Machine Learning Homework 2 Report

一、 K-Means

1. Noah Dataset Using x, y as Features

我們這組使用 python 作為撰寫 k-means 模型以及處理資料的語言，並利用 seaborn 和 matplotlib 作為視覺化的工具。關於 K-means 的模型設計使用 K=3，而使用的距離公式為 L2 範數(L2 norm)也就是歐幾里得距離(Euclidean distance)因為希望將這些散布在由 x, y 組成的二維特徵空間的樣本分成三類，也就是對應到 pitch type 的三個種類：CH、CU、FF。

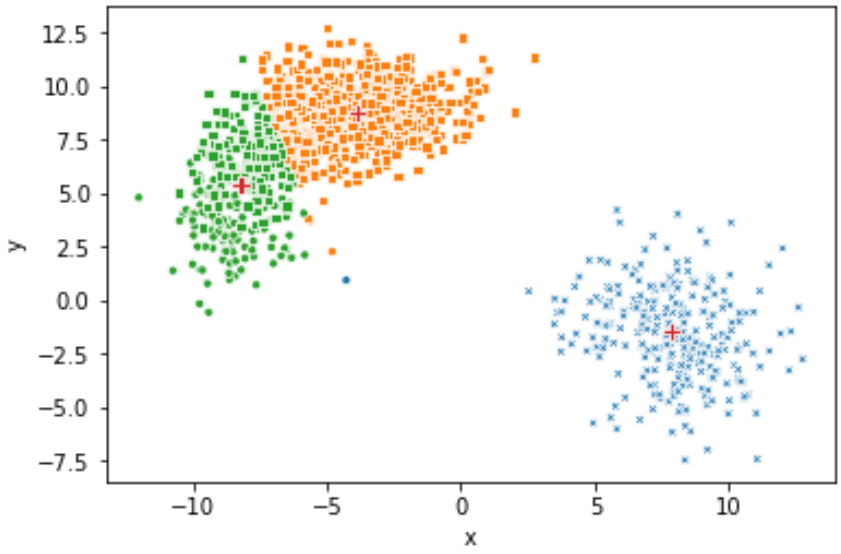


Fig.1 Result of clustering by 3-means model

紅色的點代表了最終穩定下來的 centroid，也就是每個 cluster 的平均所在位置，在實作 code 裡面如果 centroid 移動的距離收縮到 10^{-26} 時我們就認為已經分類完成，在原版的 code 中須完全不移動才算分類完成，但是有可能導致收斂時間過長，個人認為在 clustering 中少許一兩個邊界的樣本，並不會影響模型的表現。

接著，我們衡量每個 cluster 的 Shannon entropy 來推估這個分類的好壞。

$$\text{Total Entropy} = \sum_{c \in clusters} \sum_{type \in c} -P(type) \times \lg(P(type)) \quad (1)$$

c 及代表每個類別，type 則是 c 所含的 pitch type，lg 為 $\log_2$，P(type)是在 c 中選到樣本的 pitch type 的機率。最好的情況也就是每個 cluster 剛好只含有一種 pitch type，根據(1)可以得到 Total Entropy 為 0。

換言之，當我們用 entropy 衡量 clustering 的時候，值越大表示表現越差，

因為代表有越大的亂度，反之，值越小表示表現越好，且 Total Entropy 大於等於 0。

Table1. Total Entropy of K-means with Different Features

Features	Total Entropy
x, y	0.587225
x, y, speed	1.382383
x, y, speed, az	1.535960

可以發現，當我們增加越來越多特徵時，表現反而越來越差。我個人認為有兩個可能，一個是在高維空間中或許歐幾里得距離並不是一個好的衡量方式，例如：海苔捲形狀的分布，第二個是這些 features 並不是很好，沒辦法區分出三個種類的球。

二、 kd-tree

1. Model

在這邊我們使用了 dictionary 來建立這棵樹，並利用遞迴畫出 2d -tree 的樣子。

```
{'left': {'left': {'left': {'left': None, 'point': [0, 2], 'right': None
},
                   'point': [2, 1],
                   'right': None},
          'point': [1, 3],
          'right': {'left': {'left': None, 'point': [0, 5], 'right': Non
e},
                    'point': [3, 4],
                    'right': None}},
 'point': [4, 2],
 'right': {'left': {'left': {'left': None, 'point': [6, 0], 'right': Non
e},
                    'point': [7, 1],
                    'right': None},
           'point': [6, 3],
           'right': {'left': None, 'point': [5, 5], 'right': None}}}
```

