Applied AI & ML Industry Projects Lab (AAM-IPL)
Transforming Learning into Industry Solutions

Brillium Technologies
Transforming Lives Through Technology

AAM IPL Week 8

LDA/GDA –Inferring Breast Cancer/Malignancy from Diagnosis Data and Classification Models Comparative Analysis

B.Tech – CSE(AIML)
V Semester - ML and AIUP, Sept-Nov 2024
Department of Computer Science Engineering – AI and ML (CSM)
G.Pulla Reddy Engineering College (Autonomous), Kurnool, AP

Algorithm of Application

LDA/GDA

Project Title

Inferring Breast Cancer/Malignancy from Diagnosis Data and Classification Models Comparative Analysis

Project Objective

Implement LDA/GDA for inferring breast cancer disease and to analyze all other classification models performance you have learnt in AAM-IPL on the Breast Cancer Wisconsin Dataset from scikit-learn.

Optionally, use wine data set from scikit-learn and see the comparative performance of all classification models you have learnt in AAM-IPL.

Dataset – Breast Cancer

- Description:
  - Implements an LDA/GDA classifier to predict the class of breast tumor (breast cancer) from the provided dataset – malignant or benign.
  - The dataset should be used for training a SVM classifier and evaluate its performance using various metrics such as accuracy, precision, recall, F1-score.
- Dataset Details:
  - The breast cancer dataset consists of 569 samples, each representing a patient with a set of features
- Data Source/Published By:
  - Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository
  - [PDF] Nuclear feature extraction for breast tumor diagnosis | Semantic Scholar
  - Supported by CS Department, University of Wisconsin-Madison
- Features & Data Download Link
  - Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository

- Alternatively, you can load the data directly using sklearn datasets as below

```
from sklearn.datasets import load_breast_cancer
import pandas as pd

# Load the dataset
data = load_breast_cancer()
X = pd.DataFrame(data.data,
columns=data.feature_names)
y = pd.Series(data.target, name="target")
```

Dataset – Wine Data

- Description
  - The Wine dataset contains information about chemical analysis of wines from the Italian region of Piedmont, which are produced by three different cultivars. This dataset is widely used for classification tasks, particularly for distinguishing wine types based on chemical properties.
- Details
  - Number of Instances: 178
  - Number of Features: 13 (not including the target class)
  - Target: The dataset has 3 classes, corresponding to three wine cultivars (Class 0, Class 1, and Class 2)
- Data Source/Published By
  - Published By: This dataset was made available by the UCI Machine Learning Repository.
  - Original Source: Forina, M., et al. (1991) in "PARVUS - An Extendible Package for Data Exploration, Classification and Correlation".
- Features
  - Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids
  - Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue
  - OD280/OD315 of diluted wines, Proline
  - Each feature represents a chemical property used to classify the wine samples
- Data Download Link
  - You can download the Wine dataset from the UCI Machine Learning Repository: UCI Wine Dataset
  - Alternatively, you can load the data directly using sklearn datasets as below

```
from sklearn.datasets import load_wine
import pandas as pd

# Load the dataset
data = load_wine()
X = pd.DataFrame(data.data,
columns=data.feature_names)
y = pd.Series(data.target, name="target")
```

Implementation Steps

1. Import necessary libraries
2. Load and preprocess dataset
3. Split dataset into training and test sets
4. Standardize features
5. Define models
6. Initialize results list
7. Define function to add watermark to plots
8. Train models and evaluate
9. Convert results to data frame
10. Plotting metrics and times with AAM-IPL watermark
11. Plot combined ROC curves with AAM-IPL watermark
12. Plot combined confusion matrices with AAM-IPL watermark
13. Generate the PDF of code and output of project Jupyter file

Project Files Provided

- Project shell code file - ***AAM-IPL-Wk-8-LDA-GDA-Breast-Cancer-Shell-Code-V1.ipynb*** and ***AAM-IPL-Wk-8-LDA-GDA-Wine-Classification-Shell-Code-V1.ipynb***
- Training Data – **Not Provided**
- Watermark image for plots - ***AAM-IPL-Watermark-for-Plots.png***

Project Overview, Implementation and Submission Timeline

09-11-2024 – Saturday 10:30 AM – Next Week Project Details Announcement – Topic, Data Set, Shell Code etc. **Announcement Channels – Google Class, Industry Projects WhatsApp Group.**			
14	09-11-2024 - Saturday Duration: 1.5 Hrs	Linear/Gaussian Discriminant Analysis – L/GDA – Model Overview/Recap, Project Description, and Interactive Q&A	Online – Google Class
15	10-11-2024 - Sunday Duration: 1.5 Hrs	Linear/Gaussian Discriminant Analysis – L/GDA – Model Building, Output Demonstration, Q&A	Online – Google Class
14-11-2024 – Thursday 11:59 PM - Deadline to upload the project code submission by all students in Google Class.			

Guest Lecture Timings:
Saturdays: 10:30 AM IST – 12:00 Noon IST
Sundays: 10:30 AM IST – 12:00 Noon IST
Mondays: 6:30 PM IST – 8:00 PM IST

Development Environment

- Computing Language – Python
- IDE – Visual Studio Code with Jupyter Notebook

Instructor

Instructor	
Venkateswar Reddy Melachervu (alumnus of GPREC, ECE Class of '92) CTO, Brillium Technologies, Bengaluru	Visiting Faculty

Email: **venkat.reddy.gf@gprec.ac.in** Profile: LinkedIn	

Coordination

All the activities of this programme – lecture venues, weekly projects details announcements, general announcements, changes in lecture timings, etc. will be coordinated by CSM faculty member Sri V.Suresh.

Channels of Communication and Announcements

- Google Classroom
- Whatsapp group **- Applied AI & ML Industry Projects Lab**
- Emails (Strictly GPREC email addresses only)

Programme Coordinator	
Prof. V.Suresh Email: **vsuresh.ecs@gprec.ac.in**	Faculty Member, CSM

Reference Books

- Pattern Recognition and Machine Learning by Chris Bishop, 2006 – PDF Link
- Machine Learning using Python by Manaranjan Pradhan and U Dinesh Kumar, Wiley 2019 – PDF Link

Policies

- **Attendance:** All sessions are expected to be attended by all the enrolled students. In case of inability to attend, prior information is expected to be provided by the student to the coordinator with a copy to the visiting faculty
- **Project Submissions:** Duly completed projects (Jupyter NB file and a PDF of the Jupyter NB file) are expected to be submitted through google class prior to the deadline. In case of inability to complete due various unforeseen circumstance, students are expected to seek extension for the submission deadline.
- **Academic Integrity:** Students are expected to uphold the highest standards of academic integrity in all assignments for the Applied AI & ML Industry Projects Lab. Each assignment must be the student's own work, and all sources and collaborators must be properly acknowledged. By submitting their completed project source code, students confirm that they have adhered to this integrity policy and completed their work in an honest and ethical manner.

Additional Information

For students interested in engaging with special projects in the field of Gen AI, please reach out to the visiting faculty at **venkat@brillium.in** for further details and opportunities.

Contact Information

For any questions or concerns or further details on this programme, please contact **Program Coordinator** during office hours or via email.

-- End of the Document--