Machine Learning 2017 HW4 Report

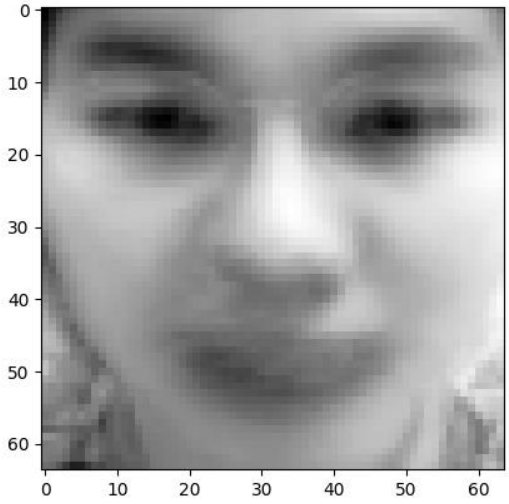
Predict PM 2.5

B03901156 Yu Xuan Huang

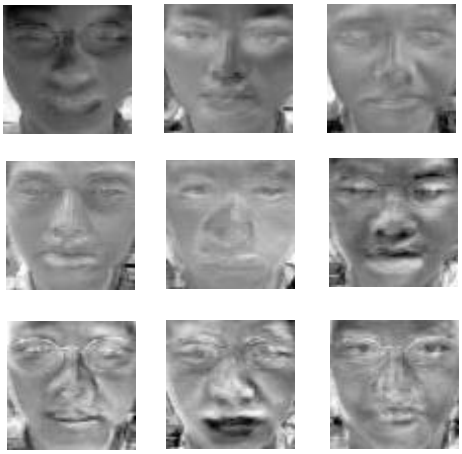
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

ANS.

Average Face



Eigenfaces



1.2. Dataset 中前 10 個人的前 10 張照片的原始圖片和 reconstruct 圖 (用前 5 個 eigenfaces):

1.3. Dataset 中前 10 個人的前 10 張照片投影到 top k eigenfaces 時就可以達到 < 1% 的 reconstruction error.

Ans. From my experiment, even when k = 100, the reconstruction error is still >1%; when k =100, the smallest reconstruction error = 0.017834625142 can be obtained.

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

ANS. train : the file path of the training data

output: the file path of the output saved model

size: 向量的維數，此為 100

window: 上下文窗口大小，此為 default 值=5

negative: 負例的數目 (用於 negative sampling), 此設為 0

min_count: 被捨棄/截斷的低頻詞閥值, 此設為 5

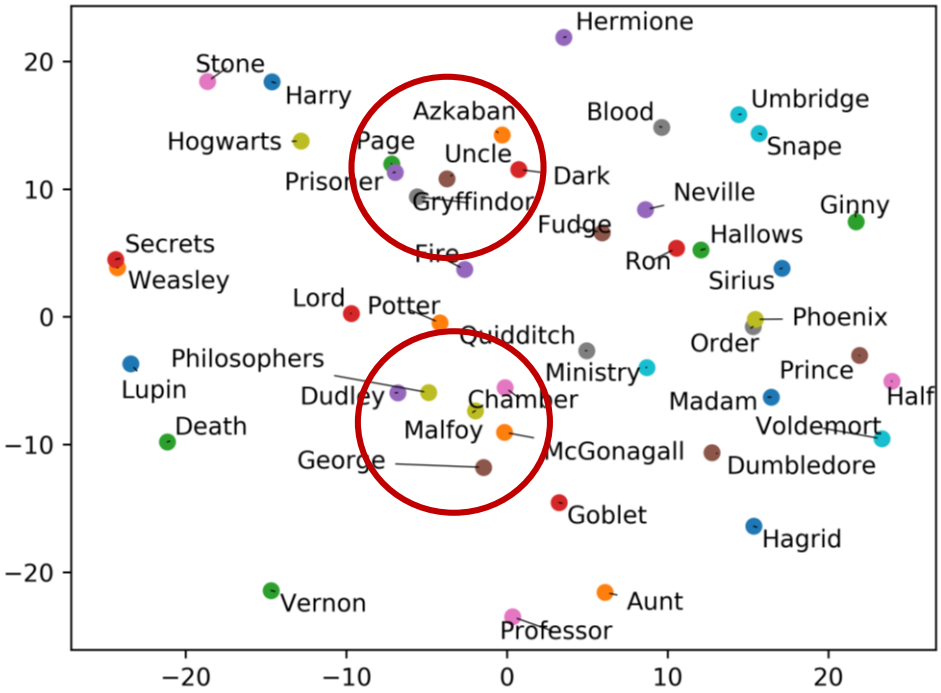
alpha: 起始的 learning rate, 此設為 0.025

iter: number of training iteration, 此設為 10000

cbow: 是否使用 CBOW 算法(0 為不使用), 此設為 1

2.2. 將 word2vec 的結果投影到 2 維的圖:

ANS.



2.3. 從上題視覺化的圖中觀察到了什麼？

ANS. 詞類較為相近的字詞理應較為接近(如上圖中紅圈處，Page 和 Prisoner, 或 Phoenix 和 Order)，但從此圖觀察可知，其實驗結果的分布點較為分散，較難看出相似詞類的集中程度

3.1. 請詳加解釋你估計原始維度的原理、合理性，這方法的通用性如何？
ANS.
1. 根據題目所給的 dataset 資訊，如維度分布、oracle network 結構產生一些 training data

2. 產生不同維度的 training data

```
# generate some data for training
X = []
y = []
for i in range(60):
    dim = i + 1
    print(i)
    for N in [10000, 12000, 14000, 16000, 18000, 20000, 22000, 24000, 26000, 28000, 30000, 32000, 34000, 36000,
    38000, 40000, 42000, 44000, 46000, 48000, 50000, 52000, 54000, 56000, 58000, 60000, 62000, 64000, 66000, 68000,
    70000, 72000, 74000, 76000, 78000, 80000, 82000, 84000, 86000, 88000, 90000, 92000, 94000, 96000, 98000, 100000]:
        #print(N)
        layer_dims = [np.random.randint(60, 80), 100]
        data = gen_data(dim, layer_dims, N).astype('float32')
        eigenvalues = get_eigenvalues(data)
        X.append(eigenvalues)
        #print(X)
        y.append(dim)
        #print(y)

X = np.array(X)
y = np.array(y)

np.savez('train4_data.npz', X=X, y=y)
```

3. 從所產生的每個 dataset 中隨機選取 20 個 sample points，對於每個 point 找出 200 個 nearest neighbors

4. 從 data subset 中計算平均的 eigenvalues

5.利用線性的 SVR 去預測 testing dataset 的維度

合理性: 藉由產生類似原始資料性質的 data 作為 training data，模擬其資料特性，再利用 SVR 預測未知的資料維度，相信這樣的 model 是能有效反映資料特性

通用性: 由於此 model 是根據此 dataset 的原始資料性質所設計的，若用於估計具相似特性的 dataset 效果將會不錯，但若性質差異較大，則預計其估計結果較差

3.2. 將你的方法做在 hand rotation sequence datatset 上得到什麼結果？合理嗎？請討論之。
答：
我認為其結果較為不合理，可能是因為此 model 是根據此 dataset 的原始資料性質所設計的，而 hand rotation data 的資料性質與之差異較大，其估計結果較差