Machine Learning Project 4

• b03902089 資工三 林良翰

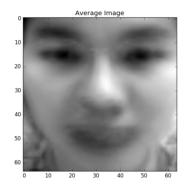
Eigenfaces with PCA

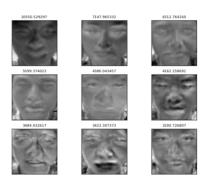
Data Sets

- Carnegie Mellon University Electrical & Computer Engineering
- Project: Eigenflow Based Face Authentication
 (http://chenlab.ece.cornell.edu/projects/FaceAuthentication/Default.html)
- Download: link (http://chenlab.ece.cornell.edu/projects/FaceAuthentication/download.html)

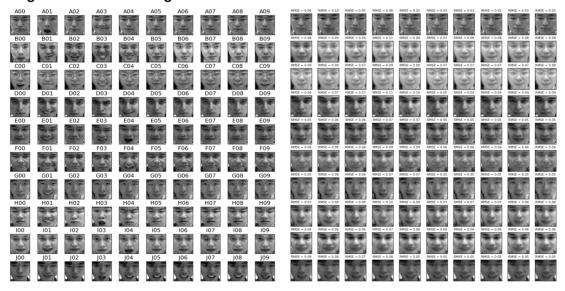
Questions

1. Perform PCA using the first 10 faces of the first 10 subjects to obtain the eigenfaces. Plot the average face. Also plot the top 9 eigenfaces in a figure.





2. Project the 100 faces onto the top 5 eigenfaces, and then reconstruct the original images. Plot the 100 original faces and the recovered faces.



- 3. In 2., we can choose top k eigenfaces and check the reconstruction error (RMSE). Find the smallest k such that the error is less than 1%.
 - \circ Smallest k = 56, RMSE = 0.98%

Visualization of Word Vectors

Data Sets

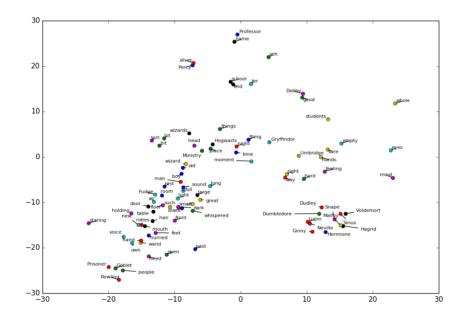
· Corpus: Harry Potter Series

Download: link

(https://archive.org/compress/Book5TheOrderOfThePhoenix/formats=DJVUTXT&file=/Book5TheOrderOfThePhoenix.zip)

Questions

- 1. Train word vectors with the toolkit. Report the parameters you used and explain what the parameters mean.
 - word2vec(): size=50, convert words into vectors of 50 dimensions.
 - TSNE(): n_component=2, reduce the dimension of vectors to 2.
- 2. Plot the visualization of word vectors on 2D space. Show the figure in your report.



- 3. Discuss your observations from the visualization.
 - Names(labeled None) and other nouns(NN, NNP, NNS) are seperated clearly.
 - There are some words not in the two main clusters, but the meaning of words are highly related with the clusters.

Estimation of Intrinsic Dimension

Data Sets

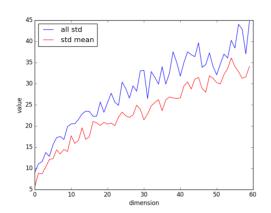
- There are 200 sets $[S_1 \dots S_{200}]$ of data. Each set contains 10k-100k datapoints in \mathbb{R}^{100} .
- Each Set of data are generated from oracle network: $i \in [1, 200]$

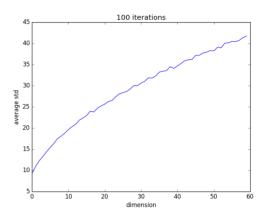
$$\mathbb{R}^{d_i} \xrightarrow{ELU} \mathbb{R}^{h_i} \xrightarrow{ELU} \mathbb{R}^{100} \xrightarrow{Linear} \mathbb{R}^{100}$$

where $h_i \in [60, 79]$ uniformly, and each layer performs a transformation f(Wx + b), both

Questions

- 1. Please elaborate your method and why you used that method. Discuss the results in detail.
 - o I modified gen.py to run on [1, 60] dimensions, with random sample size $N \in [10^4, 10^5]$, $h_i \in [60, 79]$, and found that the dimension of input d_i and the standard deviation of output σ_i are highly positive related.





- o Method 1: K-Means Clustering
 - According to the positive correlation between input d_i and output σ_i , use K-means clustering to find 60 clusters $[k_1, \ldots, k_{60}]$ and label the dimension of each cluster with respect to the mean of standard deviation $[\sigma_{k_1}, \ldots, \sigma_{k_{60}}]$.
 - Error on Kaggle public test: 0.15632
- o Method 2: K-Means Clustering with Initial Centers
 - Besides K-means clustering, I generate 60 averaged centers of output σ_i from input $d_i \in [1, 60]$ for 100 iterations, and let them be the initial centers of k-means clustering.
 - Error on Kaggle public test: 0.13157
- Method 3: Initial Centers ONLY!
 - Simply trust the centers generated by myself, and find the closest center for each data set for labeling dimensions.
 - Error on Kaggle public test: 0.11435
- 2. Download the hand rotation sequence dataset, try to estimate the intrinsic dimension of this dataset and discuss your result.
 - Download: link (http://vasc.ri.cmu.edu//idb/html/motion/hand/index.html)