9th May Group 42 Meeting Notes

Review of brief

Goal

Create a pipeline to impute missing values of a dataset, and evaluate the effectiveness of the approach.

Steps

PGDip (Part A)

Find a public dataset with a related research paper detailing an approach for handling missing instance values for a classification problem.
Use the dataset and create two imputed datasets, one using a mode category value (control) and another using Naive Bayes (test).
Choose two classification models to make predictions on the two imputed datasets (i.e. 2 models trained on 2 datasets each - 4 models), and generate applicable performance metrics.
Compare the performance of the two models between the two different imputation techniques, and write up the results.
Compare the performance differences with the techniques discussed in the research paper.

MEng (Part B)

Find a public dataset with a related research paper detailing an approach for handling missing instance values for a regression problem - same or different dataset to the one used in Part A.
Use the dataset and create two imputed datasets, one using a mean feature value (control) and another using k-NN regression (test).
Choose two regression models to make predictions on the two imputed datasets (i.e. 2 models trained on 2 datasets each - 4 models), and generate applicable performance metrics.
Compare the performance of the two models between the two different imputation techniques, and write up the results.
Compare the performance differences with the techniques discussed in the

research paper.
In addition, create a flow-chart explaining when to use each approach.

Present findings in a written report and video presentation per part, i.e. 2 videos, 2 reports.