Indian Institute of Technology Madras

ID5055 Foundations of Machine learning

Assignment I

Due date: August 24, 2023

# Instruction

1. Assignment shall be submitted on the due date. Late submissions will not be entertained. If you cannot submit the assignment due to some reasons, please contact the instructor by email.
2. All the assignments must be the student's own work. The students are encouraged to collaborate or consult friends. In the case of collaborative work, please write every student's name on the submitted solution.
3. If you find the solution in the book or article or on the website, please indicate the reference in the solutions.

# Problems

1. Let us consider the MNIST, a dataset of hand-written digits [Click here to download the dataset]. The objective here is to de-noise the images provided. Then, perform the following tasks.
   (a) Load the dataset and visually see the images ($16 \times 16$ pixels).
   (b) Split the obtained dataset into training and testing samples with the training set having 80% of samples, and the remaining 20% of samples in the test set. Use the training set for applying PCA to perform de-noising of the images. Write your own codes to perform PCA using the singular value decomposition (SVD) algorithm. You can use the in-built functions in Python for the SVD.
   (c) Do PCA on the training set and choose the appropriate number of principal components by visually analyzing the de-noised images (You have to reconstruct the entire image from the principal components)
   (d) Plot the randomly chosen de-noised test images and note down your observations.
2. Given that the data is not standardized, obtain the relation between the singular values obtained by applying SVD on the data matrix and the eigenvalues of the covariance matrix.
3. Generate a data matrix, $\mathbf{X}$ with 200 samples and 10 features. (You can assume any distribution).
   (a) Perform PCA on the matrix and get the first 2 principal components for all the samples
   (b) Calculate the sum of the distance between all the samples and the plane formed by the obtained 2 principal components.
   (c) Generate 50 random planes and calculate the sum of distances between the samples and each one of the planes. Verify that the sum distances are the least for the plane obtained from principal components

   IRIS dataset has four features: sepal length, sepal width, petal length, and petal width that can be used to identify the kind of IRIS species.

(a) Perform PCA on the given dataset and reduce the dimension to two

(b) Plot the newly obtained features (Principal components) discriminating the species type

(c) Identify which one of the four features can be used to discriminate between the species (Use the coefficient of principal components to answer)

4. A team of scientists has analyzed 90 samples that are a mixture of four different species with different concentrations. It has been established that each specie has its own pure component fingerprint. Three different spectroscopy instruments (InsA, InsB, and InsC) have been used for the same. The dataset can be found in [Dataset, click here to get dataset]. The rows indicate the samples and the columns indicate the wavelengths at which they have been measured.

   (a) Perform PCA on each of the datasets and choose the appropriate number of principal components for de-noising the dataset. Use the Scree plot to choose an appropriate number of principal components.

   (b) Another 10 samples were collected and analyzed on InsA (Test_data.csv). It is been given that one sample has contaminants in it. Identify which sample among them has been contaminated.

5. Consider a set of $D$ variables, $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$. The data matrix $\mathbf{X} \in \mathbf{R}^{D\times N}$ where $N$ is the number of observations.

   - The singular value decomposition of $\mathbf{X}$ is given by

     $$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T + \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$$

     where $\mathbf{U}_1 \in \mathbf{R}^{D\times M}$ and $\mathbf{V}_1 \in \mathbf{R}^{N\times M}$ are the appropriate eigenvectors corresponding to the largest $M$ singular values, $\mathbf{\Sigma}_1$ is the $M-$dimensional diagonal matrix of the largest $M$ singular values, $\mathbf{U}_2 \in \mathbf{R}^{D\times(D-M)}$ and $\mathbf{V}_2 \in \mathbf{R}^{N\times(D-M)}$ are the appropriate eigenvectors corresponding to the remaining $D-M$ singular values, $\mathbf{\Sigma}_2$ is the $(D-M)-$dimensional diagonal matrix of with the singular values being zeros.
     Then, prove that the $\mathbf{u}_{2,i}^T x = \mathbf{0}$ where $\mathbf{u}_{2,i}$ is the $i$th column of $\mathbf{U}_2$.

   - The singular value decomposition of $\mathbf{X}$ is given by

     $$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

     the matrices have the usual interpretation as the previous part. Let us define a transformation of the data as follows

     $$\mathbf{Z} = \mathbf{\Sigma}^{-1/2}\mathbf{U}^T\mathbf{X}$$

     . Then, show that the covariance of the transformed data ($\mathbf{Z}$) is an identity matrix.