Homework-1 (Chapter 3. Classification)

110321006 蔡秉翰

110321007 劉竑毅

為了使用 SGDClassifier 進行 MNIST 數據集的多類別分類，首先進行資料的初步整理，將 70000 筆資料分成了 60000 和 10000，其中 60000 用來訓練模型，10000 用來測試，先用 SGDClassifier **訓練出模型後，**使用交叉驗證評估模型性能（cv=3）來測量準確性，最後在 10000 筆資料的測試集上進行測試得到了以下結果

```
Cross-Validation Scores: [0.87365 0.85835 0.8689 ]
Mean Accuracy: 0.8669666666666668
Test Set Accuracy: 0.874
```

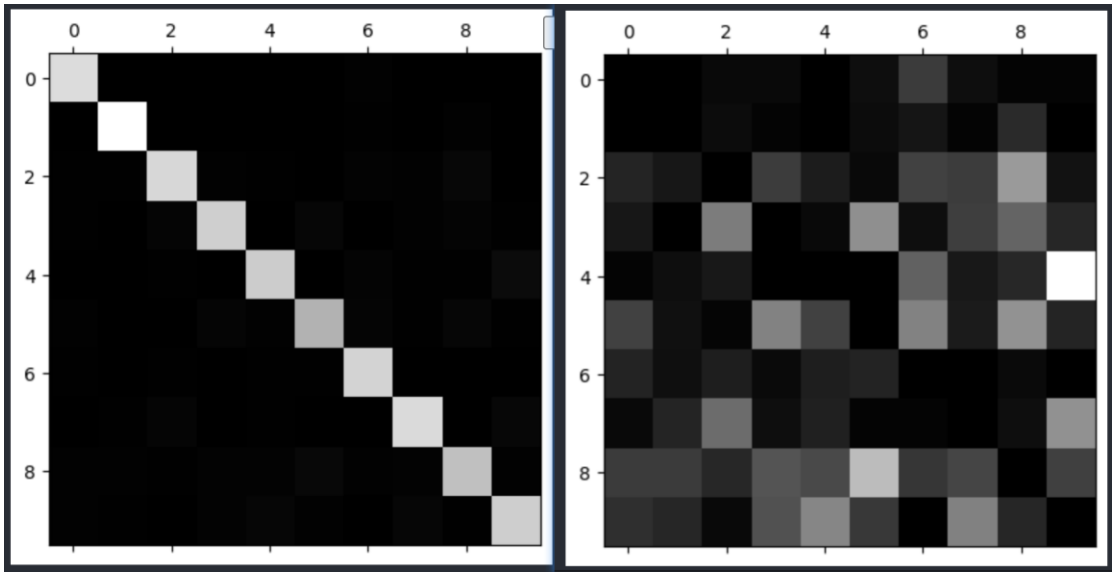
訓練集平均交叉驗證和測驗集準確率差不多，可能代表資料平均，訓練得很好，測試集準確率也有到 87，有機會可以更好，以下是 confusion matrix

```
Confusion Matrix:
[[ 902    0    8   11    1   13    2    4   39    0]
 [   0 1095    2    3    0    2    4    1   28    0]
 [   1   10  803   69    6    4    4   10  122    3]
 [   0    1    6  931    1   21    3    7   35    5]
 [   2    2    9   15  778    4    2    9   62   99]
 [   6    2    1   71    3  709   12   12   67    9]
 [   5    3   12   13    5   21  854    0   45    0]
 [   0    3   18   20    3    4    1  919   18   42]
 [   3    5    2   30    4   43    5    5  872    5]
 [   3    5    2   33    7    5    0   20   57  877]]
```

