Machine Learning – Assignment 5 - Coding

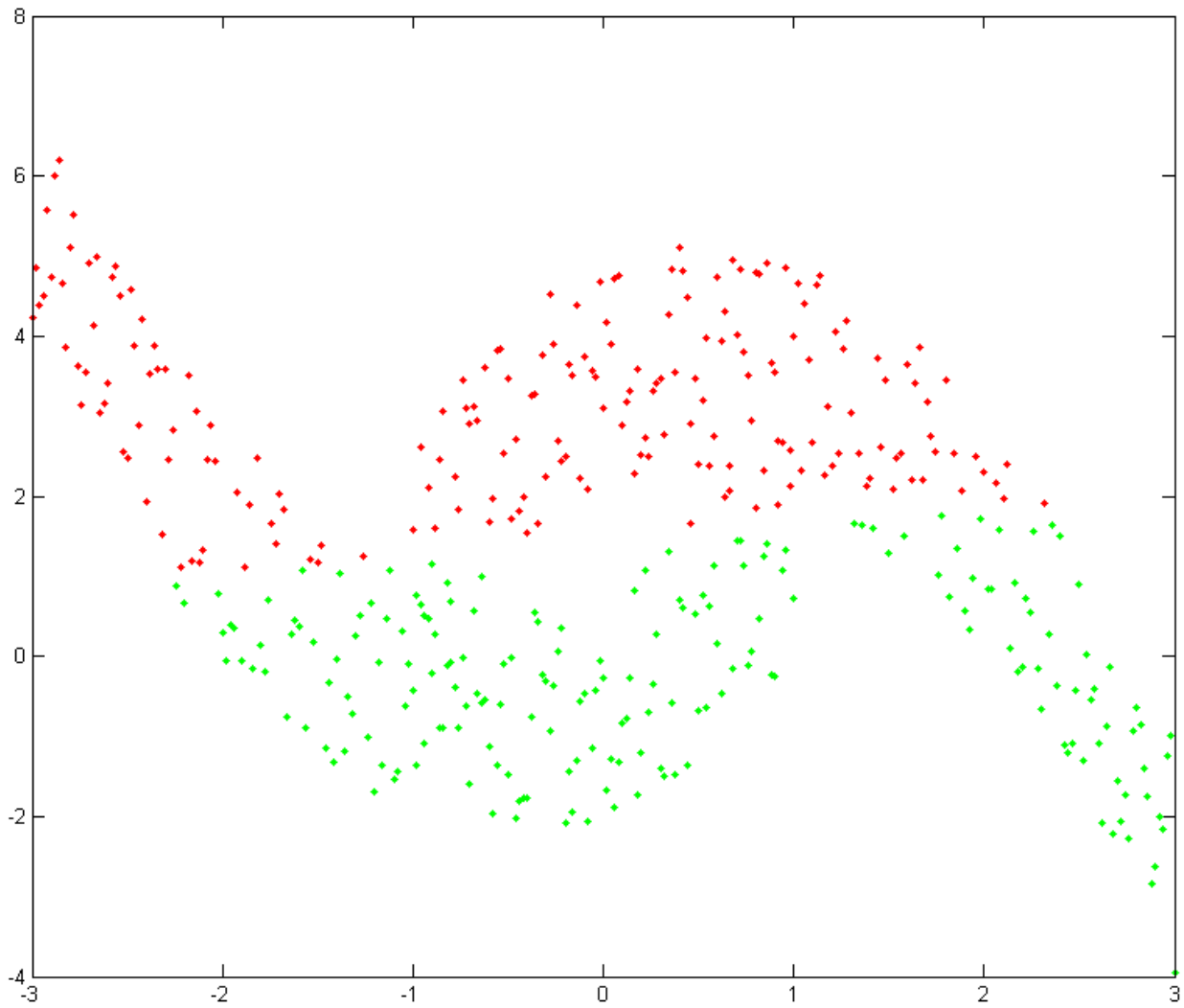
103062528 林玉山

1. 實驗結果

(a) K-Means Clustering

