

## Unsupervised Learning

Implement k-means and AGNES clustering algorithms for the dataset.

The assignment file you will submit is assignment5.py. Please complete the methods within the class definitions provided, you can add helper methods and classes as well. All the code will be run through the run\_assignment5.py file. You can modify this file for your own testing, but **you can only upload the assignment5.py file** at the end so make sure all your code is in that file. You can expect that any variable in the run file could be changed during evaluation except for the data type which will be the same. The shape of the data could change though, there may be a different number of points or a different number of features. **If you don't hardcode values, you should be fine.**

For evaluation we will let the run file run for up to a minute, but you should really aim for under 10 seconds total (assuming an average modern laptop). Using NumPy vectors is not required but there are a lot of vector operations that you can run over data in NumPy and it will keep your code clean while massively improving performance.

Your code should not print anything to the console when you submit your assignment.

### Data:

The attached csv file contains all the data. The run file handles importing it and converting it to NumPy arrays.

### K-Means:

Use Euclidean distance between the features. Use a maximum number of iterations, t. Choose a k value and use k-means to split data in k clusters. The k value is provided to the k\_means class. Please implement train method, which should return an n-elements array (with n the number of data points in X) with the cluster id corresponding to each item.

The distance function is provided for you and you can assume all data is continuous. In case of a tie, you can pick one.

### AGNES:

Use the Single-Link method (distance between cluster a and b)=distance between closest members of clusters a and b) and the dissimilarity matrix.

For this exercise, use k as the number of clusters. Stop when number of clusters == k. The k value is provided to the AGNES class.

Use Euclidean distance between the features. The distance function is provided for you and you can assume all data is continuous. In case of a tie, you can pick one.