INFO 284, Spring 2018, Second Obligatory Group Assignment

Gaussian Mixture Models for Clustering.

Deadline for feedback May 1, 2018. Final deadline June 1, 2018 (Inspera)

Outline

The goal of this group project assignment is to obtain practical knowledge of the clustering algorithms by using the scikit-learn classifier. You will apply the Gaussian Mixture Models Clustering algorithm and the k-Means clustering algorithm on the same dataset and compare the results.

You can read about Gaussian Mixture Models and how to use them in scikit-learn here `http://scikit-learn.org/stable/modules/mixture.html`

Dataset

A dataset, named Seeds Dataset of unlabelled data is available at `https://archive.ics.uci.edu/ml/datasets/seeds`. Your task is to apply both the Gaussian Mixture Models and the k-Means Clustering algorithms on this dataset. Fine-tune the parameters of the algorithms, to the best of your ability, until you get clusters that you are happy with. Compare the results. You will also need to visualise the obtained clusters in a 2D scatter plot. You may need to reduce the dimensionality of the clustered data points for the purposes of visualisation.

What to submit

All documents should be compressed in a zip file.

- The code in a digital form. Do not submit the data sets.

- A small text file on how to run the code. A necessary condition for the assignment to be admissible is that the examiner can run the code (regardless of operating system). If the examiner is unable to run the code 0 points will be given for the entire assignment!
- Between 1000 - 5000 characters of report describing how the clustering algorithms comparatively performed on the task. State your opinions as to why the results were as such. The report should also include a visualisation of the clustered data in a 2D scatter plot. You may want to reduce the dimensionality of your data before you visualise it.