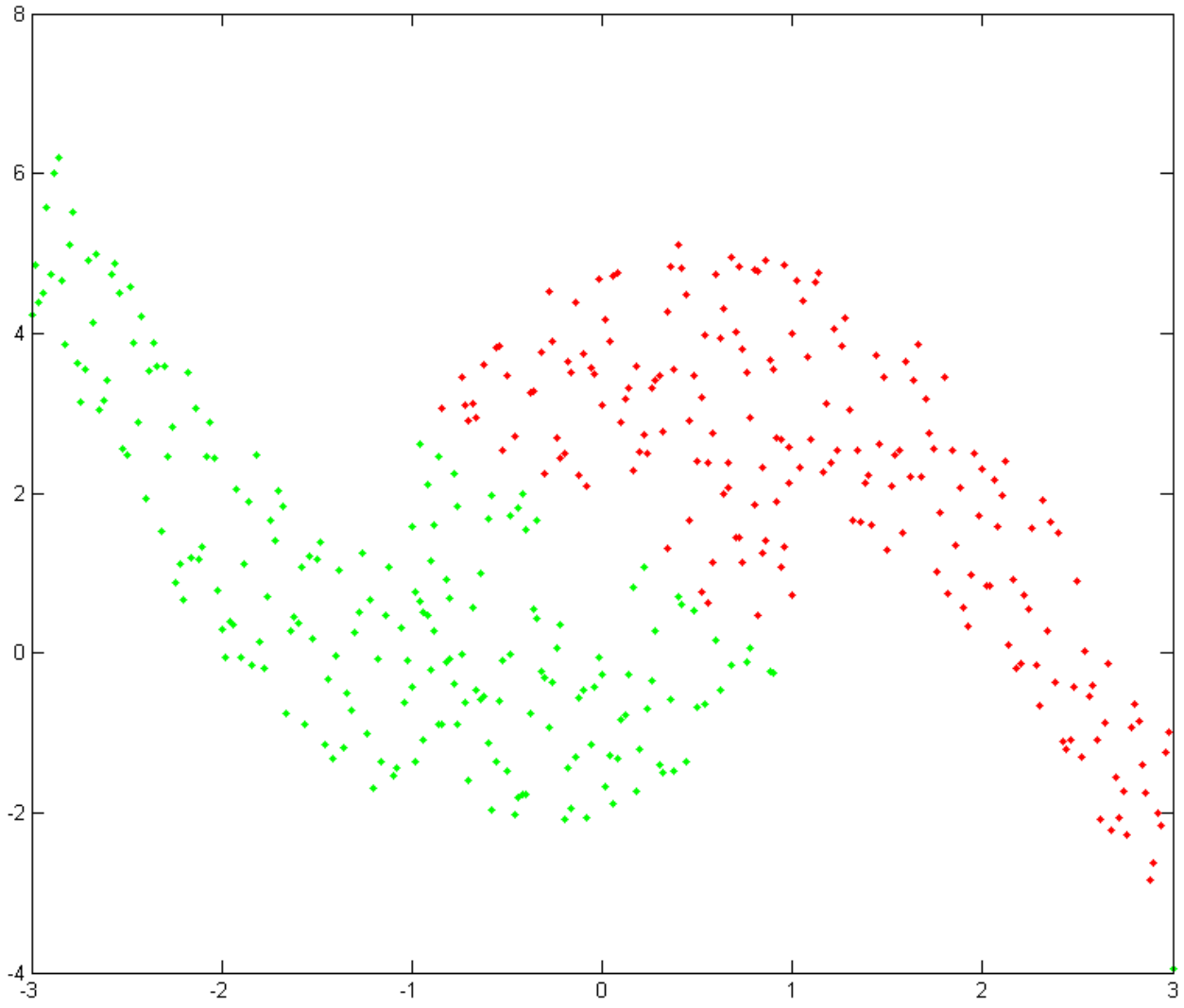
第一種結果

Quality: 0.1325



第二種結果

Quality: 1.0716



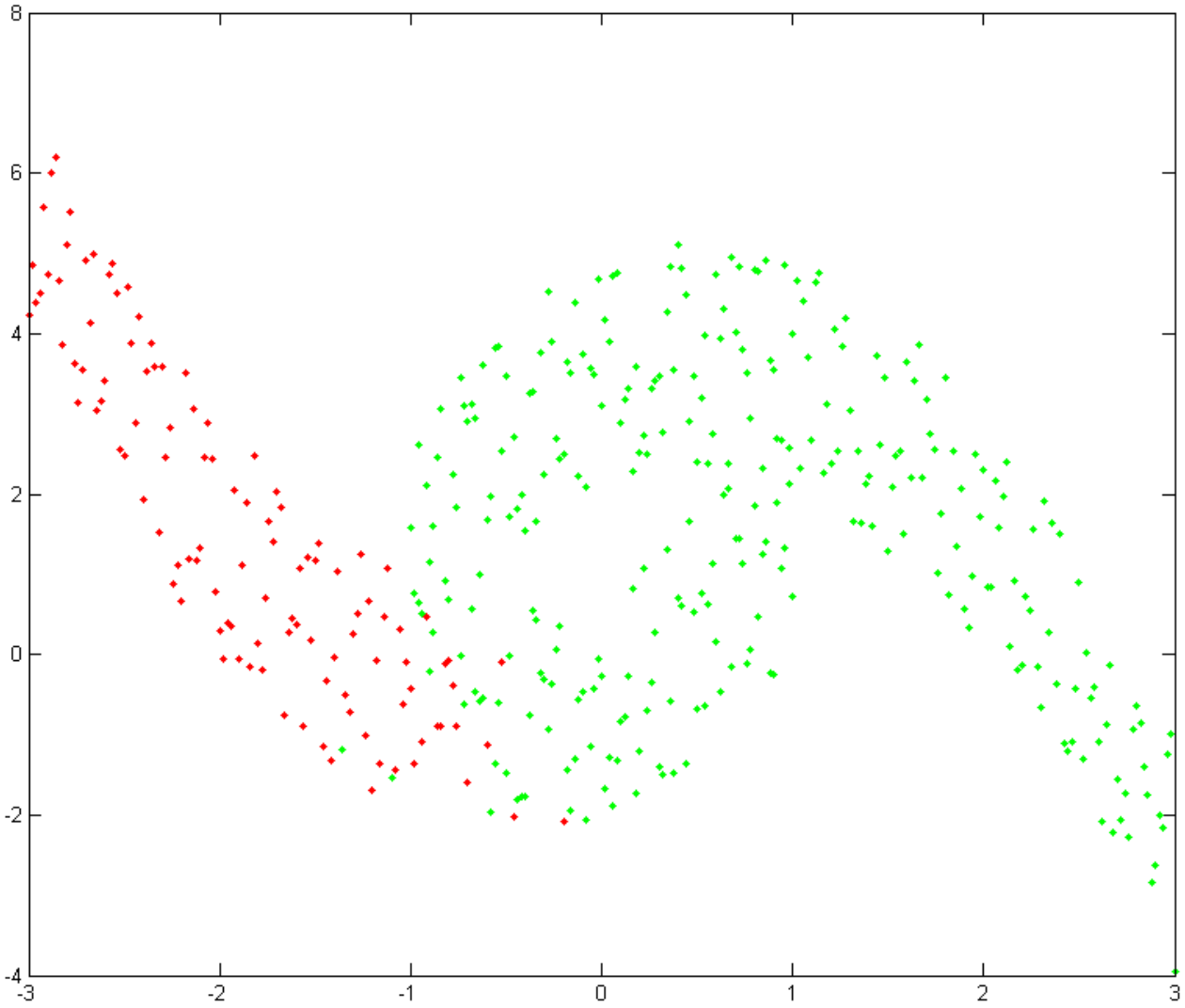
小結：

從上面的結果看的出來，k-means 的結果不大穩定。事實上，大部分的結果都偏向於第一種，第二種的出現則需要一點運氣。另外，第一種結果比較不像人類的直覺，反倒第二種比較像是人類會看出來的分群。

