Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

Exercise Sheet 7 (January, 12, 2024)

Exercise 1 (Missing Value Imputation)

LearningTime	Grade	Semester
15	NA	7
50	1.0	2
42	2.3	1
27	3.0	3
40	1.3	5
29	2.0	2
23	NA	1
25	NA	4

You are conducting a study where you capture the hours spend preparing for an exam (*LearningTime*), the final grade (*Grade*), and the current semester (*Semester*) of students. However, for three students the grade is missing for some reason.

(a) For each missing data mechanism (MCAR, MAR and MNAR), provide an explanation whether it could be present here or not.
(b) Calculate the arithmetic mean of *Grade* and *Semester*. Use the **complete-case** analysis to deal with missing values.
(c) Use mean imputation to impute the missing values and report the results
(d) Now make use of linear regression imputation. Use the following linear model which has an R^2 of 0.41:

$$Grade = 4.95 - 0.07 \cdot LearningTime - 0.18 \cdot Semester$$

For each missing value, state the imputed value.
(d) For both imputation methods, describe why and/or why not they might produce suitable imputation results.

Exercise 2 (Dealing With Missing Values in R)
In this exercise, we will look at how we can implement missing value imputation in R. Load the `mtcars` dataset that we examined in the previous exercise. This dataset does not have any missing values, so we will first create some by deleting the `disp` column for the last 5 rows. Before deletion, store the original values in a separate variable. Then, perform the following tasks:

1. Use the following imputation methods to calculate the missing values:
   (a) Mean imputation

(b) Linear regression imputation using a fitted model that follows the formula $disp = \alpha \cdot mpg + \epsilon$
(c) Predictive mean matching using the same model
(d) Quadratic regression imputation using a fitted model that follows the formula $disp = \alpha \cdot mpg + \beta \cdot mpg^2 + \epsilon$

2. Compare the imputed values to the true values and plot the predicted values for each imputation method, using mpg as the x-axis and disp as the y-axis. Examine the results. Which of the methods achieved the best results?
3. Repeat the same analysis, but instead of removing the last 5 rows, remove all rows where $mpg < 16$

To fit the linear and quadratic regression models, you can use the R function `lm`.