Fig.2 The structure of kd-tree

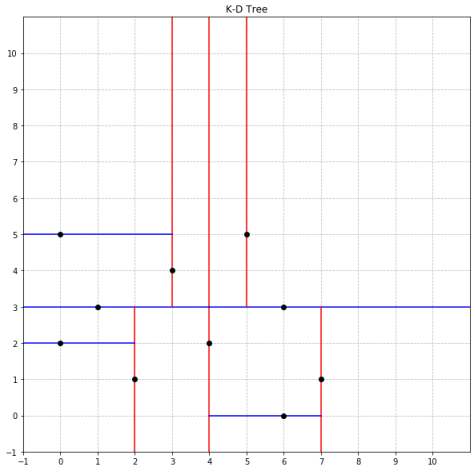
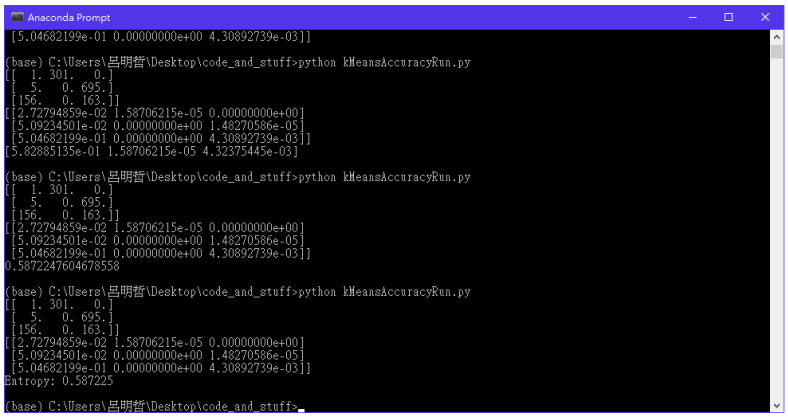
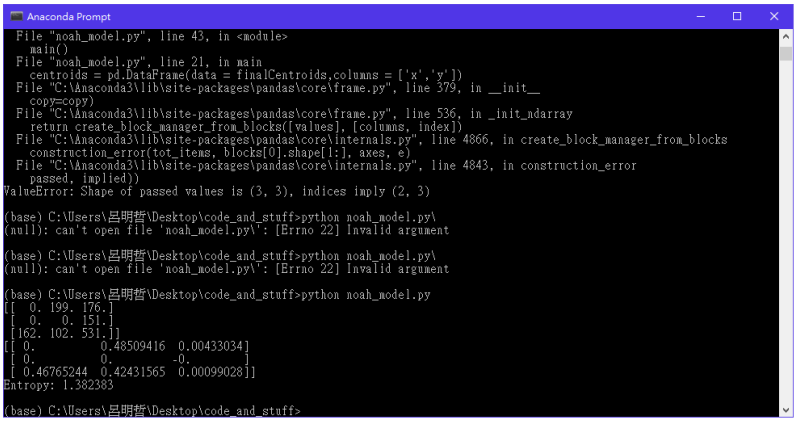
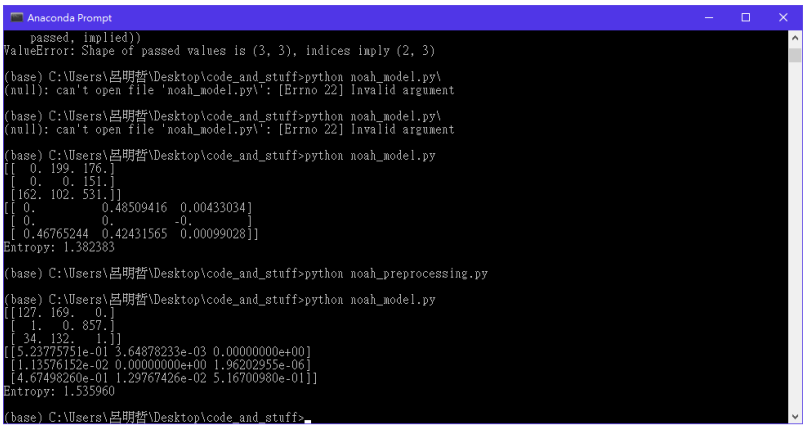


