SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

COURSE NAME :19CS407 DATA ANALYTICS WITH R

III YEAR /VI SEMESTER

Unit 2- GETTING INSIGHTS FROM DATA

Topic: Data Quality and Preprocessing

Data Quality

- ✓ even though there is a large number of robust descriptive and predictive algorithms available to deal with noisy, incomplete, inconsistent or redundant data, an increasing number of real applications have their findings harmed by poor-quality data.
- ✓ When these data are used by algorithms that learn from data – ML algorithms – the analysis problem can look more complex than it really is if there is no data pre-processing.
- ✓ The elimination or even just the reduction of these problems can lead to and improvement in the quality of knowledge extracted by data analysis processes.

Data Quality factors

- ✓ Data quality is important and can be affected by internal and external factors.
- ✓ • Internal factors can be linked to the measurement process and the collection of information through the attributes chosen.
- ✓ External factors are related to faults in the data collection process, and can involve the absence of values for some attributes and the voluntary or involuntary addition of errors to others.

Missing values

- some of predictive attribute values for some of the records may be missing in the data set. There are several causes
- • attributes values only recorded some time after the start of data collection,
- • the value of an attribute being unknown at time of collection
- • distraction, misunderstanding or refusal at time of collection
- • attribute not required for particular objects
- • non-existence of a value
- • fault in the data collection device
- • cost or difficulty of assigning a class label to an object

Handling missing values

- ✓ Ignoring the missing values- ignore the attributes and proceed with algorithm and modify algorithm to accept missing values
- ✓ Removing the objects- remove objects with missing values
- ✓ Making estimates- Filling the missing values with estimation from other values

Handling missing values

- The simplest alternative is just to remove objects with missing values in a large number of attributes.
- Objects should not be discarded when there is a risk
- of losing important data. Another simple alternative is to create a new, related, attribute, with Boolean values: the value will be true if there was a missing value in the related attribute, and false otherwise.

Handling missing values

- The filling of missing data is the most common approach. Methods are
- Fill with location- Filling with mean, median and mode
- For classification tasks use values from same class
- learning algorithm- prediction value for particular missing attribute
- the first method is the simplest and has the lowest processing cost. The second method has a slightly higher cost, but gives a better
- estimate of the true value. The third method can further improve the estimate, with a higher cost.

Example

Data with missing values			
Food	Age	Distance	Company
Chinese	51	Close	Good
			Good
Italian	82		Good
Burgers	23	Far	Bad
Chinese	46		Good
Chinese			Bad
Burgers		Very close	Good
Chinese	38	Close	Bad
Italian	31	Far	Good

Example

Data with missing values				Data without missing values			
Food	Age	Distance	Company	Food	Age	Distance	Compan
Chinese	51	Close	Good	Chinese	51	Close	Good
			Good	Chinese	53	Close	Good
Italian	82		Good	Italian	82	Close	Good
Burgers	23	Far	Bad	Burgers	23	Far	Bad
Chinese	46		Good	Chinese	46	Close	Good
Chinese			Bad	Chinese	31	Far	Bad
Burgers		Very close	Good	Burgers	53	Very far	Good
Chinese	38	Close	Bad	Chinese	38	Close	Bad
Italian	31	Far	Good	Italian	31	Far	Good

Redundant data

missing values are a lack of data, redundant data is the excess of it.
Redundant objects are those that do not bring any new information to a data set.
Thus, they are irrelevant data redundant data can be duplicate data.
Deduplication is a preprocessing technique whose goal is to identify and remove copies of objects in a data set,

Redundant data

Data with redundant objects				Data without redundant objects			
Food	Age	Distance	Company	Food	Age	Distance	Company
Chinese	51	Close	Good	Chinese	51	Close	Good
Italian	43	Very close	Good	Italian	43	Very close	Good
Italian	43	Very close	Good	—	—	—	—
Italian	82	Close	Good	Italian	82	Close	Good
Burgers	23	Far	Bad	Italian	82	Close	Good
Chinese	46	Very far	Good	Chinese	46	Very far	Good
Chinese	29	Too far	Bad	Chinese	29	Too far	Bad
Chinese	29	Too far	Bad	—	—	—	—
Burgers	42	Very far	Good	Burgers	42	Very far	Good
Chinese	38	Close	Bad	Chinese	38	Close	Bad
Italian	31	Far	Good	Italian	31	Far	Good

Inconsistent data

Inconsistent values can be found in the predictive and/or target attributes.

An example of an inconsistent value in a predictive attribute is a zip code that does not match the city name. This inconsistency can be due to a mistake or a fraud.

This can lead to ambiguity for predicting same value for two objects

Inconsistent data

Friend	Maxtemp (°C)	Weight (kg)	Height (cm)	Gender	Company
Andrew	25	77	175	M	Good
Bernhard	31	**1100**	195	M	Good
Carolina	15	70	172	F	Bad
Dennis	20	**45**	**210**	M	Good
Eve	10	65	168	F	Bad
Fred	12	75	173	M	Good
Gwyneth	16	75	**10**	F	Bad
Hayden	26	63	165	F	Bad
Irene	15	55	158	F	Bad
James	21	66	163	M	Good
Kevin	**300**	95	190	M	Bad
Lea	13	72	**1072**	F	Good
Marcus	8	83	185	F	Bad
Nigel	12	115	192	M	Good

Assessment 1

In what situations is it better to use infographics to represent information present in a data set?

References

1. João Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

Thank You