685.621 Algorithms for Data Science
Homework 2
Assigned at the start of Module 3
Due at the end of Module 5 - Tuesday October 5 (Midnight PST, 3AM EST)
Total Points 100/100

Collaboration groups will be assigned on Tuesday September 14th in Blackboard. Make sure your group starts one thread for the collaborative problems. You are required to participate in the collaborative problem and subproblem separately. Please do not directly post a complete solution, the goal is for the group to develop a solution after everyone has participated. Please ensure you have a write-up with solutions to each problem and subproblems, you are also required to submit code that will be compiled when grading the assignment. In each of the problems you are allowed to use built-in functions.

1. **Problem 1 - Module 3** ***Note this is not a Collaborative Problem***
   30 Points Total

   In this problem, implement code to analyze the Iris data sets by feature and plant class using the test statistics listed in Table 1. In Module 3 under Content in Probability document Table 1 can be used as a reference.

   (a) (5 points) Develop an algorithm to read in the iris.csv file.

   (b) (5 points) Develop an algorithm to store the data in a data structure of your choice, e.g., array, matrix, etc.

   (c) (10 points) Perform statistics of each feature and class using the test statistics listed in Table 1?

   (d) (10 points) Perform analysis and provide an explanation of what each of the statistics provides of the data.

Table 1: Data Analysis Statistics

Test Statistics	**Statistical Function** $\boldsymbol{F}(\cdot)$
Minimum	$F_{\min}(\mathbf{x}) = \min(\mathbf{x}) = x_{min}$
Maximum	$F_{\max}(\mathbf{x}) = \max(\mathbf{x}) = x_{max}$
Mean	$F_{\mu}(\mathbf{x}) = \mu(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
Trimmed Mean	$F_{\mu_t}(\mathbf{x}) = \mu_t(\mathbf{x}) = \frac{1}{n-2p}\sum_{i=p+1}^{n-p} x_i$
Standard Deviation	$F_{\sigma}(\mathbf{x}) = \sigma(\mathbf{x}) = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^2\right)^{1/2}$
Skewness	$F_{\gamma}(\mathbf{x}) = \gamma(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^3}{\sigma(\mathbf{x})^3}$
Kurtosis	$F_{\kappa}(\mathbf{x}) = \kappa(\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu(\mathbf{x})\right)^4}{\sigma(\mathbf{x})^4}$

For clarification, the analysis should be done by feature followed by class of flower type. This analysis should provide insight into the Iris data set. The analysis should be put into tables for easy understanding. The Iris data set is represented by the $[150 \times 4]$ matrix $\mathbf{X}$, $[1 \times 4]$ vector $\bar{\mathbf{X}}$ is the mean of the four features for all observations, $\mathbf{x}_1$ is the $[150 \times 1]$ vector representing the sepal length, $\mathbf{x}_2$ is the $[150 \times 1]$ vector representing the sepal width, $\mathbf{x}_3$ is the $[150 \times 1]$ vector representing the petal length, and $\mathbf{x}_4$ is the $[150 \times 1]$ vector representing the petal width. Taking the notation a step further, let $\mathbf{x}_{1,c}$ represents the vector for sepal length by class $c = [1, 2, 3]$ (Setosa, Versicolor, Virginica), specifically, $\mathbf{x}_{1,1}$ be the $[50 \times 1]$ vector representing the sepal length for class 1 (Setosa), $\mathbf{x}_{1,2}$ be the $[50 \times 1]$ vector representing the sepal length for class 2 (Versicolor), and $\mathbf{x}_{1,3}$ be the $[50 \times 1]$ vector representing the sepal length for class 3 (Virginica).

Note: The trimmed mean is a variation of the mean which is calculated by removing values from the beginning and end of a sorted set of data. The average is then taken using the remaining values. This allows any potential outliers to be removed when calculating the statistics of the data. Assuming the data in $\mathbf{x}_s = [x_{1,s}, x_{2,s}, \cdots, x_{n,s}]$ is sorted, the resulting $\mathbf{x}_{s,p} = [x_{1+p,s}, x_{2+p,s}, \cdots, x_{n-p,s}]$. the trimmed mean allows the removal of extreme values influencing the mean of the data.

2. **Problem 2** ***Note this is not a Collaborative Problem***
15 Points Total

In this problem the goal is to build a set of numerical images from a set of arrays. The data set is from the Kaggle web site will be used: https://www.kaggle.com/c/digit-recognizer/data
This data has a training.csv, test.csv and sample_submission.csv files. In this exercise the focus will be on the train.csv data. The web site has the following data description:

The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

Each pixel column in the training set has a name like pixel x, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as $x = i * 28 + j$, where i and j are integers between 0 and 27, inclusive. Then pixel x is located on row i and column j of a 28 x 28 matrix, (indexing by zero).

or example, pixel 31 indicates the pixel that is in the fourth column from the left, and the second row from the top, as in the ascii-diagram below.

This data is set up in a csv file which will require the reshaping of the data to be 28×28 matrix representing images. There are 42000 images in the train.csv file. For this problem it is only necessary to process approximately 100 images, 10 each of the numbers from 0 through 9. The goal is to learn how to generate features from images using transforms and first order statistics.

(a) (5 points) Develop an algorithm to read in the train.csv file.

(b) (5 points) Develop an algorithm to store the data in a data structure of your choice so that the data is reshaped into a matrix of size 28×28.

(c) (5 points) Plot the developed matrix for indices 1, 2, 4, 7, 8, 9, 11, 12, 17, and 22. These indices represent the numerical values from 0 to 9.

3. **Problem 3 - Module 5** ***Note this is a Collaborative Problem***
   55 Points Total

   In this problem each image from the train.csv is to be processed to generate a set of features using the discrete cosine transform and principal component analysis.

   (a) (10 points) Take the 2 dimensional Discrete Cosine Transform (DCT) of each matrix from Problem 2, the matrix represents each number.

   (b) (10 points) Extract the vertical, horizontal and diagonal coefficients from the transform.

   (c) (10 points) For each of the three sets of DCT coefficients perform Principal Component Analysis (PCA).

   (d) (10 Points) Retain either the top n number of principal components or the top principal components with maximum variance.

   (e) (10 points) Using your top principal components reduce the DCT transformed data. This will create a new data set that represents

   (f) (5 points) Save the new data in a file of your choice, *.txt, *.csv, etc.

References

[1] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006,

[2] Dillon, and Goldstein, M.. *Multivariate Analysis Methods and Applications*, John Wiley, 1984 https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[3] Fisher, R.A., *The use of Multiple Measurements in Taxonomic Problems*, Annals of Human Genetics, Vol. 7, Issue 2, pp. 179-188, 1936

[4] Hotelling, H., *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, Number 24, pp. 417–441, 1933

[5] Rao, K. P. and Yip, P., *Discrete Cosine Transform Algorithms, Advantages, Applications*, San Diego, CA: Academic Press, Inc., 1990