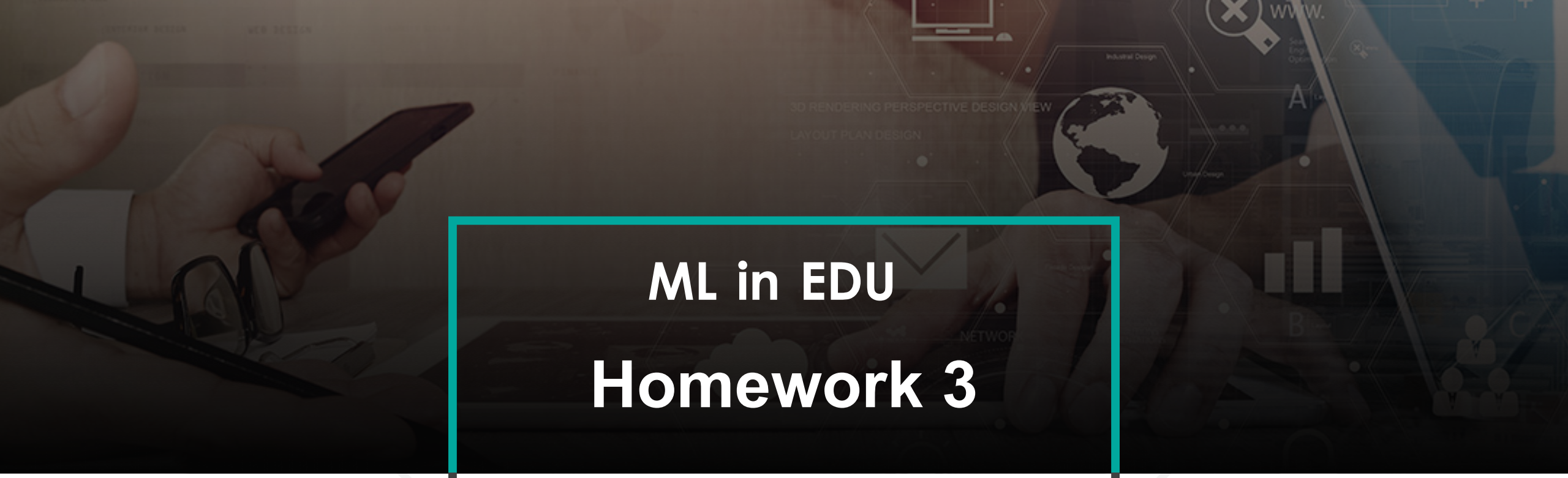


ML in EDU

Homework 3

Clustering & Dimensionality Reduction

Ref: https://towardsdatascience.com/dimensionality-reduction-toolbox-in-python-9a18995927cd

CONTENTS

Fashion MNIST

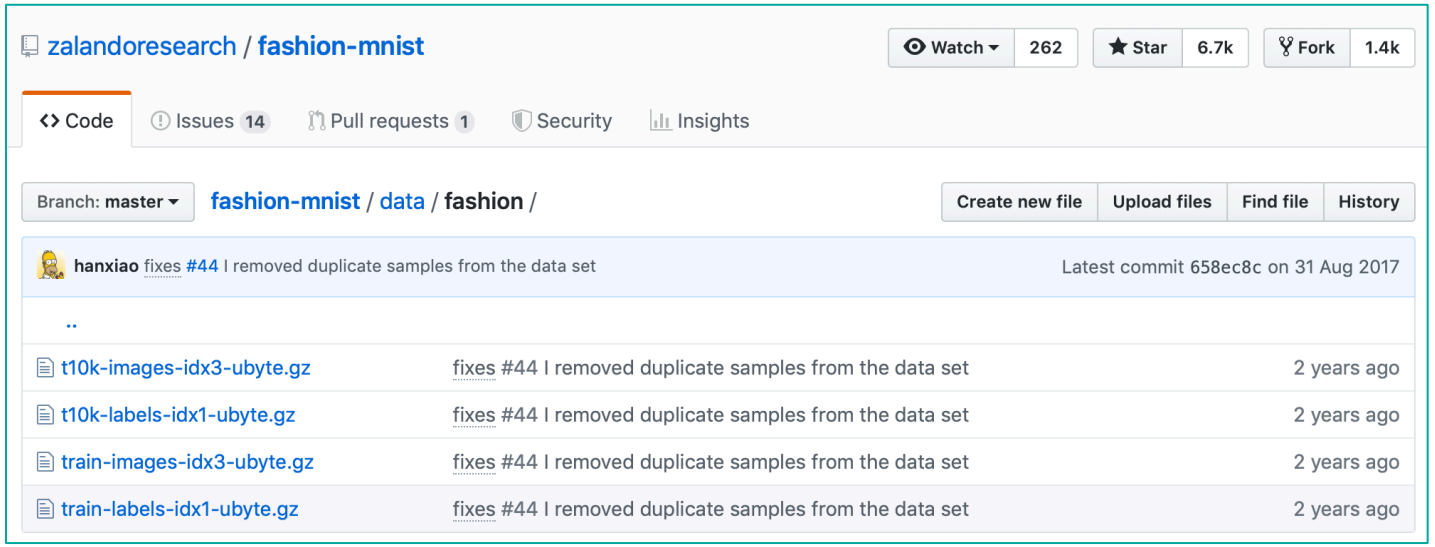
- 1-1 Load Data (10%)
- 1-2 Dimensionality Reduction (40%)