將訓練資料除以除以 255，進行 normalization，訓練模型的時間大大的縮短，然後準確率也明顯的提高到 0.91

```
Cross-Validation Scores: [0.90195 0.89985 0.90695]
Mean Accuracy: 0.9029166666666667
Test Set Accuracy: 0.9174
```

```
Confusion Matrix:
[[ 956    0    2    2    0    3   12    3    1    1]
 [   0 1112    3    1    0    3    5    1   10    0]
 [   8    5  934   13    6    2   14   13   33    4]
 [   5    0   26  902    2   30    3   13   21    8]
 [   1    3    5    0  888    0   20    5    8   52]
 [  12    3    1   24   12  777   24    5   27    7]
 [   7    3    6    2    6    7  925    0    2    0]
 [   2    8   23    3    7    1    1  949    3   31]
 [  12   12    8   17   15   38   11   14  834   13]
 [  10    8    2   17   28   12    0   27    8  897]]
```



由 confusion matrix 可以看出在 2 的表現比在其他數字上的表現更好

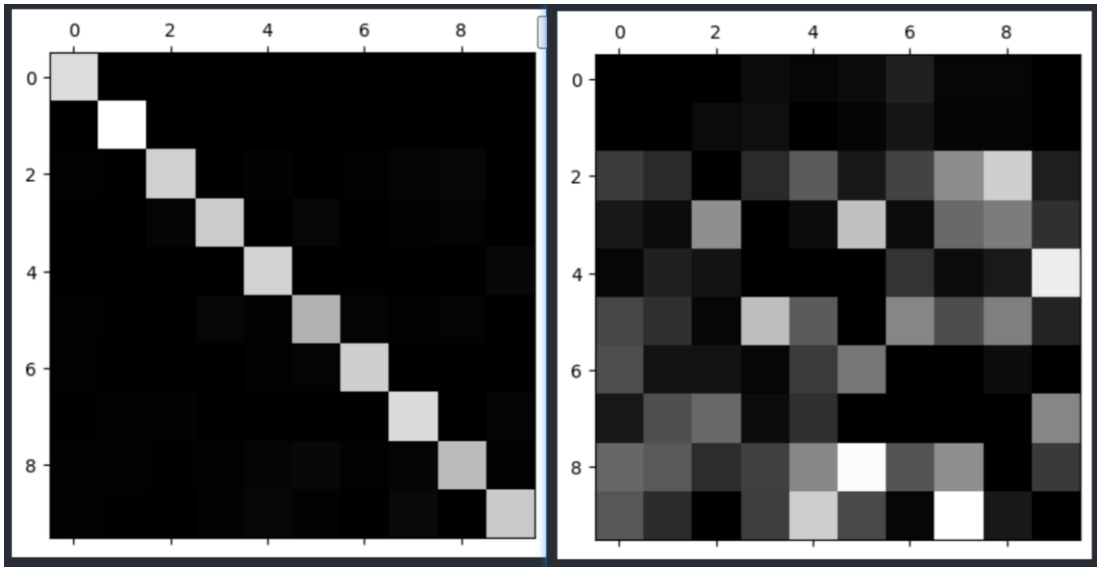
到了第二小題，我們使用人工成長訓練集來測試看看是否可以提高準確性，我們將 60000 筆 normalization 用來訓練的資料，利用系統的 scipy.ndimage 內的 shift 函數位移了上左右下各一份副本然後

合併到新的訓練集，標籤也是放入新的訓練集，依舊用 10000 個原始資料做測試，得到了以下結果

```
Cross-Validation Scores: [0.89107 0.85614 0.83964]
Mean Accuracy: 0.8622833333333334
Test Set Accuracy: 0.9193
```

可以看到準確率稍微的提升從 0.9174 上升到 0.9193 提升是微乎其微的

```
Confusion Matrix:
[[ 968    0    0    2    1    2    5    1    1    0]
 [   0 1123    2    3    0    1    4    1    1    0]
 [  10    7  916    7   15    4   11   23   34    5]
 [   4    2   23  901    2   31    2   17   20    8]
 [   1    5    3    0  922    0    8    2    4   37]
 [  10    7    1   27   13  781   19   11   18    5]
 [  12    3    3    1    9   18  910    0    2    0]
 [   4   13   17    2    8    0    0  962    0   22]
 [  16   14    7   10   21   39   13   22  823    9]
 [  14    7    0   10   33   12    1   41    4  887]]
```



