Exercises Data Cleaning and Data Transformation
Distance
-
Getting to know RapidMiner Studio and Jupyter Notebook

Exercise 1: Data cleaning means check the following

- missing data
- erroneous data
- duplicate data
- obsolete data

and after, possibly, take action. Are the issues mentioned above present in the Golf data set below? This dataset contains information about the weather at different times and dates and whether this weather is fine to play golf – attribute play.

Exercise 2: Data transformation involves one or several of the steps listed below.

- Aggregation
- Sampling
- Attribute (Dimensionality) reduction
- Attribute (Feature) subset selection
- Attribute (Feature) creation
- Discretization and Binarisation
- Attribute transformation

Aggregation: Read 3.3 data Integration in the book.

1. Compute corr(x, y), x=(-3, 6, 0), y=(1, -2, 0). Report your answer and show your calculation. What is the transformation that allows to calculate x from y?
2. Compute corr(x, a * x). Report your answer and show your calculation.
3. Consider the contingency table below and calculate with the chi square Test whether the features gender and left-handed are correlated.

	Right-handed	Left-handed	Total
Male	43	9	
Female	44	4	
Total			

Sampling: consider the dataset of the cleaning exercise. Assume all mistakes in the data have been corrected. Do the lines 1, 2, 6 and 14 form a good sample of the original dataset? Justify shortly your answer.

Attribute (Dimensionality) reduction / Attribute (feature) creation: a dataset has a column DateOfBirth and a column Age. What do you think of this? Which attribute can be created from the other? Do they have the same type?

Consider the following table that shows three students and their results in 5 modules during a semester:

	M01	M02	M03	M04	M05
Student 1	E	1.0	2.3	5.0	NE
Student 2	1.7	3.3	E	E	1.7
Student 3	2.3	2.7	1.7	1.3	2.0

The value E means that a student enrolled that module without taking the exam while NE means that a student did not enroll in that course. The continuous values 1.0 till 5.0 are the results in the exam; 1.0 is the best value one can get, 4.0 is the least value (pass) and 5.0 means fail (German marks). One is interested in analysing the success of students. Propose at least two features that could be constructed from this table.

Attribute (Feature) subset selection: Read *3.4.4 Attribute subset selection p. 104-105* in the book. Explain how the chi square Test can be used in the Stepwise forward selection and in the Stepwise backward elimination.

Discretization and Binarisation: consider the golf dataset. What could have been the original recorded data of the variables Wind and Outlook? Which one has been discretized? Binarised? Propose a method to could achieve this discretization and binarisation.

Attribute transformation: suggest a transformation for the attribute Temperature so that the mean of the transformed attribute is 0. Tipp: see book p. 114.

Exercise 3: Correct the mistakes in the *golf* dataset and store it as a .csv file excluding the last attribute TemperatureI,or retrieve the original from Sample Data in RapidMiner. Read the data with RapidMiner (.csv).
Report the mode of Outlook, Wind and Play.
Report the min, max, average, median and standard deviation of Temperature and Humidity. Are these two attributes correlated? Tipp: use the Operator "Correlation Matrix". In RapidMiner visualize the scatter matrix, which gives a good idea whether two attributes could be correlated. Include this plot in your report and report also the value of the correlation of these two attributes.
Draw the Box Plot of either Temperature or Humidity with RapidMiner and give it in the report.
Create a new attribute TemperatureN which has the normalized values (z-score) of the attribute Temperature. Report the average and standard deviation of TemperatureN and the correlation of Temperature and TemperatureN.

Outlook	Temperature	Humidity	Wind	Play	TemperatureI
sunny	85	85	f	n	81-85
sunny	25	90	70km/h	n	21-25
cloudy	83	78	f	1	81-85
rain	70	96	f	y	66-70
rain	68	80	f	y	66-70
rain	65	70	t	n	61-65
cloudy	64	65	t	y	61-65
sunny	72	?	f	n	71-75
sunny	69	70	f	y	66-70
rain	75	80	f	y	71-75
sunny	75	70	t	y	71-75
cloudy	72	90	t	y	71-75
cloudy	81	75	f	y	81-85
rain	71-75	80	t	n	71-75
rain	71-75	80	t	n	71-75

Exercise 4: Same as Exercise 3 using Jupiter Notebook.