(b) Spectral Clustering with $\boldsymbol{\epsilon - NN}$ similarity matrix

Hyperparameter: $\epsilon = 80$

Quality: 1.3270



小結：

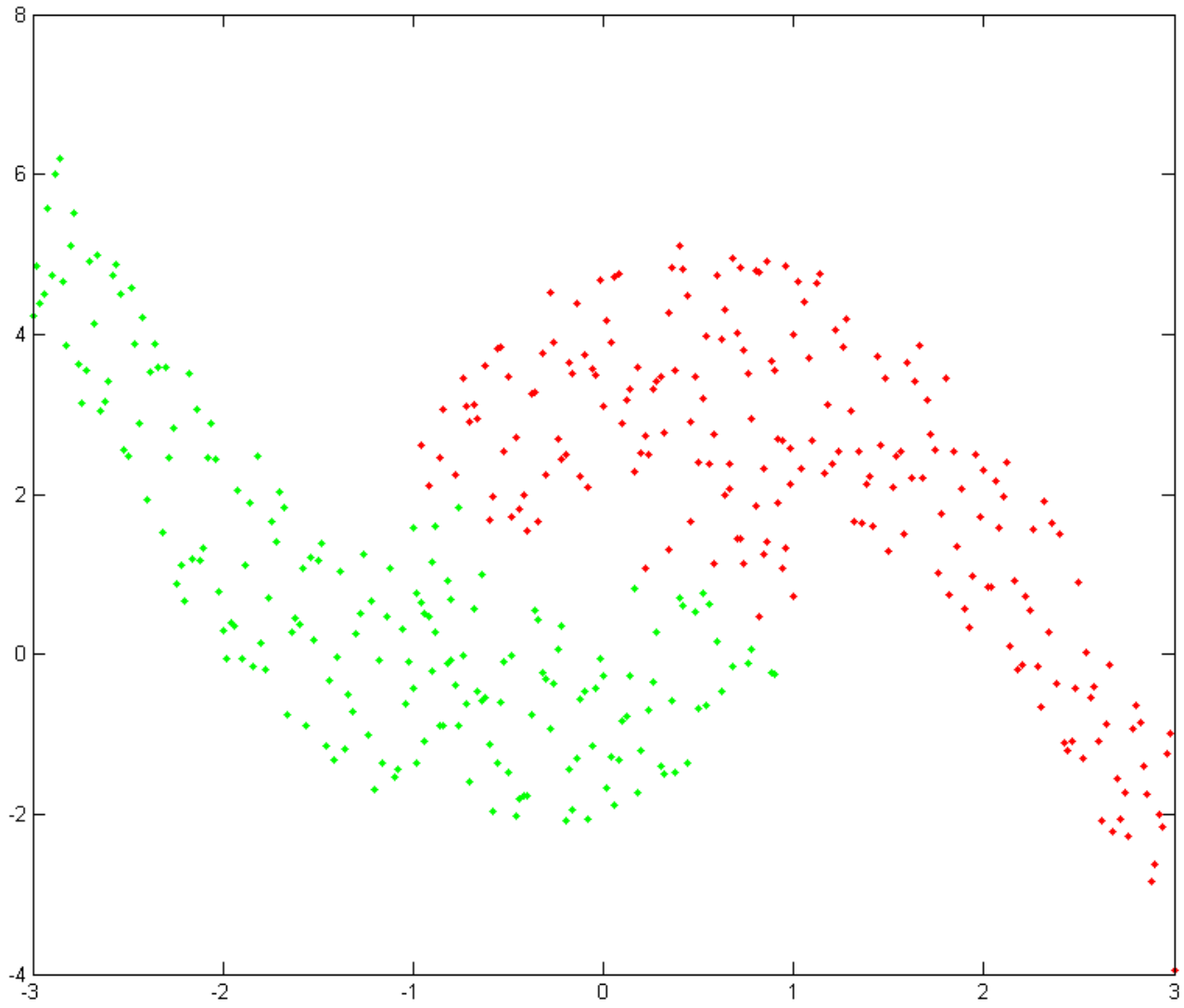
首先必須要找到合適的 hyperparameter。找到後，大概能夠分出最好的結果就是上圖。比起 K-means 穩定許多，沒有出現其他分法。只是這個分出來的結果尚有點差強人意。

