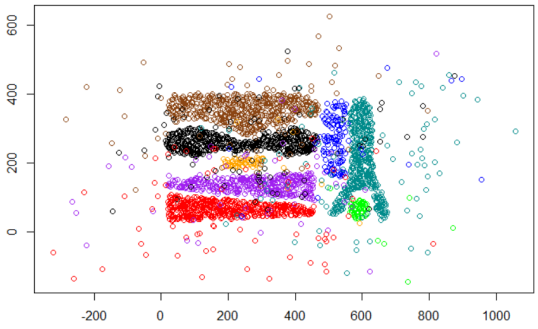
SIT384 Cyber security analytics

High Distinction Task 8.3HD: Comparing unsupervised clustering algorithms

Task description:



The standard `sklearn` clustering suite has thirteen different clustering classes alone. Sometimes we don't know what clustering algorithms we should be using. The answer is it depends on your data. A number of those thirteen classes in `sklearn` are specialised for certain tasks. If you know enough about your data, you can narrow down on the clustering algorithm that best suits that kind of data, or the sorts of important properties your data has, or the sorts of clustering you need done.

So far, we have introduced unsupervised clustering algorithms: K-Means, Agglomerative clustering, and DBSCAN.

In this task, you use all the above **three** clustering algorithms to cluster a given dataset **with different parameter settings** to find the BEST one for the given dataset.

You are given:

- a 2-dimensional dataset called Complex8_RN15, with is the variation of the Complex8 dataset with 15% gaussian noise added to the original Complex8 dataset.

  The Complex8_RN15 dataset has attributes x, y, class:

  The dataset is available in task resources zip file. It can also be obtained at: https://drive.google.com/file/d/1_geQDIMQUNHhc3d7zRxfOrOreTT3HBhU/view

- plot settings:

  fig,ax = plt.subplots(figsize=(7, 7), dpi=100)
  ax.scatter(…, alpha=0.25, s=60, linewidths=0)

- Other settings of your choice

You are asked to use:

- K-Means with proper parameters to predict clusters and visualize them using plot
- Agglomerative clustering with proper parameters to predict clusters and visualize them using plot

- DBSCAN with proper parameters to predict clusters and visualize them using plot
- create a plot using original class (y) for comparison

You must compare and interpret the results of using different approaches for the dataset. Other requirements are:

- In your report after comparing the experimental results, write a paragraph or two trying to explain/speculate why, in your opinion one clustering algorithm outperformed the others.
- Include a brief discussion in your report, how you have selected the parameters of particular algorithms.
- Finally, at the end of your report provide a 1-2 paragraphs summary that summarizes the most important findings of this task

The github website of the prescribed textbook has quite some useful supplemental material (**code examples**, IPython notebooks, etc.), available at https://github.com/amueller/introduction_to_ml_with_python, especially chapter 3.

Submission:

Submit the following files to OnTrack:

1. Your program source code (e.g. task8_3.py)
2. A screen shot of your program running
3. Your result analysis report

Check the following things before submitting:

1. Add proper comments to your code