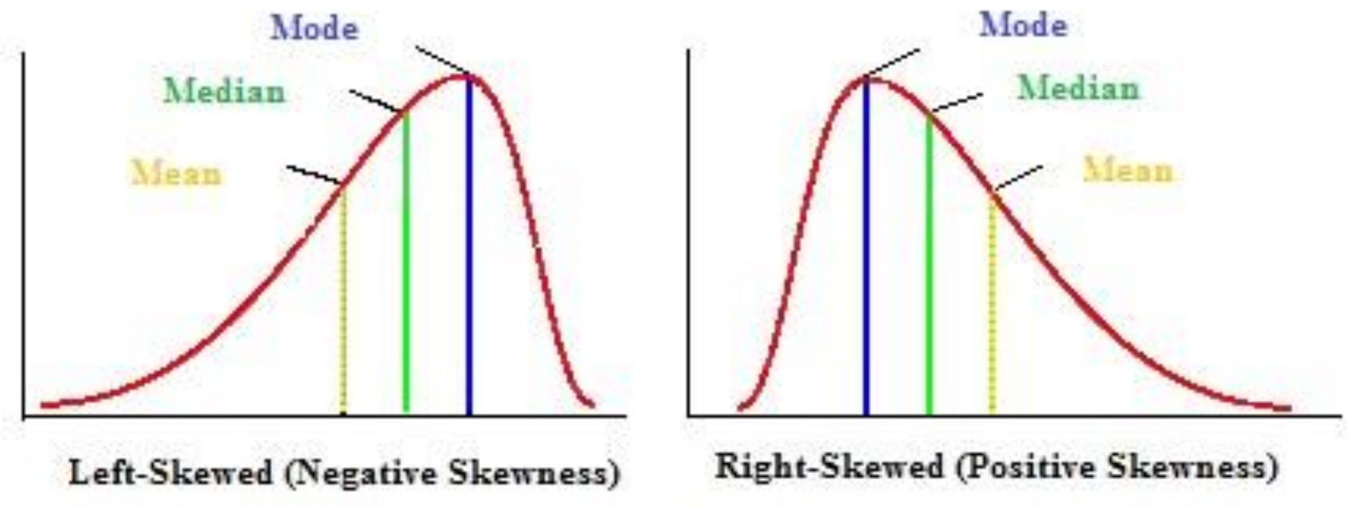
- Converts data by based on the decimal point of the for attribute X
- The number of decimal points depends on the maximum absolute value of X
- This is done by using the following formula:

$$v'_i = \frac{v_i}{10^j}$$

- Where j is the smallest integer such that $\max\left(v'_i\right) < 1$.

iv. Normaling data distribution:

- This type of transformation needs to be performed if the data is skewed to the right (positive) or to the left (negative)



- This transformation involves a mathematical function against the values of attribute.

Example: log10, square root , and etc.

- Some of the mathematical functions used in the normalization of data distribution are:

