Week2 - Introduction to NumPy, MatPlotLib, Pandas

Ungraded Lab for DS3010 - Machine Learning

# OVERVIEW & PURPOSE

Introduction to NumPy, MatPlotLib, Pandas, and Data preprocessing

# Instructions

1. Please submit the assignment through Moodle in .ipynb format (python notebook)
2. The submission should contain a single notebook containing all the solutions, including the requested documentation, observations, and findings.
3. You must adequately comment the code to improve its readability.
4. This ungraded lab is due on Aug 18th 5.00pm

# Lab

1. **Numpy**
   1. Create a 2D array with size 4 $$ \times $$ 100 such that
      1. Fill the first row with random values with approximate mean 2
      2. Fill the second row with integers in increasing order.
      3. Fill the third row with randomly chosen alphabets.
      4. Fourth with randomly formed alphanumeric strings of length 4.
   2. Create a function to operate the array such that numeric values are squared and strings are doubled.
   3. Add another row with a linear combination of the first two rows.
   4. Transpose the above array to 100 $$\times$$ 5.
2. **Pandas**
   1. Create the Pandas DataFrame from the above created numpy array with the first 4 columns.
   2. Perform standardization on numerical valued columns using the function created by you. (Standardization = (X – μ) / σ μ = mean σ = Standard deviation. You can try to use nested functions too.)
   3. Add the fifth column as a randomly chosen label among three labels of your choice.
   4. Now consider this dataframe represents your classification data.
   5. Plot the data from the first column vs second column with different colors where color represents its class.
3. **Data Preprocessing**
   1. Download the dataset in the csv file accompanying the pdf.
   2. Read the dataframe using pandas.
   3. Among date columns, keep any one column and delete others.
   4. Perform all preprocessing steps needed such as deleting unwanted columns, changing the format of some column data and other required changes. Take care of missing data. Justify your choices.
   5. After preprocessing, recognize the output column and split the data into input and output. Further, split the dataset into train/test.