SGDClassifier 的 normalization 或超參數調整 ，可以讓模型的準確率上升

使用 SGDClassifier 進行 Fashion MNIST 數據集的多類別分類，首將 70000 筆資料分成了 60000 和 10000，其中 60000 用來訓練模型，10000 用來測試，先用 SGDClassifier 訓練出模型後，使用交叉驗證評估模型性能（cv=3）來測量準確性，最後在 1000 筆資料的測試集上進行測試

```
Cross-Validation Scores: [0.78315 0.81355 0.82255]
Mean Accuracy: 0.8064166666666667
Test Set Accuracy: 0.8014
```

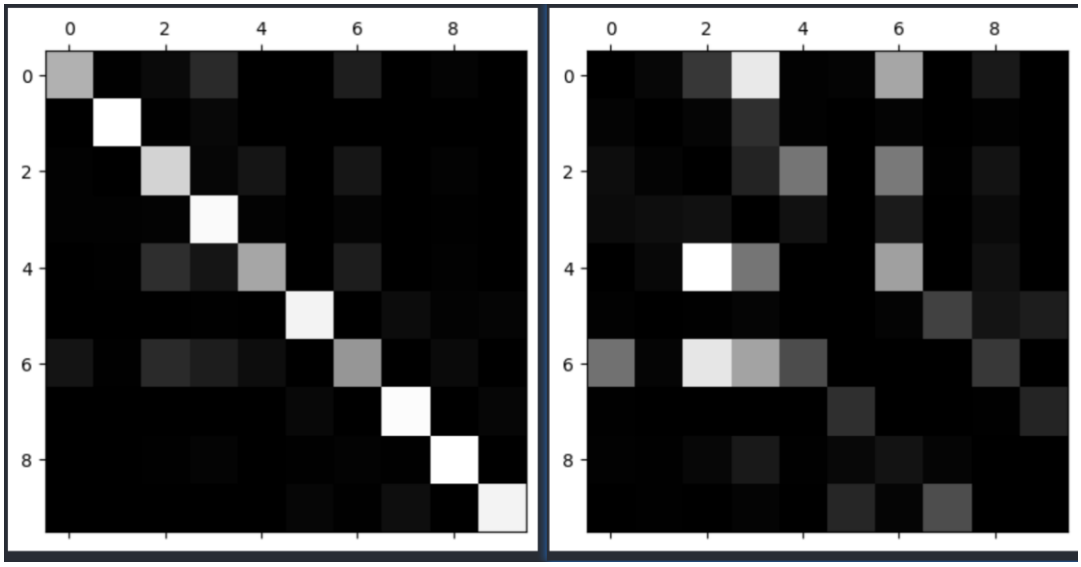
訓練集平均交叉驗證和測驗集準確率幾乎一致，相較於一開始的 MNIST 數據集，Fashion MNIST 的 Cross-Validation Scores 三項都比較低，感覺是圖案比較複雜，，以下是 confusion matrix

```
Confusion Matrix:
[[666  12  11  70  11   0 203   0  24   3]
 [  2 959   0  24   4   0   7   1   3   0]
 [  5   8 494  11 152   0 301   0  29   0]
 [  7  33  13 833  38   0  58   1  16   1]
 [  0   3  46  30 699   0 201   0  21   0]
 [  3   3   4   0   0 816   1  66  41  66]
 [ 92   5  49  45  77   0 689   0  42   1]
 [  0   0   0   0   0  12   0 932   4  52]
 [  1   1   0   6   1   4  19   5 963   0]
 [  1   0   0   0   0   3   1  30   2 963]]
```

將訓練資料除以除以 255，進行 normalization，準確率提升到了 0.82

```
Cross-Validation Scores: [0.8444 0.8359 0.8435]
Mean Accuracy: 0.8412666666666667
Test Set Accuracy: 0.8231
```

```
Confusion Matrix:
[[663   5  38 159   2   3 113   0  17   0]
 [  3 954   4  32   2   0   3   0   2   0]
 [  9   3 787  24  80   0  83   1  13   0]
 [  8  10  12 932  12   0  19   0   7   0]
 [  0   6 174  80 620   0 109   0  11   0]
 [  2   0   1   4   0 911   3  45  14  20]
 [ 78   4 157 111  52   0 559   0  39   0]
 [  1   0   0   0   0  32   0 941   1  25]
 [  2   1   5  17   1   6  13   4 951   0]
 [  0   1   0   3   0  26   4  53   0 913]]
```



