Data Science Toolbox Portfolio Questions

04 Non-Parametrics and Missing Data

Daniel Lawson — University of Bristol

Block 4

Portfolio 04

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R04.1: Conduct a Nearest Neighbour imputation of the dataset presented in Workshop 4.3, plus any other imputation that you wish to try, and document it as part of your appendix. Then consider that you are tasked with "tidying up" this dataset for onwards analysis as part of an anonymised health data competition. What do you choose to present and why?

Question R04.2: Consider the paper "Kernel Methods in Machine Learning". Write a simple explanation suitable for Masters' level class in Data Science describing how a **polynomial kernel** would be used in this context, and explain its use case.

Question R04.3: Run the analysis of missing data described in the finalfit vignette from which our `colon` dataset came. Contrast their findings to those in the workshop. Make sure to document any work you do beyond the verbatim code, e.g. if you run their analysis with more variables or ours with fewer.