(c) Spectral Clustering with $\boldsymbol{\epsilon - Ball}$ similarity matrix

Hyperparameter: $\epsilon = 1$

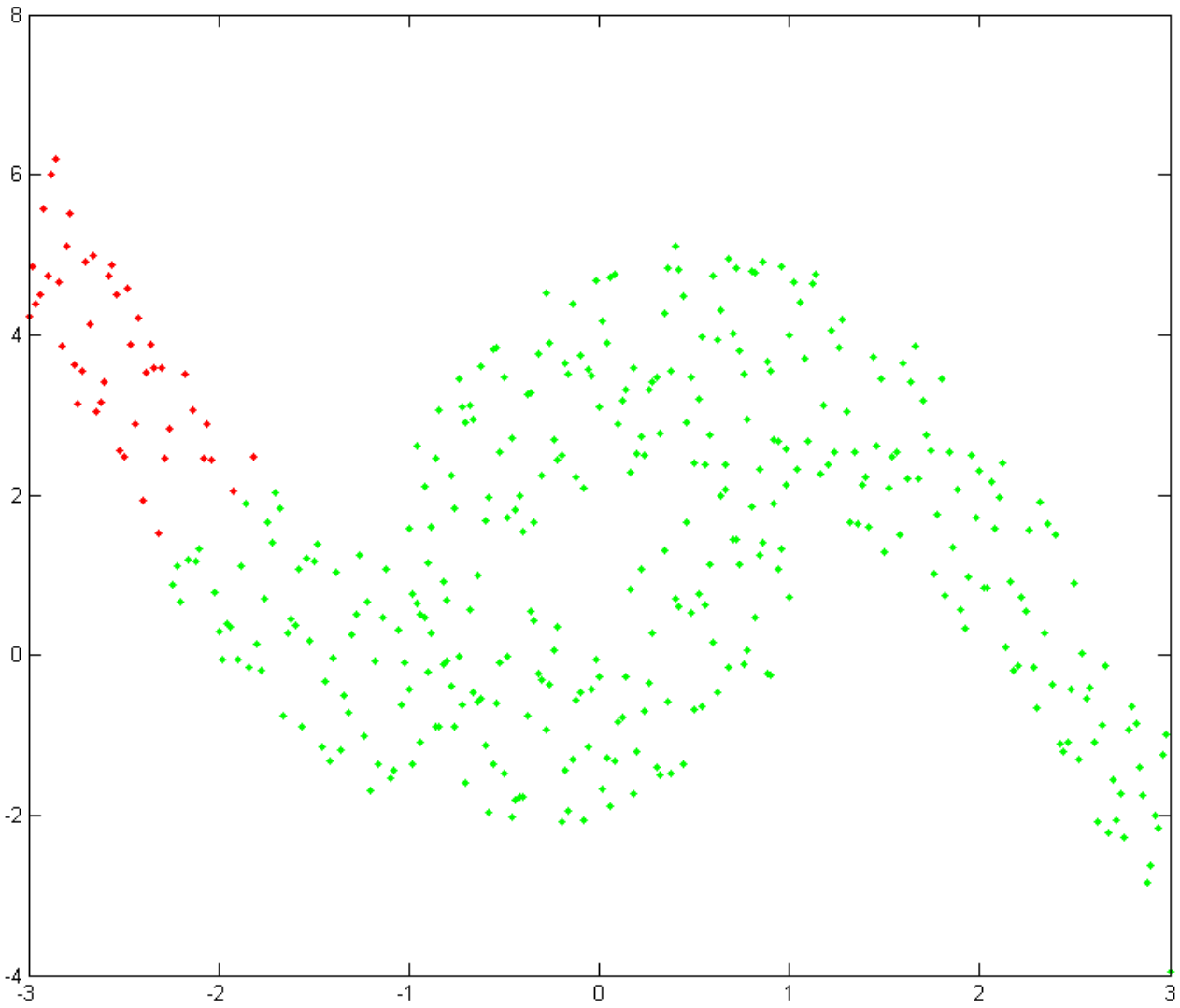
第一種結果

Quality: 1.0217



第二種結果

Quality: 2.4294