Face

- 2-1 Load Data
- 2-2 Clustering (50%)

1-1 Load data (10%)

1. 載入 fashion-mnist 資料，可使用下面任意方式
   1) 至官方 GitHub Repo 下載：https://github.com/zalandoresearch/fashion-mnist/tree/master/data/fashion



   2) 至 newE3 下載壓縮檔：https://e3new.nctu.edu.tw/pluginfile.php/534208/mod_assign/introattachment/0/fashion-mnist.zip?forcedownload=1
   3) 使用套件內的 function load data：

1-1 Load data (10%)

1. 載入 fashion-mnist 資料，可使用下面任意方式
   1) 至官方 GitHub Repo 下載：https://github.com/zalandoresearch/fashion-mnist/tree/master/data/fashion
   2) 至 newE3 下載壓縮檔：https://e3new.nctu.edu.tw/pluginfile.php/534208/mod_assign/introattachment/0/fashion-mnist.zip?forcedownload=1

   3) 使用套件內的 function load data：

1-1 Load data (10%)

1. 載入 fashion-mnist 資料，可使用下面任意方式
   1) 至官方 GitHub Repo 下載：https://github.com/zalandoresearch/fashion-mnist/tree/master/data/fashion
   2) 至 newE3 下載壓縮檔：https://e3new.nctu.edu.tw/pluginfile.php/534208/mod_assign/introattachment/0/fashion-mnist.zip?forcedownload=1
   3) 使用套件內的 function load data：
      e.g. `keras.datasets.fashion_mnist.load_data()`
      e.g. other packages: https://github.com/zalandoresearch/fashion-mnist#loading-data-with-other-machine-learning-libraries

```
from keras.datasets import fashion_mnist
((trainX, trainY), (testX, testY)) = fashion_mnist.load_data()
```

```
Using TensorFlow backend.

Downloading data from http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/train-labels-idx1-ubyte.gz
32768/29515 [==============================] - 0s 9us/step
Downloading data from http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/train-images-idx3-ubyte.gz
26427392/26421880 [==============================] - 4s 0us/step
Downloading data from http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/t10k-labels-idx1-ubyte.gz
8192/5148 [===============================================] - 0s 0us/step
Downloading data from http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/t10k-images-idx3-ubyte.gz
4423680/4422102 [==============================] - 6s 1us/step
```

1-1 Load data (10%)

1. 載入 fashion-mnist 資料
2. 印出 trainX, trainY, testX, testY 的 shape (10%)

```
shape = [[trainX.shape, trainY.shape], [testX.shape, testY.shape]]
col_name = ['image(X)', 'label(Y)']
row_name = ['train', 'test']
pd.DataFrame(shape, columns=col_name, index=row_name)
```

	image(X)	label(Y)
train	(60000, 28, 28)	(60000,)
test	(10000, 28, 28)	(10000,)

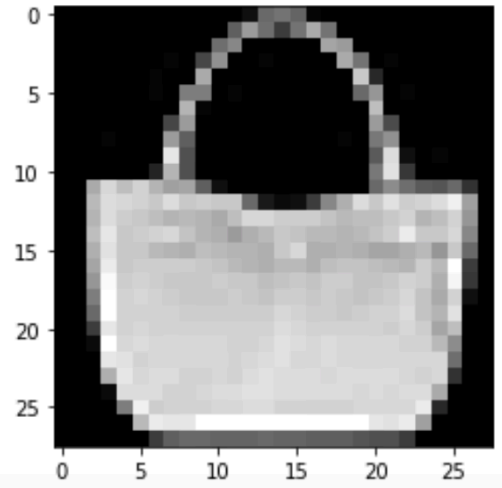
Try to get familiar with data

你可以試著用 matplotlib.pyplot.imshow() 把資料畫出來看看

- 詳細可以參考官方 repo: https://github.com/zalandoresearch/fashion-mnist
  - 也有人寫了中文版的 README: https://github.com/zalandoresearch/fashion-mnist/blob/master/README.zh-CN.md

```
import matplotlib.pyplot as plt
plt.imshow(trainX[100], cmap='gray')
```

```
<matplotlib.image.AxesImage at 0x7f38fb77b240>
```



Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

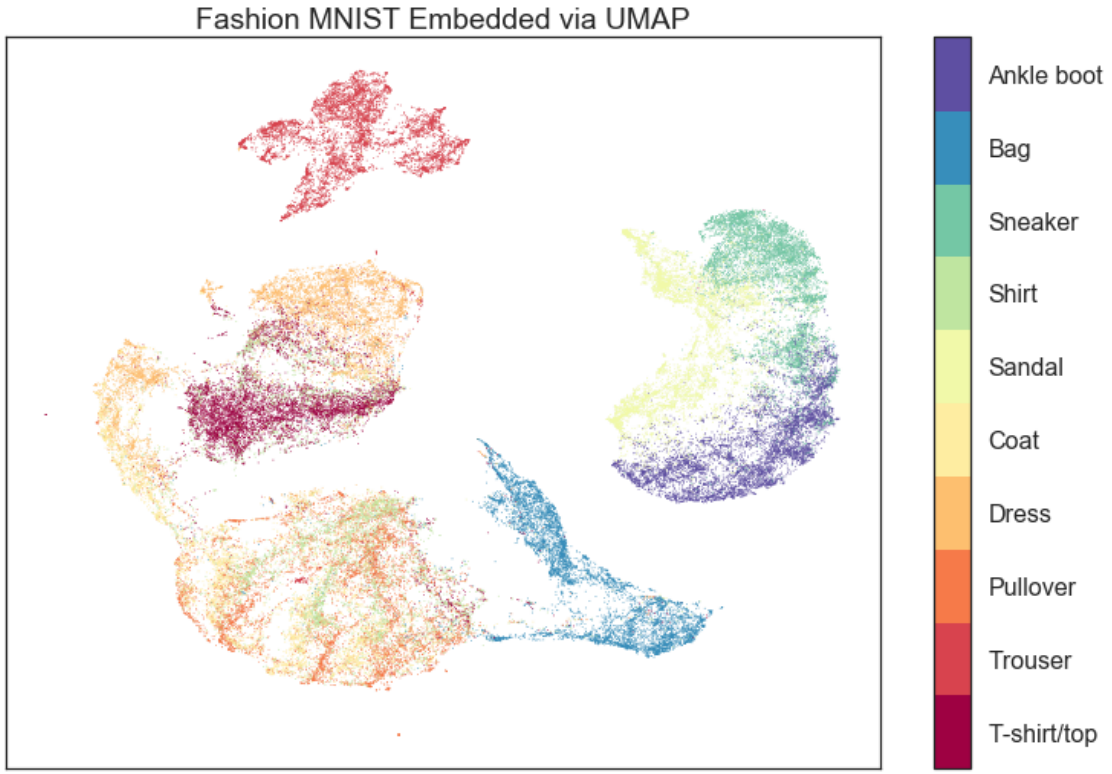
1-2 Dimensionality Reduction (40%)

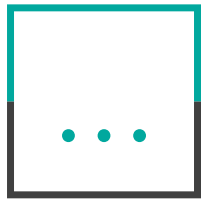
因原本資料維度過高 (28*28)，可以試著降維再處理。

1. 請使用至少兩種方式降維，並比較其差異。
   - 不限制降到幾維
   - e.g. LDA, PCA, t-SNE, UMAP, AutoEncoder…

