Data Science Toolbox Portfolio Questions

03 Latent Structures, PCA, and Clustering

Daniel Lawson — University of Bristol

Block 3

Portfolio 03

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 R03.1: Imagine that you are trying to understand the Cyber Security data from Workshop 3.3 for the purposes of predicting whether traffic is "normal". Consider the advantages and disadvantages of enriching the feature set via a) dimensionality reduction, and b) clustering, for the purpose of passing to a classifier. You may wish to perform experiments (and cite results placed in your appendix) for this task.

Question R03.2: Describe the vanilla UPGMA (Average Linkage Clustering) algorithm and compare it to an efficient and more scalable approach, for example Sparse UPGMA, paying specific attention to how it can be made more efficient than $O(N^3)$.

Question R03.3: Read the documentation about how HDBSCAN works. Reflect on the importance of dimension in this for the construction of the nearest-neighbour step. You might want to refer to results in the literature such as "When is Nearest Neighbor meaningful?".