Fig.3 Visualization of the 2d-tree

附錄


Anaconda Prompt
[5.04682199e-01 0.00000000e+00 4.30892739e-03]]
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python kMeansAccuracyRun.py
[[1. 301. 0.]
[5. 0. 695.]
[156. 0. 163.]]
[[2.72794859e-02 1.58706215e-05 0.00000000e+00]
[5.09234501e-02 0.00000000e+00 1.48270586e-05]
[5.04682199e-01 0.00000000e+00 4.30892739e-03]]
[5.82885135e-01 1.58706215e-05 4.32375445e-03]
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python kMeansAccuracyRun.py
[[1. 301. 0.]
[5. 0. 695.]
[156. 0. 163.]]
[[2.72794859e-02 1.58706215e-05 0.00000000e+00]
[5.09234501e-02 0.00000000e+00 1.48270586e-05]
[5.04682199e-01 0.00000000e+00 4.30892739e-03]]
0.5872247604678558
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python kMeansAccuracyRun.py
[[1. 301. 0.]
[5. 0. 695.]
[156. 0. 163.]]
[[2.72794859e-02 1.58706215e-05 0.00000000e+00]
[5.09234501e-02 0.00000000e+00 1.48270586e-05]
[5.04682199e-01 0.00000000e+00 4.30892739e-03]]
Entropy: 0.587225
(base) C:\Users\呂明哲\Desktop\code_and_stuff>


Anaconda Prompt
File "noah_model.py", line 43, in <module>
main()
File "noah_model.py", line 21, in main
centroids = pd.DataFrame(data = finalCentroids,columns = ['x','y'])
File "C:\Anaconda3\lib\site-packages\pandas\core\frame.py", line 379, in __init__
copy=copy)
File "C:\Anaconda3\lib\site-packages\pandas\core\frame.py", line 536, in _init_ndarray
return create_block_manager_from_blocks([values], [columns, index])
File "C:\Anaconda3\lib\site-packages\pandas\core\internals.py", line 4866, in create_block_manager_from_blocks
construction_error(tot_items, blocks[0].shape[1:], axes, e)
File "C:\Anaconda3\lib\site-packages\pandas\core\internals.py", line 4843, in construction_error
passed, implied))
ValueError: Shape of passed values is (3, 3), indices imply (2, 3)
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py\
(null): can't open file 'noah_model.py\': [Errno 22] Invalid argument
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py\
(null): can't open file 'noah_model.py\': [Errno 22] Invalid argument
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py
[[0. 199. 176.]
[0. 0. 151.]
[162. 102. 531.]]
[[0. 0.48509416 0.00433034]
[0. 0. -0.]
[0.46765244 0.42431565 0.00099028]]
Entropy: 1.382383
(base) C:\Users\呂明哲\Desktop\code_and_stuff>


Anaconda Prompt
passed, implied))
ValueError: Shape of passed values is (3, 3), indices imply (2, 3)
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py\
(null): can't open file 'noah_model.py\': [Errno 22] Invalid argument
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py\
(null): can't open file 'noah_model.py\': [Errno 22] Invalid argument
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py
[[0. 199. 176.]
[0. 0. 151.]
[162. 102. 531.]]
[[0. 0.48509416 0.00433034]
[0. 0. -0.]
[0.46765244 0.42431565 0.00099028]]
Entropy: 1.382383
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_preprocessing.py
(base) C:\Users\呂明哲\Desktop\code_and_stuff>python noah_model.py
[[127. 169. 0.]
[1. 0. 857.]
[34. 132. 1.]]
[[5.23775751e-01 3.64878233e-03 0.00000000e+00]
[1.13576152e-02 0.00000000e+00 1.96202955e-06]
[4.67498260e-01 1.29767426e-02 5.16700980e-01]]
Entropy: 1.535960
(base) C:\Users\呂明哲\Desktop\code_and_stuff>