SEIS 763 Machine Learning
Assignment 5

Individual effort

You will be implementing dimensionality reduction techniques on the dataset provided.

Dataset:
- First two columns are the features.
- Last column is the class label for every instance.

What you need to do: Create a jupyter notebook called **Assign5.ipynb**. Write code for each of the following questions by having a separate cell for every question. Copy the actual question in a markdown cell and right below that you should have a code cell.

1) Load the dataset and display a scatter plot with all the data instances. Column 1 needs to be plotted on the X-axis and Column 2 on the Y-axis. Color code the points based on their class label. That is, instances of the same class should be displayed with the same color. Another color needs to be used for instances of the second class. This plot will show you the distribution of the two classes.

Split the dataset into training and testing (70-30 split). To do this, when you call the train_test_split function make sure you specify the random_state parameter like this, **random_state=0**. This is to ensure that we all get the same train and test splits.

2) Normalize the data using standardization. Fit a logistic regression model. What is the model accuracy on the test set?

3) Perform PCA on the dataset and retain 2 principal components. Visualize the 2-dimensional training data with a scatter plot like before. (Same color for class labels as above).

4) Fit a logistic regression model on this dataset. What is the model accuracy on the test set?

5) Now perform PCA and retain only one principal component and visualize the training data like before. This time your plot should be a horizontal line since you have just one dimension.

6) Fit a logistic regression model on this dataset. What is the model accuracy on the test set?

7) Perform LDA. Your dataset should have just one dimension. Visualize this training data like before.

8) Fit a logistic regression model on this dataset. What is the model accuracy on the test set?

9) Perform Kernel PCA and retain two components. Visualize this training data like before.

10) Fit a logistic regression model on this dataset. What is the model accuracy on the test set?

Submission:
- Make sure each of the cells have been run along with the output shown right below. Now, export the notebook as .html file.
- Submit the **.html** file and **.ipynb** notebook on Canvas.

Note: Do not submit the data. Your code should be referencing the data file when you load the data in your code.