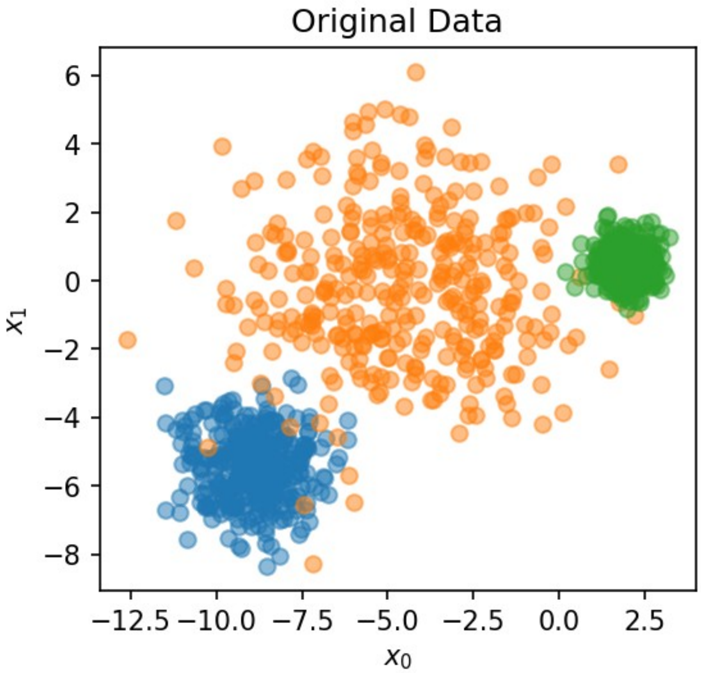
M11-L1 Problem 3

In this problem you will use the `sklearn` implementation of hierarchical clustering with three different linkage criteria `('single', 'complete', 'average')` to clusters two datasets: a "blob" shaped dataset with three classes, and a concentric circle dataset with two classes.

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn.datasets import make_blobs, make_circles
from sklearn.cluster import AgglomerativeClustering

## DO NOT MODIFY
def plotter(x, labels = None, ax = None, title = None):
    if ax is None:
        _, ax = plt.subplots(dpi = 150, figsize = (4,4))
        flag = True
    else:
        flag = False
    for i in range(len(np.unique(labels))):
        ax.scatter(x[labels == i, 0], x[labels == i, 1], alpha = 0.5)
    ax.set_xlabel('$x_0$')
    ax.set_ylabel('$x_1$')
    ax.set_aspect('equal')
    if title is not None:
        ax.set_title(title)
    if flag:
        plt.show()
    else:
        return ax
```

First we will consider the "blob" dataset, generated below. Visualize the data using the provided `plotter(x, labels)` function.

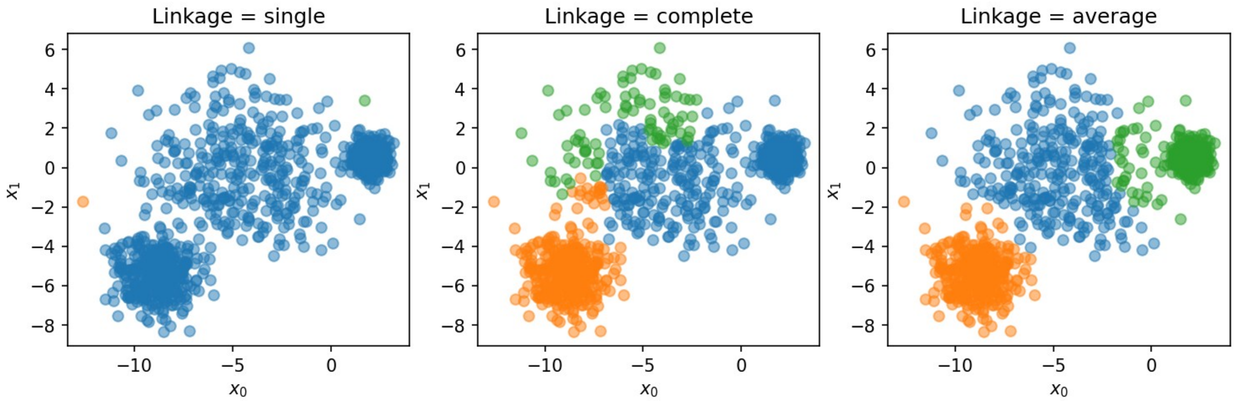
```
## DO NOT MODIFY
x, labels = make_blobs(n_samples = 1000, cluster_std=[1.0, 2.5, 0.5],
random_state = 170)

