Exercises 7: Clustering

Introduction to data science and Python programming for medical research (2025)

If you have not done so already you need to install the sklearn package. For windows users this would require you to open your conda terminal, found in the start menu, and then typing: `conda install sklearn`

1 Blob data

- Generate 2D ***blob*** data and test one or more of the methods available in sklearn of your choice to cluster it.
- Plot the generated data points using `scatter` plots. Use two subplots, one plot for the ground truth labels and another to show the labels assigned by the cluster method.

```
from sklearn.datasets import make_blobs
n_samples = 500
n_centers = 4
X, y = make_blobs(n_samples=n_samples,
                  n_features = 2,
                  centers=n_centers,
                  random_state=0)
```

See the documentation for the data generation at:
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html

2 Moon data

- Generate 2D ***moon*** data and test one or more of the methods available in sklearn of your choice to cluster it.
- Test spectral clustering and try different settings for `affinity` and `n_neighbors`. Does the method always successfully cluster the data?
- Plot the generated data points using `scatter` plots. Use two subplots, one plot for the ground truth labels and another to show the labels assigned by the cluster method.

```python
from sklearn.datasets import make_moons
n_samples = 500
X, y = make_moons(n_samples=n_samples,
                  noise=0.05,
                  random_state=0)
```

See the documentation for the data generation at:
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html

See the documentation for the SpectralClustering:
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html

3 Circular data

- Generate 2D ***circle*** data and test one or more of the methods available in sklearn of your choice to cluster it.
- Plot the generated data points using `scatter` plots. Use two subplots, one plot for the ground truth labels and another to show the labels assigned by the cluster method.

```python
from sklearn.datasets import make_circles
n_samples = 500
X, y = make_circles(n_samples=n_samples,
                    factor=0.5,
                    noise=0.05,
                    random_state=0)
```

See the documentation for the data generation at:
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_circles.html