HW2 成果分析

本次作業使用 KNN 進行 MNIST 手寫數字分類。

資料集狀況 Dataset

MNIST dataset 的資料是數筆 28*28 pixels 的灰階影像

```
Shape of training data: (60000, 28, 28)
Shape of training labels: (60000,)
Shape of test data: (10000, 28, 28)
Shape of test labels: (10000,)
Number of classes: 10
```

KNN 實現 Practice

KNN 的作法就是找到k個最近的資料，以k個資料當中的**眾數**作為該點預測值。
距離計算的方法有兩種：l1和l2演算法。

```
# l1 distance
for i in range(num_test_data):
    for j in range(self.num_train_data):
      dists[i][j] = np.sum(np.abs(test_data_flatten[i] - self.train_data[j]))

# l2 distance
for i in range(num_test_data):
    for j in range(self.num_train_data):
      dists[i][j] = np.sqrt(np.sum(np.square(test_data_flatten[i] -
self.train_data[j])))
```

找到距離矩陣之後，就可以找最近的k個數值，並且取其眾數。

```
for i in range(num_test_data):
     k_closest_index = np.argsort(dists[i])[:k]
     k_closest_label = self.train_label[k_closest_index]
     preds[i] = np.argmax(np.bincount(k_closest_label))
```

執行結果與分析 Results

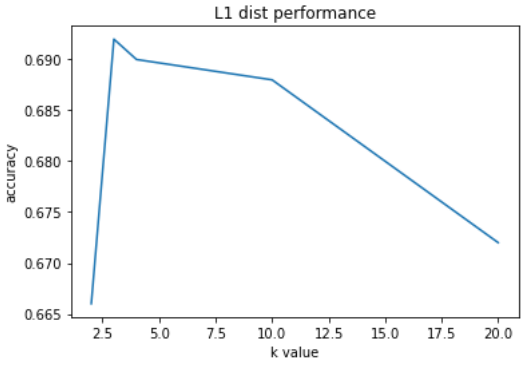
我們將資料經過的影像降至一維(28 ,28) => (784)，除此之外**不經過任何前處理**，直接丟入 knn 模型當中進行預測。

少樣本數預測

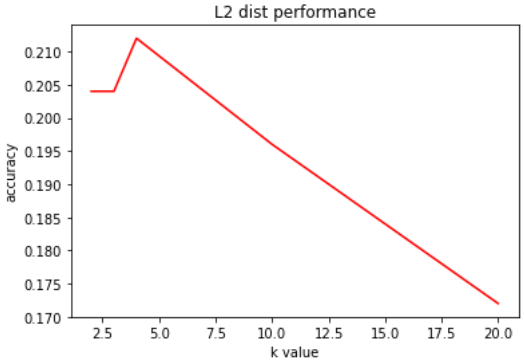
先取用少量樣本進行預測，找到最佳 k 值。

```
# Number of training data: 5000
# Number of test data: 500

Using L1 distance metric:
k = 2, accuracy = 0.666
k = 3, accuracy = 0.692
k = 4, accuracy = 0.69
k = 10, accuracy = 0.688
k = 20, accuracy = 0.672
```




```
Using L2 distance metric:
k = 2, accuracy = 0.204
k = 3, accuracy = 0.204
k = 4, accuracy = 0.212
k = 10, accuracy = 0.196
k = 20, accuracy = 0.172
```



程式執行約5分鐘。
在小樣本的情況，我們可以發現使用 l1、k=4的設定是最優秀的，準確率來到了0.69。

全部樣本預測

現在，使用這個設定來預測全部的資料。

```
Using L1 distance metric, k = 4
Accuracy = 0.7371
```

程式執行約11分鐘，這是因為 knn 進行一次預測時要運算很多次距離，當訓練資料有N個時，演算法效率為 $O(N)$。

改良 Improvisation

這次的改良重點決定要**提升準確率**。
我認為 knn 這類簡單的模型預測時，必須儘量減少雜訊的可能，因為簡單的模型容錯空間較少。像素太高的影像可能會反而增加雜訊，減少資料的穩定度。
我的作法是使用 `cv2.resize()` 對影像直接壓縮。

```
import cv2

def pre_processing(data, n=15):
  num_data = data.shape[0]
  out_data=np.zeros([num_data,n,n])
  for i in range(num_data):
    out_data[i] = cv2.resize(data[i], (n, n))

  return out_data
```

這次最佳結果(L2, k=4)的準確率來到了 0.93 ，已經接近 keras sequential model 的成績了。

```
Using L2 distance metric, k = 4
Accuracy = 0.9325
```

後續 Next Step

猜測可以使用卷積層的概念對影像進行前處理，減少資訊量或是強化特徵，進而增加準確度。
而速度的部份，可以使用課堂說的faiss演算法改進。