HW4

學號：r06921002 系級：電機所碩一 姓名：張哲誠

A. PCA of colored faces

A.1. (.5%) 請畫出所有臉的平均。

Ans:

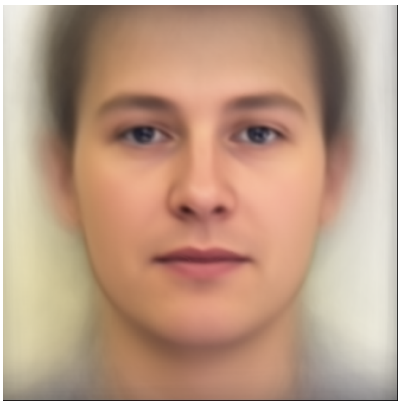

fig. 1 所有臉的平均

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

Ans: Eigenfaces_0~ Eigenfaces_3 由大到小排序

fig. 2 Eigenfaces_0

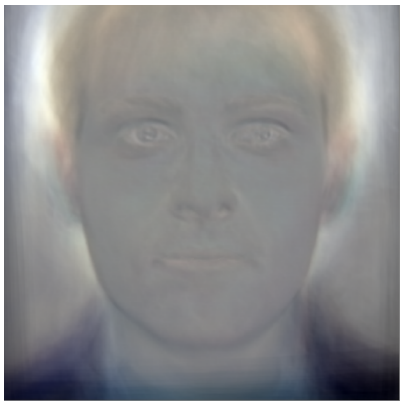

fig. 3 Eigenfaces_1

fig. 4 Eigenfaces_2

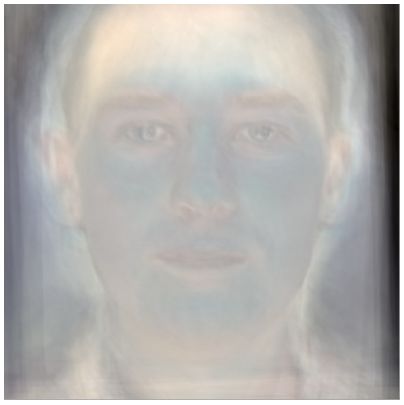

fig. 5 Eigenfaces_3

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。

Ans: src_, y_後的編號代表從 Aberdeen 資料庫中讀取的代號。

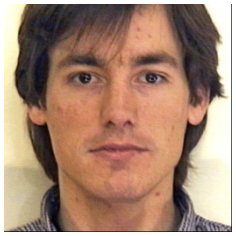

fig. 6 src_0

fig. 7 y_0

fig. 8 src_3

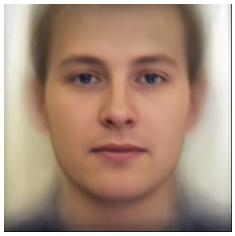

fig. 9 y_3

fig. 10 src_4

fig. 11 y_4

fig. 12 src_8

fig. 13 y_8

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。

Ans: 此題將 A.3 中的四張圖之前四大 eigenfaces 的比重列舉

前四大 Eigenfaces	s
Eigenfaces_0	4.1
Eigenfaces_1	3.0
Eigenfaces_2	2.4
Eigenfaces_3	2.2

B. Image clustering

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)

Ans: 此題我分別使用 **PCA 降維到 401 維度加上 k-means 分群**與 **autoencoder 加上 k-means 分群**(參考 ML2018 Spring HW4 TA Hours.pdf) 兩種 feature extraction 做比較。下表為兩者在 kaggle 上的成績比較。

Feature extraction method	Kaggle score
PCA + k-means	0.99994
Autoencoder + k-means	0.98357

兩者的差別在於，PCA 僅限於**線性運算**的分類，而 autoencoder 具有**非線性編碼器/解碼器**，而具有線性傳遞函數的單層 autoencoder 幾乎等同 PCA。

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

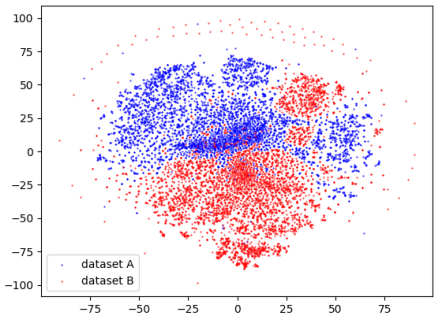

fig. 14 visualization.npy ground truth

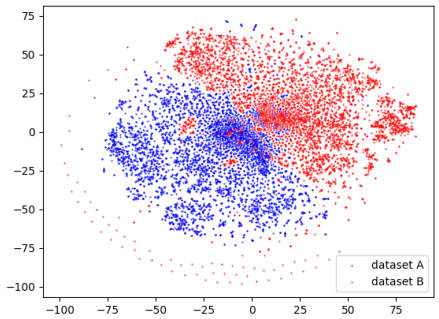

fig. 15 visualization.npy 預測結果圖

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。

Ans: fig. 14 中，嘗試使用助教所提中的 sample code，畫出 visualization.npy 使用 PCA 降維再使用 t-sne 作視覺化的結果，把前 5000 筆 label 成藍色並訂為 dataset A、後 5000 筆 label 成紅色並 label 成 dataset B；fig. 15 使用我在本題所建構出來的 PCA model，先行把 visualization.npy 降至 401 個維度，再使用 k-means 分成兩群，並 label 上 0 與 1，最後根據每本資料畫出 dataset A 與 dataset B 的 scatter 圖。兩種的結果都可以順利的將兩個資料群分成兩個群，如 fig. 14、fig. 15 所示。我原先認為，兩張圖的分佈情形應該要極為類似，但左圖較分成左上與右下，而右圖分成左下與右上兩群。

C. Ensemble learning

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）

Ans: 本題我實作 ensemble learning 中的 **adaboost classifier** 應用於 hw2 的 income prediction 上。Training data 與當初我所繳交作業二的

hw2_best_train.py 一樣，除了先做標準化以外，還多加了 age, sex_Female, sex_Male, capital_gain, capital_loss 與 hours_per_week 的平方項當作新的 training data 餵進 adaboost classifier。而我 adaboost classifier 的寫法如下：

```
# Create and fit an AdaBoosted decision tree
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=1),
                         algorithm="SAMME",
                         n_estimators=200)
bdt.fit(trainX, trainY)
sig = bdt.predict(testX)
```

下圖表比較我使用 hw2_best_test.py 與本次 ensemble learning 所使用的 adaboost classifier 在 kaggle 上的分數表現：

	Public Score	Private Score
Hw2_best (logistic)	0.85749	0.85616
Hw4_adaboost	0.85958	0.85738

從這個結果來看，adaboost 在這個情境下的分類情形較 logistic regresion 好。