Machine Learning Homework 7

Due Date 23:55 13th June 2022

> 📢 Packages allowed in this assignment:
>
> > You are only allowed to use `numpy`, `scipy.spatial.distance`, and I/O related functions (like `cv2.imread()`, `csv`, `matplotlib` etc.)

> 📢 Important: `scikit-learn` and `SciPy` are not allowed.

---

1. Kernel Eigenfaces

> 😊 In this section, you are going to do face recognition using **eigenface and fisherface**.
>
> > Reference: https://www.csie.ntu.edu.tw/~mhyang/papers/fg02.pdf

- ### Data
  - The `Yale_Face_Database.zip` on E3 contains 165 images of 15 subjects (subject01, subject02, etc.). There are 11 images per subject, one for each of the following facial expressions or configurations: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.
  - These data are separated into training dataset(135 images) and testing dataset(30 images). You can resize the images for easier implementation.
- ### Tasks for You
  - **Part1:**
    - Use PCA and LDA to show the first 25 eigenfaces and fisherfaces, and randomly pick 10 images to show their reconstruction. (please refer to the lecture slides).
  - **Part2:**
    - Use PCA and LDA to do face recognition, and compute the performance. You should use k nearest neighbor to classify which subject the testing image belongs to.
  - **Part3:**
    - Use kernel PCA and kernel LDA to do face recognition, and compute the performance. (You can choose whatever kernel you want, but you should try different kernels in your implementation.) Then compare the difference between simple LDA/PCA and kernel LDA/PCA, and the difference between different kernels.

---

2. t-SNE

> 😎 Here are nice implementations of t-SNE in different programming languages:
>
> > https://lvdmaaten.github.io/tsne/

- ### Data & reference code
  - Download `tsne_python.zip` from E3
    - You can also download it from https://lvdmaaten.github.io/tsne/code/tsne_python.zip
    - `mnist2500_X.txt`: contains 2500 feature vectors with length 784, for describing 2500 mnist images.
    - `mnist2500_labels.txt`: provides corresponding labels
    - `tsne.py`: reference code
- ### Tasks for You
  - **Part1:**
    - Try to modify the code a little bit and make it back to symmetric SNE. You need to first understand how to implement t-SNE and find out the specific code piece to modify. You have to explain the difference between symmetric SNE and t-SNE in the report (e.g. point out the crowded problem of symmetric SNE).
  - **Part2:**
    - Visualize the embedding of both t-SNE and symmetric SNE. Details of the visualization:
      - Project all your data onto 2D space and mark the data points into different colors respectively. The color of the data points depends on the label.
      - Use videos or GIF images to show the optimize procedure.
  - **Part3:**
    - Visualize the distribution of pairwise similarities in both high-dimensional space and low-dimensional space, based on both t-SNE and symmetric SNE.
  - **Part4:**
    - Try to play with different perplexity values. Observe the change in visualization and explain it in the report.

---

3. Report

> 💀 Submit a report in **pdf format**. The report should be written in **English**.
> 💀 Please follow the **report format**. If you skip some sections in the report format, you score will be affected. Additional content outside the format is welcome (but we may not be able to give you extra points).
> 💀 Please don't explain the code line by line. You need to explain it clearly and well structured. For example, explain which part you done in the function.
> 💀 Since this homework is mainly graded by report, please spend more time on it. (e.g. well organized) We won't give you any points if you just finish the code.

- **Report format:**
  - **A. Code (with detailed explanations) (40%)**
    - Paste the screenshot of your functions with comments and explain your code. Explain the process to clustering and show different kernels.
    - If you don't explain your code, you cannot get any point in section B and C either.
    - **Kernel Eigenfaces**
      - Part1 (10%) Also, simply explain how you do PCA & LDA (what are the steps?)
      - Part2 (5%)
      - Part3 (10%) Also, simply explain how you do Kernel PCA & kernel LDA (what is the step of it?)
    - **t-SNE**
      - Part1 (10%) Also, show the formula of tsne & ssne
      - Part2 (2%)
      - Part3 (2%)
      - Part4 (1%)
  - **B. Experiments Results (35%) & Discussion (15%)**
    - Show everything we asked you to show (experiment should include both setting and result)
    - **Kernel Eigenfaces**
      - Part1 (5%)
      - Part2 (5%)
      - Part3 (5%) & (5%) Please discuss the observation in this part (You can compare the result with PCA/LDA)
    - **t-SNE**
      - Part1 (5%) & (5%) Please discuss the observation in this part
      - Part2 (5%)
      - Part3 (5%)
      - Part4 (5%) & (5%) Please discuss the observation in this part
  - **C. Observations and Discussion (10%)**
    - Anything you want to discuss, such as the meaning of eigenfaces or trying different dimension reduction methods, comparing the advantages and disadvantages of them. (It is noticed that the score of this part is different from the discussion in Section B. You can summarize the observation or try to discuss more about the project)

---

4. Turning In Your Work

> To Submit the Homework, you should do the following:

- Zip your contents in one file, including
  - Report (.pdf)
  - Source Code
  - Video of GIF images of clustering procedure
- Name the zip file as `ML_HW7_{your student id}_{your name}.zip`. (e.g. ML_HW7_0856XXX_王小明.zip)
  - If the zip file name has format error or the report is not in pdf format, there will be a penalty (we are considering -10).
- Submit your homework *in time*
  - After deadline, you can still sumit in the following 7 days, you will get only 70% of original score.
  - Starting from the seven'th day after the deadline, you can not submit your homework and you will get 0 score.
  - Whenever you submit your homework, the latest submission will be used for grading. (so don't accidently submit something after deadline, you will get 70% discount no matter what)