Assignment 2 Due Date 26/10/2023

## Python/Numpy basics

Dataset download and learn about characteristics

1. Download wheat seeds Dataset from here.
2. Read the dataset using panda data frame. Convert to numpy array
3. Plot the data points using the first two dimensions (area, perimeter). Use three different shapes (triangle, square, circle) to plot datapoints  for three different classes. You should use the class information from class label and use them when you decide on shapes (see slide 6 and slide 10)
4. Calculate the mean data point for each class and show them with similar shape with the larger size on the above plot.
5. Calculate the **centered version** of area, perimeter. Plot them in a new plot with tile: “Centered data”
6. Now, plot a line (l) in this plot with line equation on the plot done in step 5. The line is l  = span{$[-1.75\ 1.75\ ]$}. Therefore, the equation of the line is:  $x_1 = - x_2$ (See slide 10)
7. Now calculate the projection of each data points on the line l (spanned by the vector $[-1.75\ 1.75\ ]$). And plot the projected point on the line using the same shape but smaller size. So all smaller shapes would be on the line. (See slide 10)

In 3D:

1. Plot the data points using the first three dimensions (area, perimeter, compactness) .
2. Use three different shapes (triangle, square, circle) to plot datapoints  for three different classes. You should use the class information from class label and use them when you decide on shapes (see slide 6 and slide 10)
3. Calculate the mean data point (3D) for each class and show them with similar shape with the larger size on the above plot.
4. Calculate the **centered version of**   area, perimeter, compactness. Plot centered data points in 3D in a new plot.
5. Now, plot the plane (with yellow plane) spanned by two normal vector( $[1 - 2,\ 1\ ]^T$, $[2, 1, 0]^T$ )  on the plot done in step 4.
6. Now calculate the projection of each data point on the plane  (spanned by the vector ( $[1 - 2,\ 1\ ]^T$, $[2, 1, 0]^T$ ). And plot the projected point on the plane  using the same shape but smaller size in the plot done in step 4. So all smaller shapes would be on the line  (See slide 10)

## Numeric Data Analysis

1. Download wheat seeds Dataset from here.
2. Read the dataset using panda data frame. Convert to numpy array.
3. Discard the last class label column. We don’t need them.

Write a script to answer the following questions.

1. Compute the multivariate mean vector

2. Compute the sample covariance matrix as inner products between the columns of the centered data matrix (see **Eq. (2.38)** in chapter 2).
3. Compute the sample covariance matrix as outer product between the centered data points (see **Eq. (2.39)** in chapter 2)
4. Compute the correlation between Attributes 1 and 2 by computing the cosine of the angle between the centered attribute vectors. Use vector notations shown in class. Plot the scatter plot between these two attributes.

**Analytical Problem:**

Let $\overline{\boldsymbol{D}}$ represent the centered data matrix

$$\overline{\boldsymbol{D}} = \boldsymbol{D} - \boldsymbol{1} \cdot \hat{\boldsymbol{\mu}}^T = \begin{pmatrix} \boldsymbol{x}_1^T - \hat{\boldsymbol{\mu}}^T \\ \boldsymbol{x}_2^T - \hat{\boldsymbol{\mu}}^T \\ \vdots \\ \boldsymbol{x}_n^T - \hat{\boldsymbol{\mu}}^T \end{pmatrix} = \begin{pmatrix} - & \overline{\boldsymbol{x}}_1^T & - \\ - & \overline{\boldsymbol{x}}_2^T & - \\ & \vdots & \\ - & \overline{\boldsymbol{x}}_n^T & - \end{pmatrix}$$

**Show that**

**a.**

$$\frac{1}{n}\left(\overline{\boldsymbol{D}}^T\,\overline{\boldsymbol{D}}\right) = \frac{1}{n}\begin{pmatrix} \overline{X}_1^T\overline{X}_1 & \overline{X}_1^T\overline{X}_2 & \cdots & \overline{X}_1^T\overline{X}_d \\ \overline{X}_2^T\overline{X}_1 & \overline{X}_2^T\overline{X}_2 & \cdots & \overline{X}_2^T\overline{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X}_d^T\overline{X}_1 & \overline{X}_d^T\overline{X}_2 & \cdots & \overline{X}_d^T\overline{X}_d \end{pmatrix}$$

b.

$$\frac{1}{n}\begin{pmatrix} \overline{X}_1^T\overline{X}_1 & \overline{X}_1^T\overline{X}_2 & \cdots & \overline{X}_1^T\overline{X}_d \\ \overline{X}_2^T\overline{X}_1 & \overline{X}_2^T\overline{X}_2 & \cdots & \overline{X}_2^T\overline{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X}_d^T\overline{X}_1 & \overline{X}_d^T\overline{X}_2 & \cdots & \overline{X}_d^T\overline{X}_d \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n}\overline{\boldsymbol{x}}_i \cdot \overline{\boldsymbol{x}}_i^T$$

Where $\overline{X}_1$ $\overline{X}_2$ ….. $\overline{X}_d$ are centered attribute vectors. N.B You should use pen and paper to show the LHS = RHS for a and b here. Use vector notations shown in class. Take the image, make pdf and submit with your ipynb file.