2. 請將降維的結果作圖（未必要將每一筆資料都畫出來）

參考結果：https://github.com/zalandoresearch/fashion-mnist#visualization





Future work…

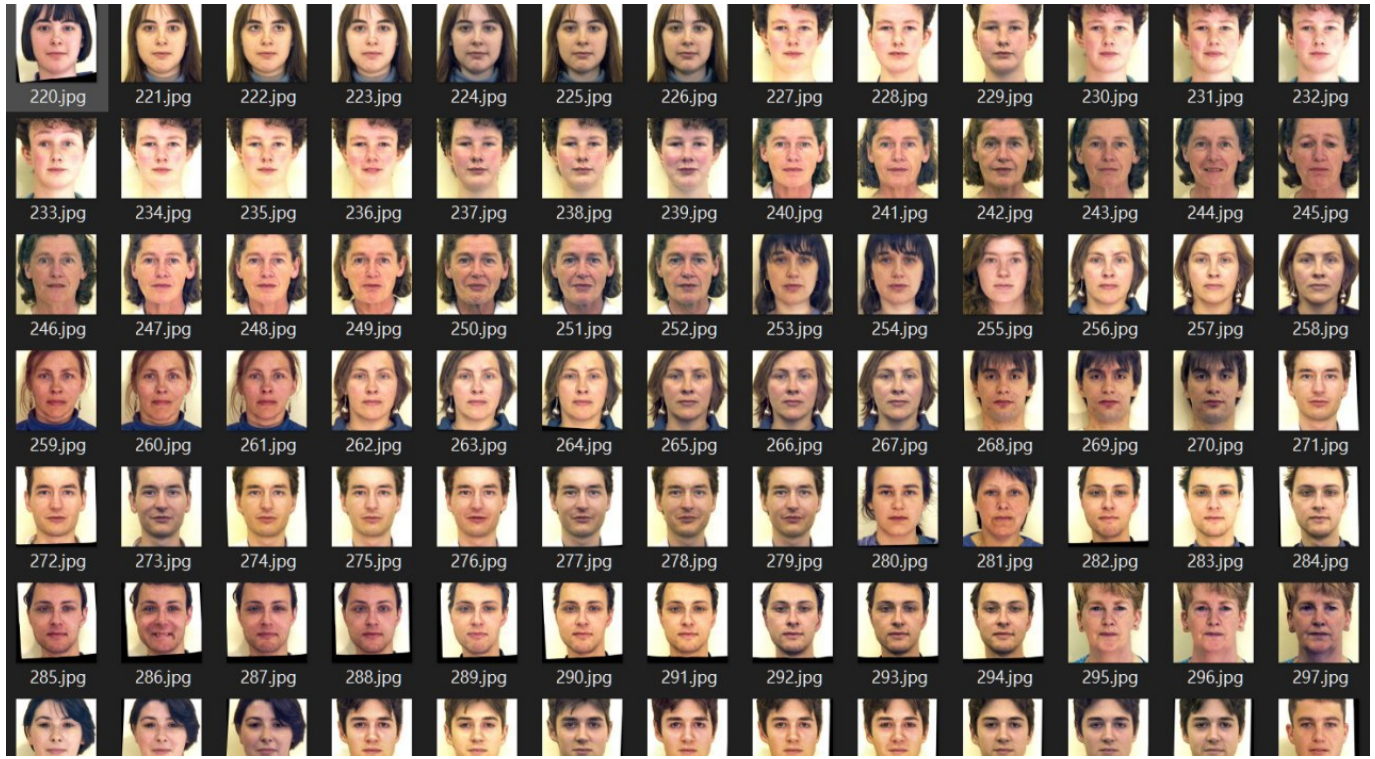
降維完可以做什麼？

1. 若 label 未知，可試圖 clustering
2. 若 label 已知，可做 classification
3. 可試圖 reconstruct 降維後資料（即圖片壓縮）

…

2-1 Load data

1. 請至 newE3 下載指定的資料 (Aberdeen_HungYiLee.zip)
   - Aberdeen University 的 Prof. Ian Craw 所收集，由台大李宏毅教授及其研究生所整理。
   - 415 張 600 x 600 x 3 的臉部彩圖



2-2 Clustering (50%)

1. 請至 newE3 下載指定的資料 (Aberdeen_HungYiLee.zip)
2. (50%) 請用至少兩種分群法對此資料做分群，群數可自訂，並解釋分群結果。
   - e.g. k-means, hierarchical clustering…
     其他分群法可參考：https://scikit-learn.org/dev/auto_examples/cluster/plot_cluster_comparison.html?fbclid=IwAR3ufFA5w5NHNff8F-aQOtmNVThoULsoCYBKy3qxN9OirBs-Lpngp2UWwak#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py

   - 決定分兩群
     - 長髮/短髮、喜悅/悲傷
   - 決定分多群
     - 喜怒哀樂…

2-3 Recommended (0%)

1. 請至 newE3 下載指定的資料 (Aberdeen_HungYiLee.zip)
2. (50%) 請用至少兩種分群法對此資料做分群，群數可自訂，並解釋分群結果。

3. 你可以自行 label（可以用上面分群的結果作為 label 參考、或是手動標記）
   並再加入監督式學習進行分析
   （此題不強制要做）

Hand in your homework to e3

Hand in your report & code to e3(https://e3new.nctu.edu.tw)
Briefly describe how your code works and show results. Make sure TA could run your code.

You should only hand in 2 files:

`hw3_<student_id>.pdf`
`hw3_<student_id>.zip`

(e.g. hw3_0123456.pdf, hw3_0123456.zip)

Due: 11/12 (Tue.) 11:00 a.m.

TA:
蔡旻均 dollars9256741@gmail.com
劉昱劭 ysl@cs.nctu.edu.tw