Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603 110
(An Autonomous Institution, Affiliated to Anna University, Chennai)

UCS2612 Machine Learning Laboratory

Academic Year: 2023-2024 Even **Batch: 2021-2025**
Faculty In-charges: Y.V. Lokeswari & Nilu R Salim **VI Semester A & B**

A. No. : 9. **Applications of dimensionality reduction techniques**

Download the Wine Quality dataset from the link given below:

https://archive.ics.uci.edu/dataset/186/wine+quality

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: http://www.vinhoverde.pt/en/ or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

The data can be used to test (ordinal) **regression** or **classification** (in effect, this is a **multi-class** task, where the clases are **ordered**) methods. Other research issues are **feature selection** and **outlier detection**. The data includes two datasets:

- winequality-red.csv - red wine preference samples;
- winequality-white.csv - white wine preference samples;

The **datasets** are available here: winequality.zip

Quality is the target regression column.

Develop a python program to perform dimensionality reduction using PCA and LDA. Visualize the features from the dataset and interpret the results obtained by the model using Matplotlib library. **[CO1, K3]**

Use the following steps to do implementation:

1. Loading the dataset.
2. Pre-Processing the data (Handling missing values, Encoding, Normalization, Standardization, Outlier Detection).
3. Exploratory Data Analysis.
4. Feature Engineering techniques.
5. Split the data into training, testing and validation sets.
6. Train the model.
7. Test the model.
8. Measure the performance of the trained model.
9. Compare the results of each ensemble model using graphs.
10. Represent the ROC of training and test results in the graphs.

Upload the code in GitHub and include the GitHub main branch link in the assignment PDF.

Hints to do the assignment:

Do the following:

1. Load the dataset.
2. Pre-Processing the data (Handling missing values, Encoding, Normalization, and Standardization).
3. Exploratory Data Analysis
4. Feature Engineering techniques.
   Refer to
   https://machinelearningmastery.com/feature-selection-machine-learning-python/
   https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
   https://www.datacamp.com/tutorial/feature-selection-python

5. Apply dimensionality reduction techniques on the dataset and obtain the reduced feature set. Apply PCA and LDA techniques and evaluate the reduced features with linear / logistic regression model.
6. Combine the red and white wine dataset together. Include the class label as red / white. Apply PCA and LDA techniques and evaluate the reduced features with any classification algorithm.
7. Compare the results of both techniques by evaluating the features using classification / regression algorithm.

   Refer to the following sources.

PCA

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

https://www.javatpoint.com/principal-component-analysis-with-python

https://github.com/tirthajyoti/Machine-Learning-with-Python/blob/master/Clustering-Dimensionality-Reduction/Principal%20Component%20Analysis.ipynb

https://www.geeksforgeeks.org/principal-component-analysis-with-python/

https://www.kdnuggets.com/2023/05/principal-component-analysis-pca-scikitlearn.html

LDA

https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html

https://developer.ibm.com/tutorials/awb-implementing-linear-discriminant-analysis-python/

https://www.statology.org/linear-discriminant-analysis-in-python/

https://www.mltut.com/linear-discriminant-analysis-python-complete-and-easy-guide/

8. Upload python project in GitHub and explore all git commands. Git Commands Tutorial : https://git-scm.com/docs/gittutorial

   Upload IPython to GitHub
   https://reproducible-science-curriculum.github.io/sharing-RR-Jupyter/01-sharing-github/

Additional Reference:
https://www.youtube.com/watch?v=LlrKTV4-ftI

Upload the code in GitHub and include the GitHub main branch link in the assignment PDF.

Upload python project in GitHub and explore all git commands.

Git Commands Tutorial : https://git-scm.com/docs/gittutorial