STUTERN

Feature Engineering (2)

SGA07_DATASCI

4th Febraury 2020

Module Overview

- Data Integration & Reduction
- Data Transformation & Discretisation

Book Keeping

- Peer Reviews
- Guest Lectures
- Sort into 3 Groups
- Catch up on Tasks/Practice Labs so far

Outcome

After this Module, you will;

- Overview of data integration and reduction as part of the preprocessing tasks.
- Explore some techniques for data integration and reduction: Attribute matching and correlation analysis
- Overview of data transformation and discretisation as concluding part of preprocessing tasks
- Explore how to write a custom function for equal-width approach to discretisation

Data Integration

- Attribute matching
- Correlation analysis
- Tuple duplication
- Data Value detection

> This step can help to reduce and avoid redundancies as well as inconsistencies in your data as well improve accuracy and speed of other steps in the data science process.

STUTERN

Attribute Matching

Data integration can be done by referring to the metadata of various sources to effectively match attributes from various sources.

Jss 1 Science

Student_ID	Height
3	167
8	148
15	135

Jss 1 Art

ID	Height
1	165
2	135
5	176

Jss 1

Student_ID	Height
1	165
2	135
3	167
5	176
8	148
15	135

Correlation Analysis

- Nominal Data
- Numeric Data

> Correlation analysis enables the detection of redundancy in your data, when one attribute is derived from another with implicit rules attached

Nominal Correlation Analysis

> The chi-square or Pearson Statistic is used to evaluate correlation relationship between two nominal attributes

$$\tilde{\chi}^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

A	B
a1	b3
A3	b2
A4	B4
A1	B1
A2	B3
A3	B2
A2	b4
A4	b1
A3	B2

	B1	B2	B3	B4	Total
A1	1		1		2
A2			1	1	2
A3		3			3
A4	1			1	2
	2	3	2	2	9

Numeric Correlation Analysis

> The correlation coefficient or Pearson's product moment coefficient is used to evaluate correlation relationship between two numeric attributes

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Data Reduction

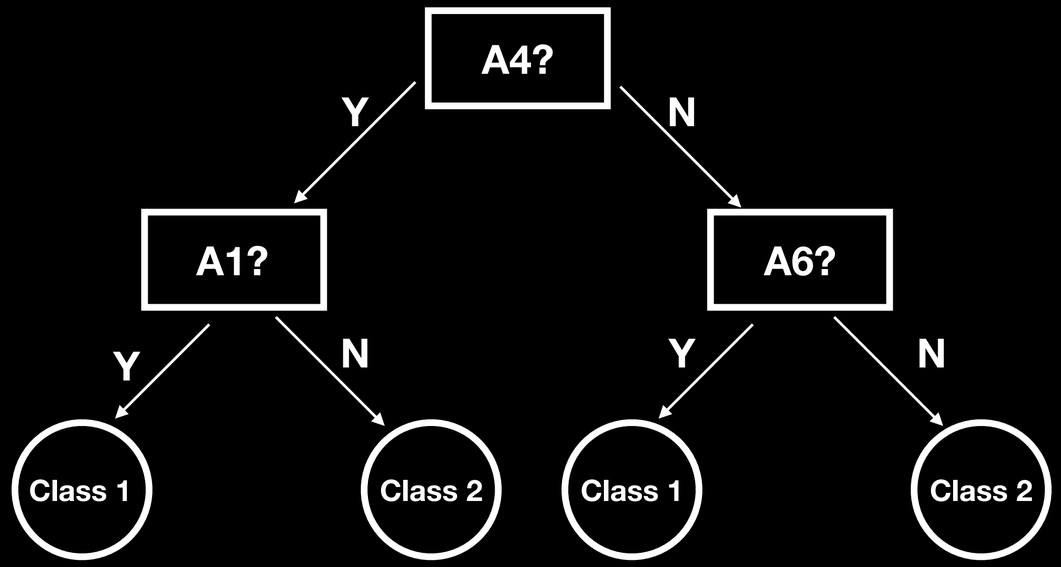
- Attribute Subset Selection
- Wavelet Transforms
- Principal Component Analysis

> This step helps to obtain a representation of the data set that is much smaller in volume, yet maintains the integrity of the original data..

Attribute Subset Selection

> Reduces the data set size by removing irrelevant or redundant attributes (or dimensions).

Initial Attribute Set $(A1, A2, A3, A4, A5, A6)$



Reduced Attribute Set $(A1, A4, A6)$

Principal Component Selection

- The input data are normalised, so that each attribute falls within the same range.
- PCA computes k orthonormal vectors that provide a basis for the normalised input data.
- The principal components are sorted in order of decreasing "significance" or strength.
- Because the components are sorted in decreasing order of "significance," the data size can be reduced by eliminating the weaker components, that is, those with low variance.

> PCA searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$.

Practice Lab

Quick Examples of Data Integration and Reduction

Use the following Instructions:

- Use the merge package in R to join two data frames together
- Use the chisq.test package in R to explore correlation relationship between gender and selected ice-cream

Data Transformation

- Smoothing
- Feature Engineering
- Data Aggregation
- Normalisation
- Discretisation

> Transform the data so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

Normalisation

> Help avoid dependence on the choice of measurement units, the data should be normalised or standardised, transforming the data to fall within a smaller or common range such as [−1,1] or [0.0, 1.0].

Min-Max Normalisation

$$v_i' = \frac{v_i - min_A}{max_A - min_A}$$

Jss 1

Student_ID	Height
1	165
2	135
3	167
5	176
8	148
15	135

Jss 1-1

Student_ID	Height
1	0.73
2	0.00
3	0.78
5	1.00
8	0.32
15	0.00

Discretisation

- Discretisation by binning
- Discretisation by histogram analysis
- Discretisation by correlation analysis
- Discretisation by decision tree

> Raw values of a numeric attribute (e.g. age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).

Discretisation by Binning

- Get maximum value
- Get minimum value
- Create width interval based on specified bin value
- Create cut value based on width interval between max and min values
- For each numeric value replace with minimus cut value

> Binning is a top-down splitting technique based on a specified number of bins.

Practice Lab

Quick Example of Data Discretisation

Use the following Instructions:

- Create a custom function to discretise a numeric attributes using equal-width approach

Recap/Summary

At the end of this Module, you should understand;

- Overview of data integration and reduction as part of the preprocessing tasks.
- Explore some techniques for data integration and reduction: Attribute matching and correlation analysis
- Overview of data transformation and discretisation as concluding part of preprocessing tasks
- Explore how to write a custom function for equal-width approach to discretisation

Suggested Material

- Data Mining Concepts and Techniques (3rd Edition) by Jiawei Han, Micheline Kamper and Jian Pei: Chapter 3 - Data Preprocessing
- https://uc-r.github.io/pca
- http://www.dataintegration.ninja/data-integration-techniques-and-its-challenges/
- http://www.programmingr.com/tutorial/left-join-in-r/
- https://en.wikipedia.org/wiki/Data_integration