SM2001/FSM3001 DATA-DRIVEN METHODS IN ENGINEERING

Homework #2

Problem 1

The matrix in the dataset stored in `data.mat` contains results from tensile tests performed on paperboard specimens. The tests were carried out on different occasions and contain 30 tests in total. The problem is that after all the tests were performed it was realized that the humidity in the test room varied significantly during the day, a fact that affected the measurements since paperboard is a highly hygroscopic material. Your task is to see if the data can be grouped into distinct clusters which would correspond to the same humidity. The matrix $\boldsymbol{X}$ contains test data typically used by the paperboard industry. The first column is the strength [MPa], the second strain to failure [%], and the third column is the elastic modulus [GPa]. Please consider the tasks below:

a) Program the K-means method and apply it to the supplied data set. You should use a maximum of three clusters ($K = 3$). Make sure that the algorithm does not get stuck in one of the local minima and show evidence of that being addressed. Plot the color-marked features (stress vs strain) for the result of the clustering. *Hints:* Make sure that you normalize the features according to good machine-learning practice, and bring them back to the physical scale when visualizing the results. There is a good summary of the method and examples of implementation here: `https://youtu.be/Ev8YbxPu_bQ`

b) Apply the hierarchical-clustering algorithm for the same data set, which you can use in Matlab using the built-in functions `clusterdata`, `pdist`, `linkage` and `dendrogram`. Compare the results with the outcome of the K-means method and propose an explanation of the observed differences if any. Explain how one should interpret the dendrogram output for the given data.