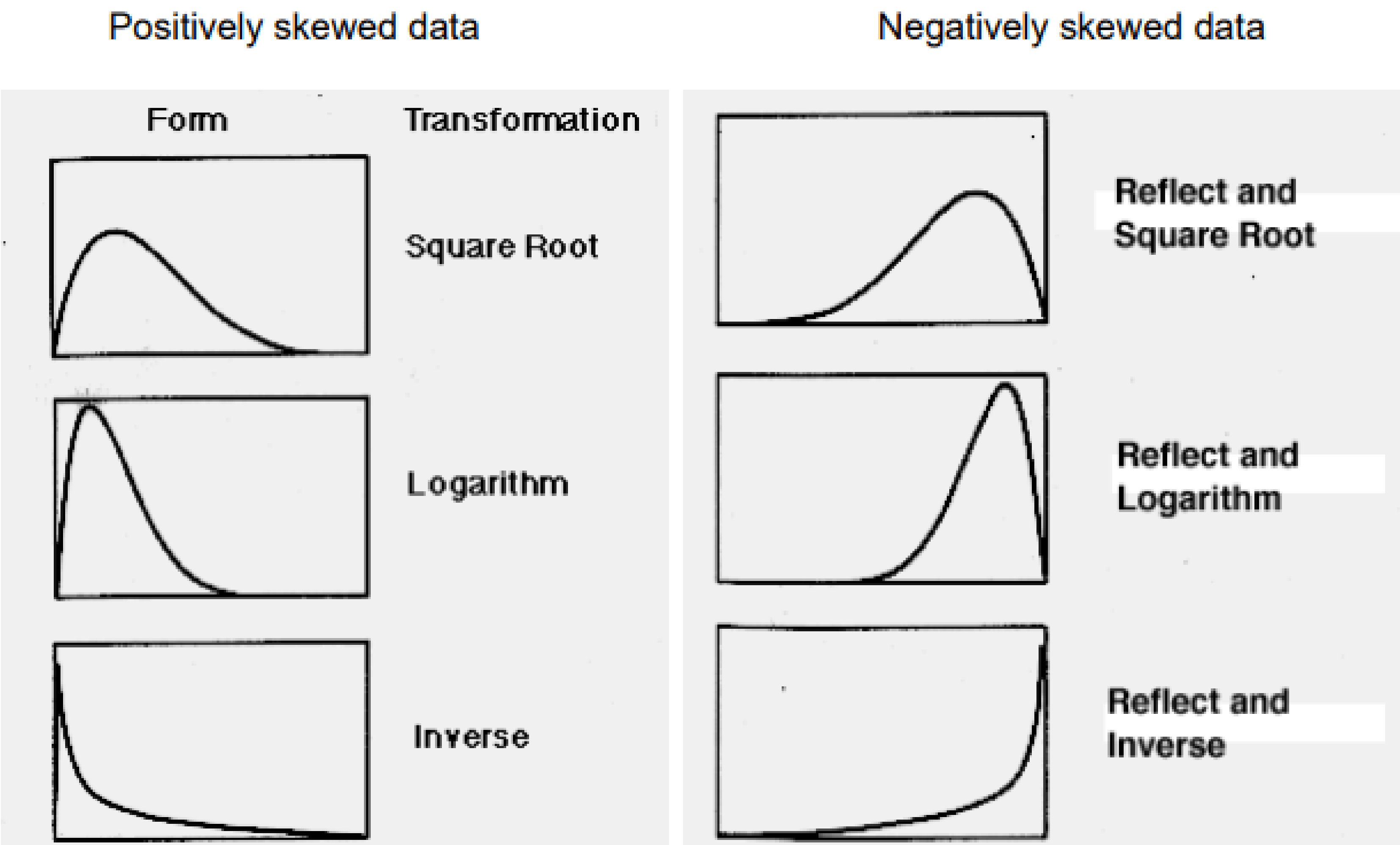
a) Logarithm:

- Transformation through the log (x) function is appropriate if the variance of the data is found to increase against the mean of the data.
- It is also suitable for growth rate data that typically have exponential distributions.

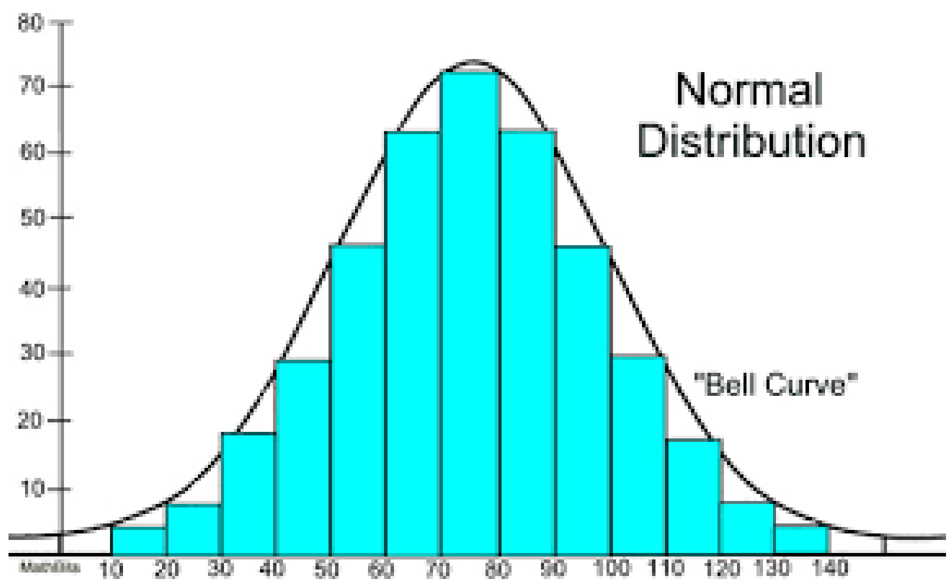
b) If the logarithm is not suitable, several other functions can be tried, for example:

- Reciprocal Transformation
- Square Root Transformation ($x^1/2$).
- Arcsine Transformation (asin(x)): also known as angular transformation and is useful for percentage or proportion type data.

