EEE418: Advanced Pattern Recognition **Spring 2020**

Lab 2: Feature Extraction

Lecturer: Xiaobo Jin *Unit: EEE Dept. of XJTLU*

Disclaimer:

1. Lab reports deadlines are strict. University late submission policy will be applied.

2. Collusion and plagiarism are absolutely forbidden (University policy will be applied).

3. Report is due 14 days from the date of running this lab

2.1 Objectives

- Implement the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) algorithms
- In this experiment, we will use the publicly dataset to verify our algorithm. Download the UCI iris dataset: https://archive.ics.uci.edu/ml/datasets/Iris

2.2 Principal Component Analysis (PCA)

- (**10 marks**) Standardize the data $\hat{X} = X_{n\times d} - \mathbf{1}_{n\times 1}\text{mean}(X)_{1\times d}$
- (**10 marks**) Compute the covariance matrix $\Sigma = \hat{X}^T\hat{X}$
- (**10 marks**) Call the function **numpy.linalg.eig(a)** to find the most r maximum eigvectors listed as C. Transform the training dataset X using C

$$Y = XC$$

Algorithm 1 PCA Algorithm

```
$X$: input $n \times d$ data matrix (each row a d-dimensional sample)
Standardize the data: subtract mean of $X$ from each row of $X$
Compute the covariance matrix of X (along the row of $X$) to obtain $\Sigma = (X - \bar{X})^T(X - \bar{X})$
Find eigenvectors and eigenvalues of $\Sigma$
Compute $r$ eigenvectors with largest eigenvalues to construct the matrix $C_{d\times r}$, where the value of eigenvalues gives importance of each component
Transform $X$ using $C$
    $$Y = XC$$
    where the number of new dimensional is $r$ ($r \ll d$)
```

2.3 Linear Discriminant Analysis (LDA)

- (**10 marks**) Generate $\{X_k\}$ matrix according to the labels of the examples.
- (**10 marks**) Compute the class-wise mean and the total mean of the data matrix.
- (**10 marks**) Compute the within-class covariance matrix S_W
- (**10 marks**) Compute the between-class covariance matrix S_B
- (**10 marks**) Find $K-1$ projection eigvector of $S_B^{-1}S_W$ as the matrix C and transform the data matrix X
$$Y = XC$$
- For numerical computation, the vectorization will be beneficial to the running efficiency of the algorithm. The following examples demonstrate how to standardize the data X with the mean $\boldsymbol{u}$

```
#pseudo code on the vectorization and the non-vectorization
# non-vectorization
for i in range(N):
    X[i,:] -=  u^T

# vectorization
X -=  1 u^T
```

Algorithm 2 Multi-class LDA Algorithm

```
Let $X_1, X_2, \cdots, X_K$ be the data matrices belong to class $i (i = 1, 2, \cdots, K)$ and $n_k$ the rows of the matrix matrix $X_k$
Compute the means of the data matrices $X_k$ to get $\boldsymbol{u}_k$ and the mean of the training data $X$ to get $\boldsymbol{u}$
Compute the within-class covariance matrix $S_W = \sum_{k=1}^{K}(X_k - \mathbf{1}\boldsymbol{u}_k^T)^{(}X_k - \mathbf{1}\boldsymbol{u}_k^T)$
Compute the between-class covariance matrix $S_B = \sum_{k=1}^{K} n_k(\boldsymbol{u}_k - \boldsymbol{u})(\boldsymbol{u}_k - \boldsymbol{u})^T$
Find $K-1$ eigvectors of $S_B^{-1}S_W$ listed as the matrix $C$ that correspond to the $K-1$ largest eigenvalues.
Transform the matrix $X$ as $Y = XC$
```

2.4 Comparisons of PCA and LDA

- Reduce the dataset into 2 dimensional space by PCA and LDA algorithms, separately.
- (**10 marks**) Project the dataset into 2 dimensional space and visualize them with the python matplotlib library.
- (**10 marks**) Compare the visualization of both algorithm and analyze their differences.

2.5 Lab Report

- Write a short report which should contain a concise description of your results and observations to prove that you can understand the algorithm deeply.

- **Please insert the clipped running image into your report for the steps with the marks to prove that you have accomplished that step.**
- Submit the pdf report (no latex .tex file) and the python source code electronically into ICE.
- The report must be written with the latex typesetting language.
- The report in pdf format and python source code of your implementation should be zipped into a single file. The naming of report is as follows:

  e.g. StudentID_LastName_FirstName_LabNumber.zip (123456789_Einstein_Albert_1.zip)

2.6 Hints

Please refer to the lecture slides for more details.

- Latex IDE: texstudio
- Python IDE: pycharm
- Use the python numpy library flexibly.