Chapter 4 - Types of Data Manipulations

Study Exercise – Learning the most famous data transformation methods

Dwnload *Chapter_4_Study1_Empty.docx* from this link, and its solution file *Chapter_4_Study1.pdf* from this link

Study Exercise – data cleaning issues

Use any reference that you might find useful, including our good friend Google.com, to study and draft an answer for the following question.

1. What are the differences and similarities between Error and Noise?
2. What are the difference and similarities between Errors and Missing values?
3. What are the difference and similarities between Errors and Outliers?
4. What are the common strategies to address Missing Values?
5. Do outliers always pose an issue for our analytics? When do they do so, and why?
6. Assume that we have performed some data manipulations to address one or all of the level 3 data cleaning issues. Should we call the data manipulations Data Cleaning or Data Transformation? Explain.

Answers to questions

Q1: What are the differences and similarities between Error and Noise?

Similarity: They are both issues with the data that we need to resolve.

Difference: Error is avoidable, but noise is not as noise happens due to random irregularities in the phenomena that we measure, but the error happens due to systematic mistakes in data collection.

Q2: What are the differences and similarities between Errors and Missing values?

Similarity: They are both issues with the data that we need to resolve.

Similarity: They both happen due to systematic problems in the data collection

Difference: For errors, we have a wrong measurement, but for missing values, we lost the window of opportunity to be able to measure.

Q3: What are the difference and similarities between Errors and Outliers?

In principle, they bear no similarities. The only similarity they have is that we use the same analytic tools to detect outliers and errors. Essentially, when we detect data objects too different from the rest of the population, we check if the values may have been errors, if not, they are outliers.

Q4: What are the common strategies to address Missing Values?

- Leave as as
- Drop the row
- Drop the column
- Impute a new value

Q5: Do outliers always pose an issue for our analytics? When do they do so, and why?

Most often they do not. In situations where the large numerical differences between the outlier and the rest of the population consume the analytics capabilities, then they need to be dealt with. For instance, see Data Transformation in Action – An example with data in Chapter 3.

Q6: Assume that we have performed some data manipulations to address one or all of the level 3 data cleaning issues. Should we call the data manipulations Data Cleaning or Data Transformation? Explain.

It doesn't matter, we use these distinctive terms to learn about various data preparation that needs to be taken place. When it comes to action, one action may be doing more than one thing and we may not be able to find one distinctive name for that manipulation action.

Study Exercise – Reviewing Data Reduction tools and concepts.

Use your knowledge, the preceding study material, or any other resources to answer the following questions.

1. What are the differences between Numerosity Data Reduction and Dimensionality Data Reduction?

2. When performing data reduction we must be careful to maintain a reasonable balance between two objectives, what are those two objectives?
3. What is the Class Imbalance Problem and how it is related to data reduction?
4. What is the difference between supervised and unsupervised dimension reduction? Mention three tools for performing both supervised and unsupervised dimension reduction.

Answers to questions

Q1: What are the differences between Numerosity Data Reduction and Dimensionality Data Reduction? Numerosity Data Reduction reduces the data by reducing the data volume or the number of data objects (rows), and dimensionality data reduction reduces the data by reducing the data variety or the number of attributes (columns).

Q2: When performing data reduction we must be careful to maintain a reasonable balance between two objectives, what are those two objectives?

1. Keeping the integrity of the data and causing bias to the data as little as possible
2. Reducing the size of the data

Q3: What is the Class Imbalance Problem and how it is related to data reduction?

When the binary dependent attribute of a dataset has overwhelmingly higher repetition for one of the classes over the other. For instance, in a fraud detection dataset, most of the rows represent non-fraudulent records and that's the imbalance. This causes the algorithms to be skewed toward the more dominant class. It is related to data reduction as one of the methods to deal with this problem is by resampling which is a form of data reduction.

Q4: What is the difference between supervised and unsupervised dimension reduction? Mention three tools for performing both supervised and unsupervised dimension reduction.

The Supervised dimension reduction is also called feature selection. It is the task of finding the subset of attributes that will lead to the best classification/prediction performance. In contrast, unsupervised dimension reduction is about finding a new representation of the data that keeps the most important part of the data and cut the fluff out.

Three tools for supervised dimension reduction: Decision/Regression Tree, Random Forest, Brute-force dimension reduction

Three tools for unsupervised dimension reduction: Principal Component Analysis, Functional Data Reduction, Feature extractio