## YOUR CODE GOES HERE
# visualize the data
plotter(x, labels, title = 'Original Data')
```



Using the `AgglomerativeClustering()` function, generate 3 side-by-side plots using `plt.subplots()` and the provided `plotter(x, labels, ax, title)` function to visualize the results of the following three linkage criteria `['single', 'complete', 'average']`.

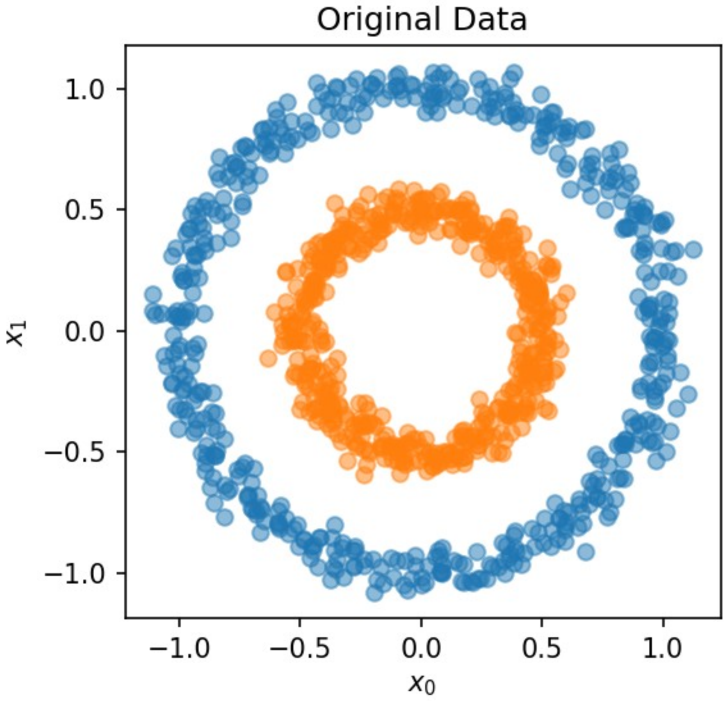
Note: the `plt.subplots()` function will return `fig, ax`, where `ax` is an array of all the subplot axes in the figure. Each individual subplot can be accessed with `ax[i]` which you can then pass to the `plotter()` function's `ax` argument.

```
## YOUR CODE GOES HERE
# 3 subplot using agglomerative clustering for 3 linkage criteria (single, complete, average)
fig, ax = plt.subplots(1, 3, dpi = 150, figsize = (12,5))
for i, linkage in enumerate(['single', 'complete', 'average']):
    model = AgglomerativeClustering(n_clusters = 3, linkage = linkage)
    model.fit(x)
    plotter(x, model.labels_, ax = ax[i], title = f'Linkage = {linkage}')
plt.show()
```



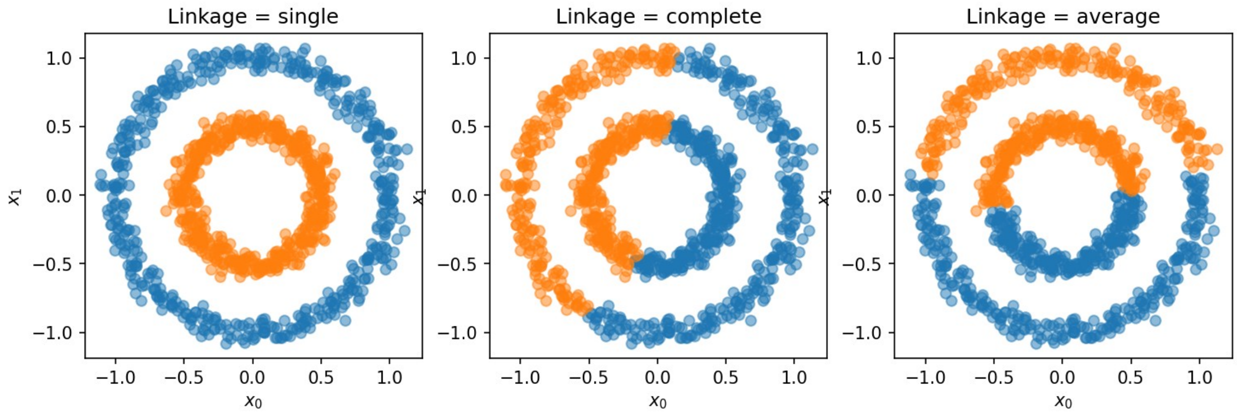
Now we will work on the concentric circle dataset, generated below. Visualize the data using the provided `plotter(x, labels)` function.

```
## DO NOT MODIFY
x, labels = make_circles(1000, factor = 0.5, noise = 0.05,
random_state = 0)

## YOUR CODE GOES HERE
# visualize the data
plotter(x, labels, title = 'Original Data')
```



Again, use the `AgglomerativeClustering()` function to generate 3 side-by-side plots using `plt.subplots()` and the provided `plotter(x, labels, ax, title)` function to visualize the results of the following three linkage criteria `['single', 'complete', 'average']` for the concentric circle dataset.

```
## YOUR CODE GOES HERE
# 3 subplot using agglomerative clustering for 3 linkage criteria (single, complete, average)
fig, ax = plt.subplots(1, 3, dpi = 150, figsize = (12,5))
for i, linkage in enumerate(['single', 'complete', 'average']):
    model = AgglomerativeClustering(n_clusters = 2, linkage = linkage)
    model.fit(x)
    plotter(x, model.labels_, ax = ax[i], title = f'Linkage = {linkage}')
plt.show()
```



Discussion

Discuss the performance of the three different linkage criteria on the "blob" dataset, and then on the concentric circle dataset. Why do some linkage criteria perform better on one dataset, but worse on others?

In the blob dataset, the average linkage criteria performed the best in data clustering. The complete linkage criteria did classify some clusters but the outcome is not satisfactory compared to the ground truth dataset. The single linkage criteria performed the worst in data clustering.

In the concentric circle dataset, the single linkage criteria performed the best in data clustering. On the other hand, the complete and average linkage criteria did classify the dataset but the outcome is not correct when compared to the ground truth dataset.

The performance of each linkage criterion will depend on the shape and distribution of the cluster in the given dataset. The single linkage criteria perform better on elongated clusters, while the complete linkage criteria perform better on compact clusters. These unique properties depend on the cluster distance calculations.