小結：

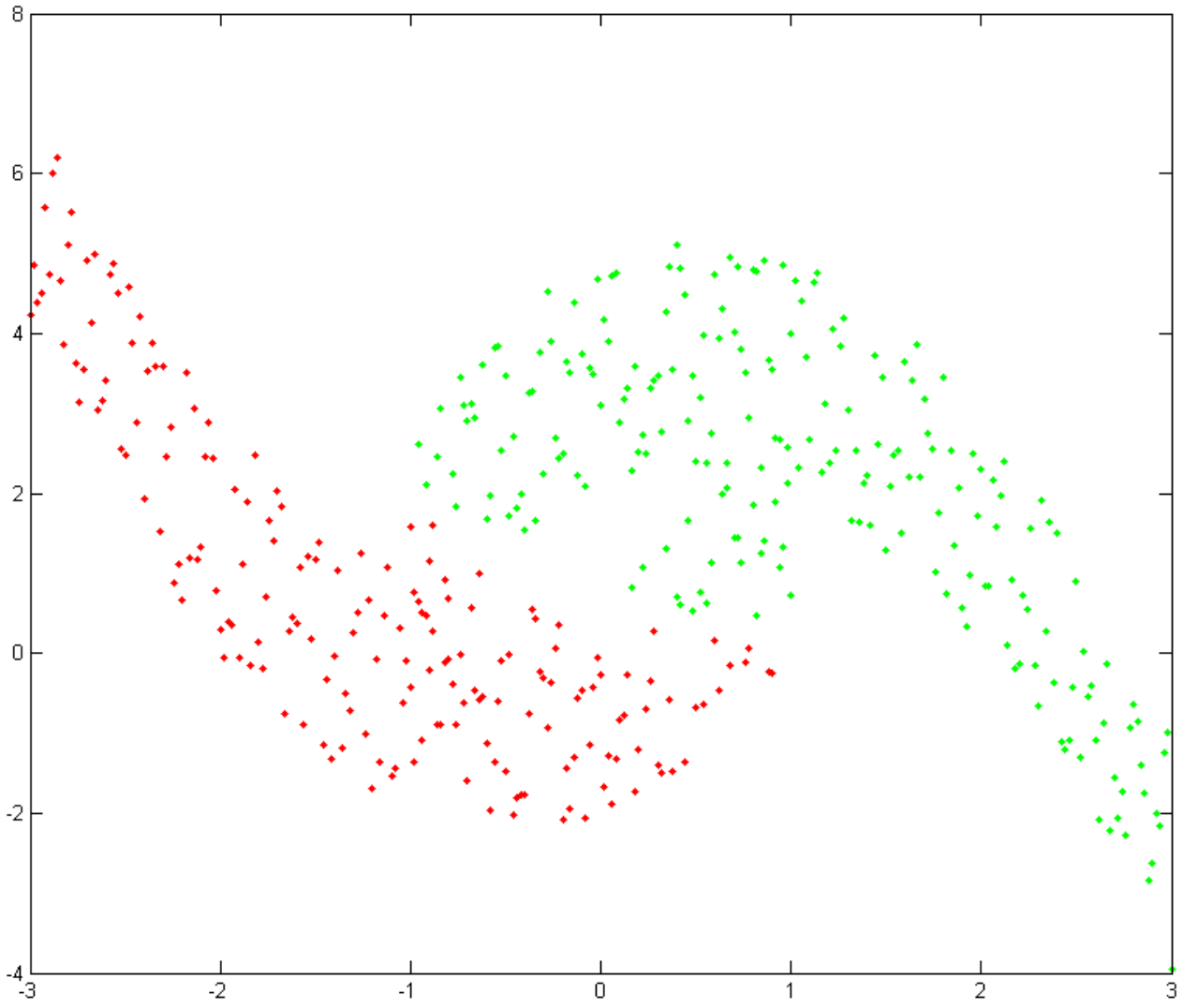
$\boldsymbol{\epsilon - Ball}$ 的表現比上一個好很多，大多可以分出第一種結果，也是人類直覺的分群。但是在少數的情況下，會分出第二種結果。此時只有一小部分被分進同一群，大多數被另一群囊括。這是由於執行 k-means 挑選初始點仍有極小的機率挑到不好的點所造成的。