- The figure suggests an appropriate mathematical function depending on the degree of skew of the original distribution data.



- Data that is skewed to the left (negative) requires a reflected transformation.
- The data needs to be reflected first before the transformation is made.
- The reflection of a variable is made by the formation of a new variable with the original value of the data subtracted by a constant, k.
- The constant k is calculated by adding 1 to the largest value of the original variable, $k = (\max(x) + 1)$.
- Next, the reflected variable (P) is compute as: $P = k - X$



- Tabachnick & Fidell (2007) and Howell (2007) provide the following procedures for data transformation based on the skewness of the original data distribution.

Original Data	Proposed Transformation Techniques
Moderate Positive Skewness	Power, $Y = X^2$
Highly Positive Skewness	Logarithm, $Y = \log_{10}(X)$
Moderate Negative Skewness	Square Root, $Y = \sqrt{k - X}$
Highly PNegative Skewness	Logarithm, $Y = \log_{10}(k - X)$

* Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston: Allyn and Bacon.

* Howell, D. C. (2007). Statistical methods for psychology (6th ed.). Belmont, CA: Thomson Wadsworth.

METHODS FOR ASSESSING NORMALITY DATA:

i. Histogram and Boxplot

ii. Normal Quantile Plot
- also known as Normal Probability Plots.

iii. Goodness-of-fit test:

i) Kolmogorov-Smirnov Test

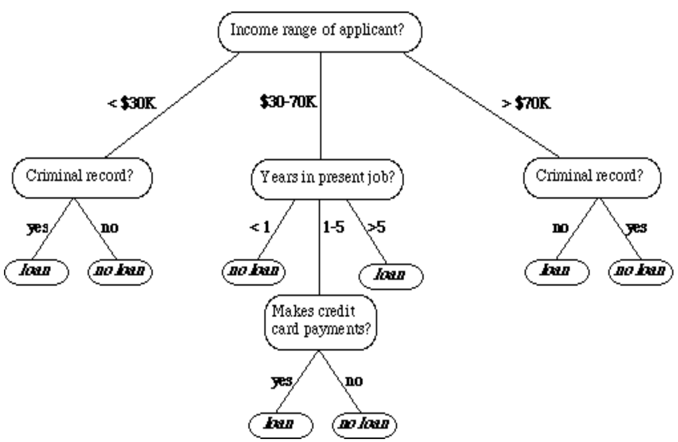
ii) Shapiro-Wilk Test

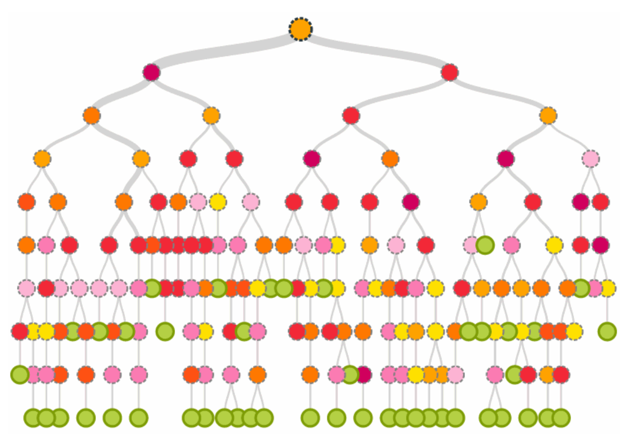
iii) Anderson-Darling Test

DISCRETIZATION:

- Discretization is the process of dividing attribute data into several intervals.

- Data in the form of intervals are used to replace an actual data.
- Some data mining methods can only be performed on discrete data. Example: Decision trees.
- Discretization is also an approach to reduce the data to make the data mining algorithm become more efficient.
- Discretization can be performed repeatedly on the same attributes.





- Through discretization, attributes in numerical form (continuous) will be converted to attributes in discrete or interval form.

Example:

Cotinuous: Total Income, 1000< X< 10000.

Interval: 1000-2000, 2000-3000, >3000.

Discrete/Categorical: 1=low income, 2=moderate income, 3=high income.

- The purpose of discretization is to reduce the number of continuous attribute values by grouping them to the number of *b*-intervals or bin.

- An important issue in discretization is how to select the number of intervals/bins.

- Two approach: supervised approach and unsupervised approach.

