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: How should we choose k in neighest neighbor methods? Consider the paper "OPTIMAL CHOICE OF k FOR k-NEAREST NEIGHBOR REGRESSION". What is the problem with using a fixed k for k-NN problems? Implement the method described in the paper and apply it to the example data from Lecture 4.1.

Question R04.2: Outlier removal using HDBSCAN was discussed in Lecture 4.2 and is recommended for the Machine Learning task of Topic modelling (which we interpret later in the course).

Use the example given in the documentation and implement a few choices for outlier removal. (It only requres `pip install bertopic` to run verbatim in colab). Present a figure summarising the impact of the outlier removal on the topic model, and discuss the results.

Question R04.3: Consider the Workshop 4.3 on Missing Data and in particular Section 5 on Evaluation, which asks how we know imputation worked, and specifically: *Q. How do we know that imputation has done a good job? Q. If the imputation went badly, how wrong could our estimates be?*

Continue the analysis and try to answer these questions, by performing an evaluation of the performance of the imputation procedure. To do this, you will need to leave further data out as cross valudation, or make adversarial imputation. Include your code and any non-essential figures as an appendix.