(d) Spectral Clustering with **Gaussian** similarity matrix

Hyperparameter: $\sigma = 0.5$

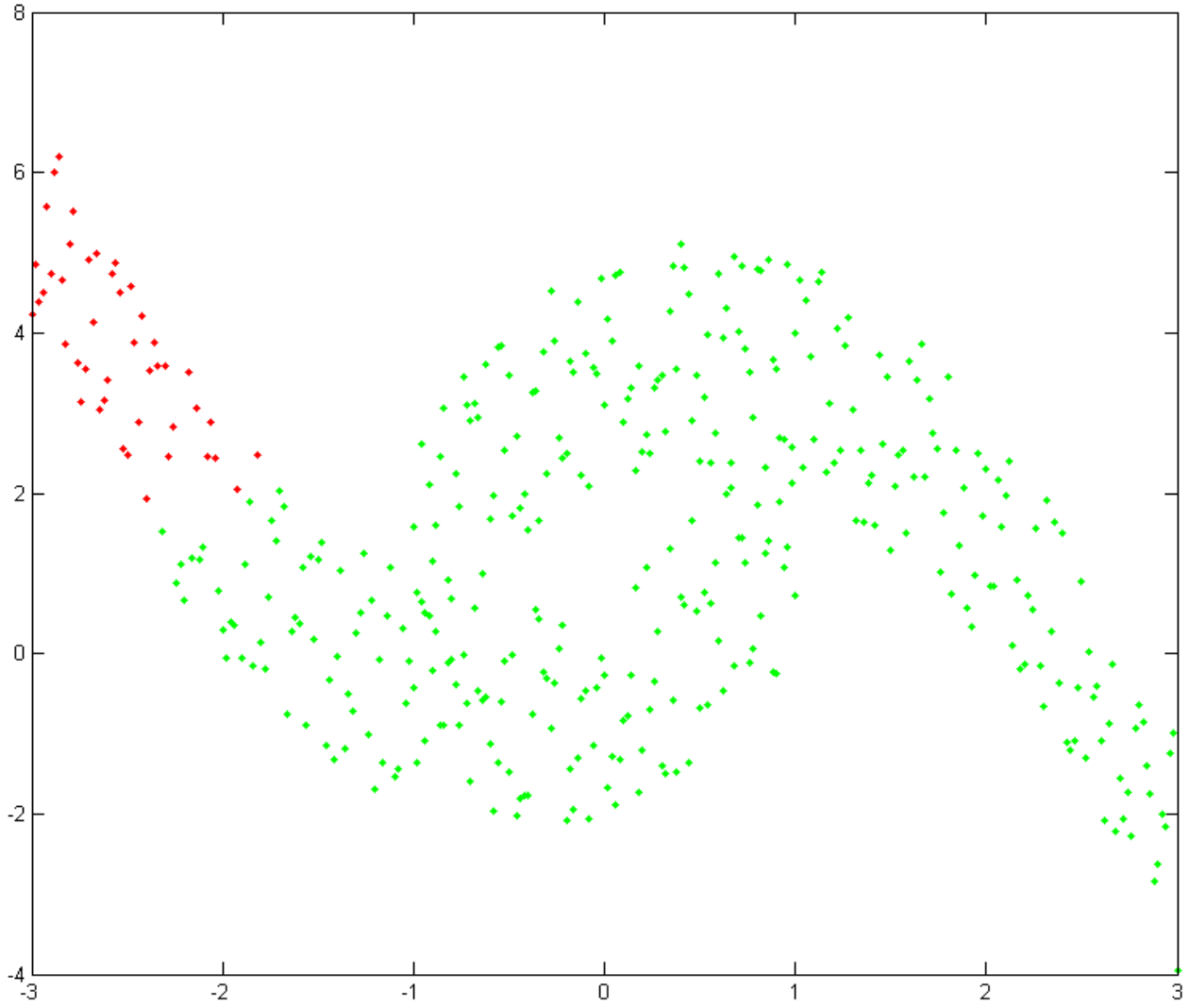
第一種結果

Quality: 1.0443



第二種結果

Quality: 2.4568



小結：

Gaussian 的結果與 $\boldsymbol{\epsilon - Ball}$ 類似，常出現的結果有上面兩種，但是第二種出現的頻率較之前高。撇出這種特殊狀況不談，可以看到 Gaussian 分的結果也是很接近人類的直覺。

2. 討論

(1) 分群結果討論

從上述的結果可以看出來，k-means 單純地找出平均點，並以距離來分群的效果很有限。因為對於這種類似吸盤狀的分布資料，只憑找平均點的話，就容易分出 k-means 的第一種結果。另一端的尾端因為距離較近，儘管並沒有直接的連結，也容易分到同一群。這是表現出了 k-means 的缺點。

相較之下，spectral clustering 就好很多。不過可以看的出來，$\boldsymbol{\epsilon - NN}$ 的效果也是有限。雖然可以根據點之間連接的狀況找到附近的點，但是仍無法完全地分開。另外兩個則表現得較符合預期的結果。

(2) Quality Evaluation 討論

從實驗結果，以及計算出來的 quality 值來看，這並不是一個好的 measurement。以 Gaussian 的兩種分群結果為例。可以明顯地看出，第一種比較是我們想要的結果，但是第二種分數卻比較高。我想原因在於，這個 evaluation 仍是單純地以每一個點的距離與平均點的距離作為衡量標準。

每一個 cluster 內的點距離 mean 越近，這個分數就會越高。但是並非每一個 cluster 都是相似的大小，而且也並非是以 mean 為中心散出。另外，兩個 cluster 之間的距離越遠，分數也會越高。這個也是不一定的。因此在某些狀況下，像是這組資料，就不是一個好的量測方式。