- Unsupervised approach: No class labels are known. The discretization interval can be run directly on the data

- Supervised approach: If the class label is known, the discretization method should take advantage of this information, and the model algorithm can be used.

- The discretization method should maximize the dependence between attribute values and class labels and minimize information loss.

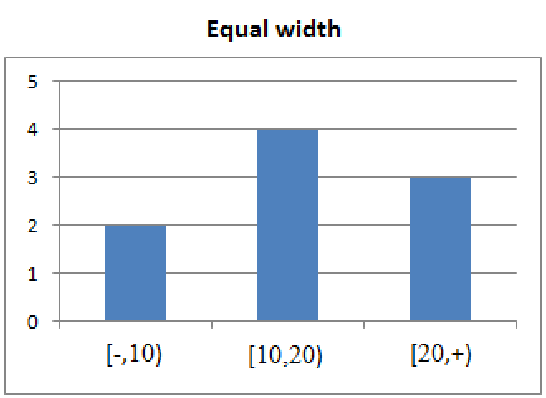
UNSUPERVISED DISCRETIZATION :

i. Data discretization through domain knowledge:

- Manual discretization.
- However, data scientists need to have appropriate arguments regarding the division of the interval.

ii. Equal-width discretization:

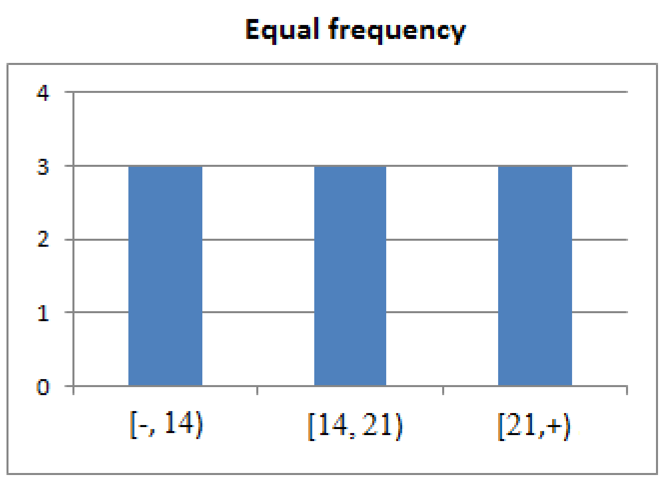
- This algorithm use the information about the minimum (A) and maksimum (B) values for attribute, X_i .
- The width of the discretization interval is compute as follow: $W = (B - A)/N$.



UNSUPERVISED DISCRETIZATION:

iii. Equal-frequency discretization:

- This algorithm use information about minimum (A) and maximum (B) value of attribute, X_i .
- Next, the value of X_i will be ordered in ascending order.
- The width of the discretization interval was determined based on the same number of observations in each interval.



SUPERVISED DISCRETIZATION:

- Supervised Discretization algorithm take into account the class information in the data set.

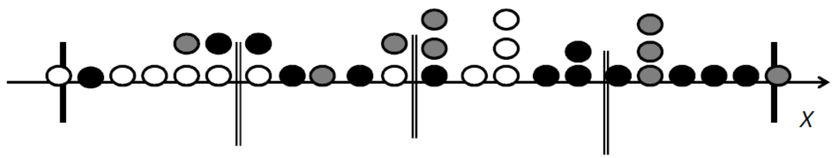


Fig. 6.2 Distribution of values belonging to three classes {white, gray, black} over variable *X*.

- Various Supervised Discretization algorithm can be carried out using R:

i) Discretization using Chi2 algorithm.

ii) Discretization using ChiMerge algorithm .

iii) Discretization using Top-down algorithm.

iv) Discretization using Minimum Description Length Principle (MDLP) algorithm.

v) And etc.

VARIOUS DISCRETIZATION ALGORITHM:

Equal Width Discretizer	**EqualWidth**
Equal Frequency Discretizer	**EqualFrequency**
No name specified	Chou91
Adaptive Quantizer	AQ
Discretizer 2	D2
ChiMerge	**ChiMerge**
One-Rule Discretizer	**1R**
Iterative Dichotomizer 3 Discretizer	**ID3**
Minimum Description Length Principle	**MDLP**
Valley	Valley
Class-Attribute Dependent Discretizer	**CADD**
ReliefF Discretizer	ReliefF
Class-driven Statistical Discretizer	StatDisc
No name specified	NBIterative
Boolean Reasoning Discretizer	BRDisc
Minimum Description Length Discretizer	MDL-Disc
Bayesian Discretizer	**Bayesian**
No name specified	Friedman96
Cluster Analysis Discretizer	**ClusterAnalysis**
Zeta	**Zeta**
Distance-based Discretizer	**Distance**
Finite Mixture Model Discretizer	FMM

