Data Analytics Foundations

Online Workshop 4

Data Preprocessing (1)

Questions from Module 3 so far?

KNIME – Preprocessing: missing values

- Download imports-mod.csv from the Workshop 4 page in Canvas and import it into KNIME.
- Use the Statistics to identify the missing values. Is there any obvious pattern? Why the missing values of price attribute are not detected by KNIME? Use String Manipulation node to fix the problem
- Use Missing Value Column Filter node to remove column with threshold of 60%? How many columns are removed? Why in some cases we need to remove the column with high missing values percentage.
- Use Missing Value node to handle the missing values. Numeric attributes replace missing values by mean and Nominal by the most frequent value

KNIME – Preprocessing: data quality

- Download census.csv from the Workshop 4 page in Canvas and import it into KNIME.
- Use Color Manager to colour the two values of Salary differently
- Add the Scatter Plot to show Fnlwgt on the x-axis and Education years on the y-axis. Can you see any outliers?
- Set up a Box Plot. Are there any obvious outliers in the numerical attributes? How would you deal with them? Discuss the options?
- Use the Numeric Outliers to remove outliers of Fnlwgt attributes.
- Add Box Plot to look into Fnlwgt after the treatment
- Is it always appropriate to remove outliers? What about interesting values? Look at Capital gain attribute?

Q&A question

1. When might you want to use normalization?

2. How would you handle a situation where an attribute “Age” was missing for 5 records out of 1000 in a dataset? What about for 50 records? 500?