Datamining & Statistics

Assignment 3: Clustering

Use the ***DMS – S6 - Assignment Template*** notebook that is available on Moodle to complete this assignment.

To get a 6: complete chapter 1 in the template

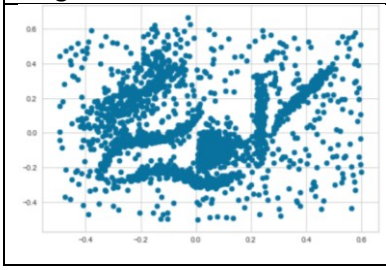
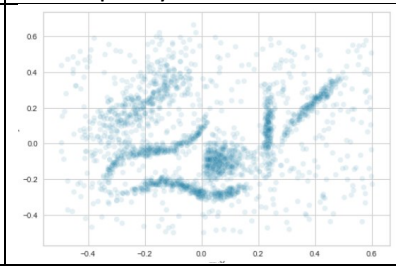
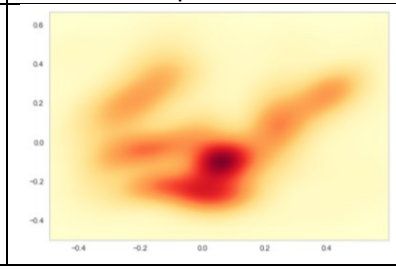
In the jupyter notebook from lecture 6, in Exercise 1 you have used the Soccer2019C.csv data to cluster the data based on 'Finishing' and 'HeadingAccuracy'. However, this was only done for the kmeans.

Create a jupyter notebook where you include the kmeans clustering on this, but also the other clustering techniques discussed. Evaluate the results by explaining why one technique is working better than others with this data.

As usual assume a student next year might get your notebook and will have to understand the steps you take, so add explanations along the way. This also goes for any of the additional steps described below.

+ 1 Visualisation: complete chapter 2 in the template

For larger datasets the scatterplot can look like a big blur. You can have a look at the UFC2019 dataset for this. Using some opacity can then sometimes resolve this. For even larger sets also the opacity is not very helpful. A heatmap can then perhaps be better suited. This often uses that famous Gaussian kernel. Add easy methods to show a scatterplot for each of these. Below is an example of each.

Original	With opacity	With heatmap
		

+1 Learn a new technique: complete chapter 3 in the template

Add one clustering techniques that is not already in the jupyter notebook from lecture 6 (anywhere, so also not in the appendix). Explain also briefly how that clustering technique works.
See for some inspiration one of the last slides ('Clustering techniques…'). The red ones are all options.

+1 Think of new types of data: complete chapter 4 in the template

What about categorical data. Figure out how to use clustering with categorical data.

+1 Elbow as number: complete chapter 5 in the template

Create a function that will return the k where the elbow is, so it can be used in a loop if I want to run multiple cluster analyses. Note that the KElbowVisualizer only visualizes the k, it does not return it's value.

Deliverables

To submit:

- Jupyter notebook file
- Dataset file(s) (or link to this in case it's too large)

Method to submit:

- One zip or rar file containing the files. Include your student number in the name of the zip/rar file.
- Upload on Moodle.

FAQ's

Q1: Can we work together?
A1: No and yes. No, it is an individual assignment, yes you can of course always help each other. But no direct copy-pasting. Copy-pasting, or copy-pasting and changing just one or two words, either from fellow students, or the internet, books, etc. is considered fraud. This will be reported to the examination board.

But…: But some lines of code will be standard, and only a few parameters or variable names will be different…..
A-but: The copy-pasting is more on the text surrounding the code, and indeed some parts of the code will be copy-pasted. This is acceptable and probably normal in IT that you copy some code from the internet and make a few adjustments to make it work for your data.

Q2: I'm not clear about the '+1', can you explain?
A2: Sure. There are some criteria to get a 6. To get to a 7 add one of the three things described at a '+1', your choice which one. To get to an 8 add two, to get to a 9 add three, and for a 10 add all four.