CS909: 2013-14

Week 3: Data pre-processing

1. Use the R hist(), boxplot() and pairs() commands to explore the distributions of each non-categorical attribute in the iris dataset.

   Comment on what you observe, including outliers in the boxplot and the scatter plots of the attribute values.

2. Import the dataset irisMissing.csv into a data frame named irisMissing in your R workspace and use an R command to discover the row numbers of the instances that have missing values.

3. Identify an R command that will drop missing values. Apply it to the irisMissing dataset to create a new data frame irisDrop.

   Briefly describe three other strategies for handling missing values.

4. Identify or write your own R functions to implement each of these three strategies.

5. Write an R function `foo()` that takes a data frame and a missing value function as arguments and returns a new data frame with the missing values replaced with values as determined by the missing value function.

6. Use the hist() and boxplot() commands to compare results of applying each missing value strategy. Based on this, comment on their relative merits.

Save any figures you generate as PDFs and include them in your Tabula submission. Make sure they are clearly labeled.

Submission deadline: Midday, Thursday 30th January.