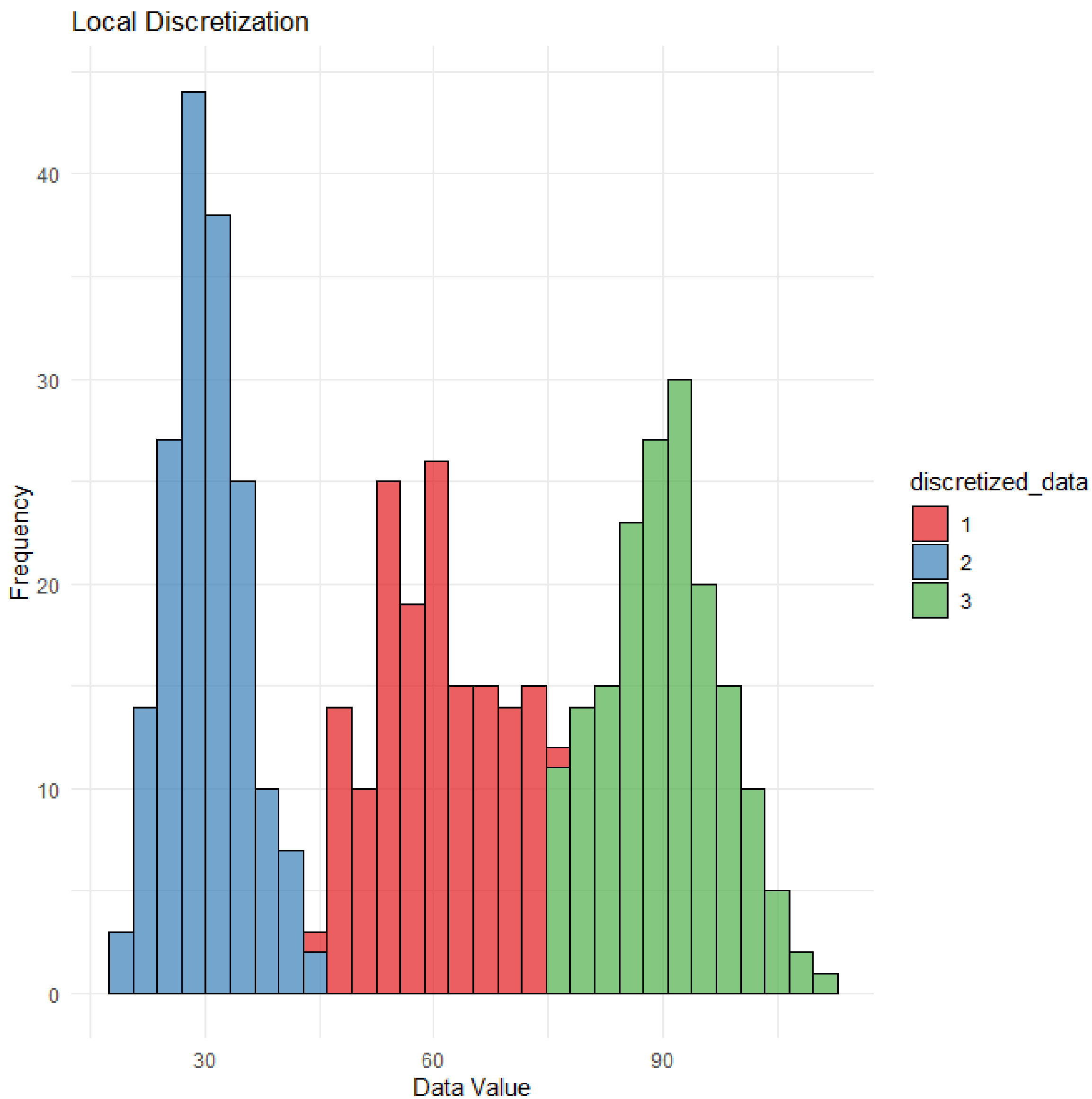
No name specified	Butterworth04
No name specified	Zhang04
Khiops	**Khiops**
Class-Attribute Interdependence Maximization	**CAIM**
Extended Chi2	**Extended Chi2**
Heterogeneity Discretizer	**Heter-Disc**
Unsupervised Correlation Preserving Discretizer	**UCPD**
No name specified	Multi-MDL
Difference Similitude Set Theory Discretizer	DSST
Multivariate Interdependent Discretizer	MIDCA
MODL	**MODL**
Information Theoretic Fuzzy Partitioning	ITFP
No name specified	Wu06
Fast Independent Component Analysis	FastICA
Linear Program Relaxation	LP-Relaxation
Hellinger-Based Discretizer	**HellingerBD**
Distribution Index-Based Discretizer	**DIBD**
Wrapper Estimation of Distribution Algorithm	WEDA
Clustering + Rought Sets Discretizer	Cluster-RS-Disc
Interval Distance Discretizer	**IDD**
Class-Attribute Contingency Coefficient	**CACC**
Rectified Chi2	Rectified Chi2

