School of Computer Science Engineering and Technology

Course- BTech	Type- Core
Course Code- 301	Course Name-AIML
Year- 2022	Semester- Even
Date- 24-02-2022	Batch- IV Sem Spl

7 - Lab Assignment # No. (7.1)

Objective: In this lab, students will use PCA to reduce a dataset to a smaller number of dimensions. The objective is for students to:

- Understand what PCA is and why it's useful
- Feel comfortable performing PCA on a new dataset
- Understand what it means for each component to capture variance from the original dataset
- Be able to extract the `variance explained` by components.
- Perform modelling with the PCA components

1. **Dataset:** Download the wine quality dataset from the link https://archive.ics.uci.edu/ml/datasets/wine+quality (5)
2. **Data Pre-processing:** For this assignment, let's say that the wine expert is curious if s/he can predict, as a rough approximation, the categorical quality -- bad, average, or great. Let's define those categories as follows:
   i. bad is when for wines that have a quality <= 5
   ii. average is when a wine has a quality of 6 or 7
   iii. great is when a wine has a quality of >= 8 (10)
3. **Data Splitting:** After the range normalization, it's time to split the data into training and testing. Split data into 80:20 ratio. (5)
4. **Standardization:** Since the features are in different ranges, each column can be normalized into 0 to 1 using different methods such as scaling, standardizing etc. Note: Normalization should not be done for the target feature. (5)
5. **Model Construction:** Train logistic regression model and test the model accuracy on train and test subsets. Also check the precision, recall, F1-score and ROC-AUC score of the model. (5)
6. **Dimensionality Reduction:** In attempt to improve performance, we may wonder if some of our features are redundant and are posing difficulties for our logistic regression model. Let's PCA to shrink the problem down to 2 dimensions (with as little loss as possible). Check the explained_variance_ratio of PCA components. (10)
7. Now again construct the Logistic regression model on 2 dimensions returned by PCA and check its score: accuracy, precision, recall, F1_score, etc. (5)
8. Fit a PCA that finds the first 10 PCA components of our training data. Use `np.cumsum()` to print out the variance we'd be able to explain by using n PCA dimensions for n=1 through 10. Again train a Logistic Regression model on this 10d data and check its score. (10)

For Fun: Import load_digits the dataset from sklearn.datasets. Use PCA to reduce the dimensions from 64 to some smaller number. Plot cumulative explained variance ratio of PCA components to find out how many components will give more than 90% of the information. Also, impute some noise in this dataset and then use PCA for removal of noise.