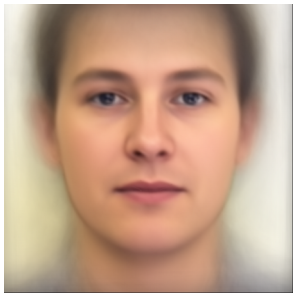
HW4

學號：b04501127 系級：土木三 姓名：凌于凱

A. PCA of colored faces(reference: hw4 手把手)

A.1. (.5%) 請畫出所有臉的平均。



A.2. (.5%) 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。

（註：因為我的第九個 eigenface 跟投影片的差一個負號，故下面 eigenface 都是乘過-1 的，左到右＝1-4）

A.3. (.5%) 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

下面圖片分別為 10.jpg, 100.jpg, 200.jpg, 300.jpg

A.4. (.5%) 請寫出前四大 Eigenfaces 各自所佔的比重，請用百分比表示並四捨五入到小數點後一位。

	pc1	pc2	pc3	pc4
si/Σsj(%)	4.15	2.95	2.39	2.21

B. Image clustering(reference: hw4 手把手)

B.1. (.5%) 請比較至少兩種不同的 feature extraction 及其結果。(不同的降維方法或不同的 cluster 方法都可以算是不同的方法)

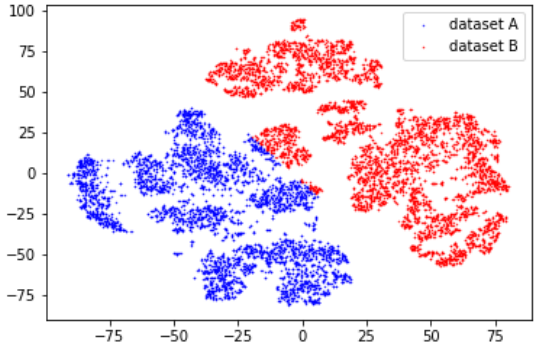
cluster: kmean(n_clusters=2, random_state=0)
dimension reduction:
1. deep autoencoded: 5 層 encoder, 5 層 decoder
詳細模型如下

```
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 784)               0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               401920
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328
_________________________________________________________________
dense_3 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_4 (Dense)              (None, 64)                8256
_________________________________________________________________
dense_5 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_6 (Dense)              (None, 64)                2112
_________________________________________________________________
dense_7 (Dense)              (None, 128)               8320
_________________________________________________________________
dense_8 (Dense)              (None, 256)               33024
_________________________________________________________________
dense_9 (Dense)              (None, 512)               131584
_________________________________________________________________
dense_10 (Dense)             (None, 784)               402192
=================================================================
Total params: 1,153,712
Trainable params: 1,153,712
Non-trainable params: 0
```

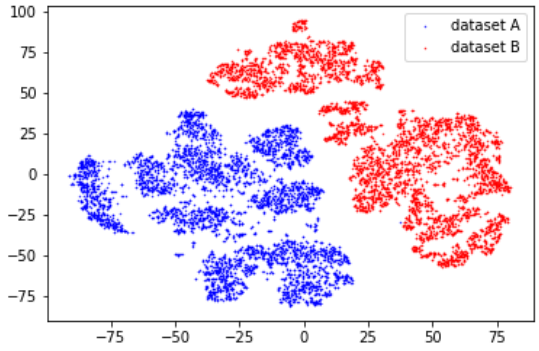
public score: 0.99404, private score: 0.99422
2. pca:利用 sklearn 的 pca(n_components=32)
public score: 0.51708, private score: 0.51783

兩個都是降維到 32 維，但成績差了有點多，可能是因為 pca 無法觀察到多維點和點之間有沒有相連的關係，使用後可能會造成資料上的損失，所以成果不佳。

B.2. (.5%) 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈。



B.3. (.5%) visualization.npy 中前 5000 個 images 跟後 5000 個 images 來自不同 dataset。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。



可以發現在正中間預設的結果和正確的不一樣，覺得比較像是 kmean 初始值的問題，用 tsne 時感覺已經把不能 label 的分很開了，所以我認為是 kmean 初始值可能是設在藍色最左下的地方，以及紅色中間，才造成中間有一塊被分類成紅色。

C. Ensemble learning(reference: hw2 手把手)

C.1. (1.5%) 請在 hw1/hw2/hw3 的 task 上擇一實作 ensemble learning，請比較其與未使用 ensemble method 的模型在 public/private score 的表現並詳細說明你實作的方法。（所有跟 ensemble learning 有關的方法都可以，不需要像 hw3 的要求硬塞到同一個 model 中）

實作在 hw2，利用 np.random.choice 隨機抽樣出 10 個 n=n’的

train set（with replacement），在個別利用 logistic regression 去做 classify，訓練出來的 10 個 model 再去 predict test set，在將結果透過 voting 的方式決定最後的成績。

model	0	1	2	3	4
public	0.85721	0.85945	0.85796	0.86068	0.85945
private	0.85603	0.85480	0.85480	0.85702	0.85542

model	5	6	7	8	9
public	0.86044	0.86130	0.85909	0.85749	0.85798
private	0.85787	0.85714	0.85824	0.85603	0.85566

將十個 model 預測的成果拿去 voting:
public: 0.86019, private: 0.85566
成績並沒有進步太多，還比 model 6 還低，推測是因為一些成績較低的在拉低分數，所以在重新用 private 成績前四名(3, 5, 6, 7)的來做 ensemble-→public: 0.86167, private: 0.85763
確實有了一些進步，但有點少，可能是因為 model 還不夠複雜的關係，造成使用 bagging 的方式成果沒有很顯著。