LOCAL DISCRETIZATION:

- Local discretization refers to a data discretization method that considers the local characteristics of the data distribution.
- This approach is useful when we are dealing with heterogeneous data, with different data segments which requiring different discretization strategies.

- Example: In a data set with varying density of data distribution, local discretization needs to construct; i) smaller bins in areas of high data density, and, ii)larger bins in areas of low data density.
- By adapting the discretization process to the local context, this method can help preserve important patterns and relationships in the data.

LOCAL DISCRETIZATION:



DATA TRANSFORMATION TO FORM NEW ATTRIBUTES:

- New attributes can be constructed from several combinations or transformation of existing attributes in the data.

- Example:

i) the attribute for "area" can be constructed from the values of the "length" and the "width" attributes.

ii) the attribute for "BMI" can be constructed from the values of the "weight" and the "height" attributes of the individual.

iii) the attribute for "net income" can be constructed from the sum of the values of all "income" related and the subtraction of all individual "debt" related attributes.

- Domain knowledge is essential to define the correct relationship between attributes.

- New attributes can also be formed through various mathematical relationships between attributes in a data set.

- Among the methods of Data Transformation to form New Attributes are:

i) Linear Transformation:

- This technique involves simple algebraic transformations such as addition, averages, rotations, and etc.
- Let $A = A_1, A_2, \ldots, A_n$ is a set of attributes, and let $B = B_1, B_2, \ldots, B_m$ is a subset for set of attribute in A.
- The new attribute Z can be formed through the following linear transformation:

$$Z = r_1 B_1 + r_2 B_2 + \cdots + r_M B_M$$

ii) Data transformation through encoding:

- This technique used to convert categorical data into a numerical format so that it can be effectively used data mining analysis.
- Many data mining algorithms, such as decision trees or linear regression, expect numerical input, and thus encoding is a crucial step when dealing with categorical variables (Example: colors or labels).

- Several Encoding methods:

i) One-Hot Encoding;

ii) Ordinal Encoding;

iii) Target Encoding;

iv) Frequency Encoding;

v) And many more.

ii) Rank Transformation:

- This transformation is carried out in order to replace the numerical value of the attribute to the value of the rank attribute.
- The attribute will change to a new attribute that contains integer values (rank, r_i) between 1 to m (in ascending or descending order).
- The rank can be transformed into data in the form of a Normal score using the following equation:

$$y_i = \Phi^{-1}\left(\frac{r_i - \frac{3}{8}}{m + \frac{1}{4}}\right)$$

iii) Box-Cox Transformation:

- The Box-Cox transformation aims to make the new attributes of the data distributed approximate to the Normal distribution through the following equation:

$$y = \begin{cases} x^{\lambda-1}/\lambda, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$

- The value of λ should be between −3.0 to 3.0. The best value of λ is selected if the distribution is found to be close to normal
- However, the above formula is limited to non -negative data. Data that have negative values, some modifications need to be done.

Other transformation methods:

- Polynomial Approximation transformation.
- Non-Polynomial Approximation transformation.
- Wavelet transformation.
- And etc.

REFERENCES:

- Aggarwal, C.C. (2015). *Data Mining. The Textbook*. Springer, New York.
- García, S., Luengo, J., Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer, New York.
- Jafari, R. (2022). *Hands-On Data Preprocessing in Python: Learn how to effectively prepare data for successful data analytics 1st Edition*. Packt Publishing
- Kuhn, M., Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall.
- Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., Herrera, F. (2020). *Big Data Preprocessing*. Springer, Switzerland.

NEXT TOPIC:

Data Reduction