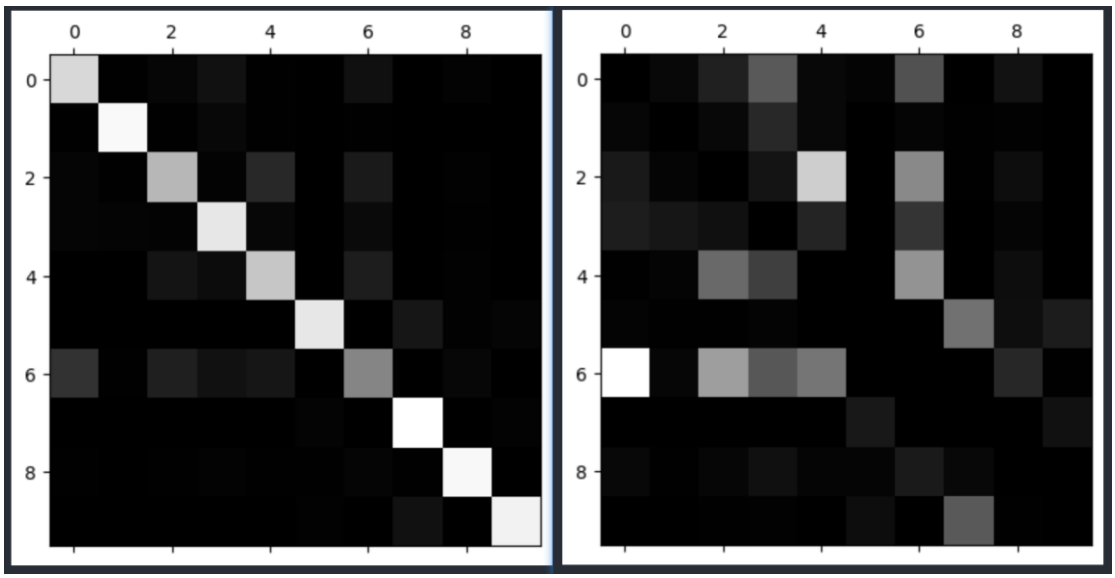
可以看到錯誤集中在 236 行

到了第二小題，我們使用人工成長訓練集來測試看看是否可以提高準確性，得到了以下結果

```
Cross-Validation Scores: [0.80176 0.75582 0.74984]
Mean Accuracy: 0.76914
Test Set Accuracy: 0.8298
```

跟一開始差不多，準確率來到 0.8298， confusion matrix

```
Confusion Matrix:
[[819   6  24  67   6   3  61   0  14   0]
 [  5 945   6  31   6   0   4   1   2   0]
 [ 19   4 696  15 153   0 102   1  10   0]
 [ 22  17  12 878  28   0  39   0   4   0]
 [  1   3  79  47 750   0 110   0  10   0]
 [  3   1   1   3   0 875   0  85  11  21]
 [190   5 118  65  87   0 505   0  29   1]
 [  0   0   0   0   0  18   0 969   0  13]
 [  6   1   5  12   4   4  20   6 942   0]
 [  0   0   1   2   0  10   0  67   1 919]]
```



可以看到第二行和第三行的錯誤減少了很多但是第一行增加了一點

以我的實驗去做結論的話，我認為 SGDClassifier 在第一個資料集上的效能比較好，在 Fashion MNIST 資料及上效果較差在， 2 個資料集在 normalization 之後性能就已經提升很多了的，而資料加強在兩個資料集上都沒有取得卓越的效果，我的結論是資料加強可能效果有限，僅限於如果我的實驗是正確的，以上是功課一的報告內容，感謝觀看