Shahid Beheshti University
Machine Learning

Assignment 3

Theoretical Section

Question 1: What is the difference between likelihood and probability? In logistic regression, why can't we use the ordinary least squares (OLS) method for parameter estimation? Provide a detailed theoretical justification, explaining the maximum likelihood estimation (MLE) approach instead.

Question 2: Why is accuracy not always a reliable metric for classification? Discuss alternative metrics such as precision, recall, and F1-score. Provide a scenario where precision is significantly more important than recall, and explain why.

Question 3: Explain the difference between Linear Discriminant Analysis (LDA) as a classifier and LDA as a dimensionality reduction technique. Explain why LDA is a special case of Naive Bayes.

Question 4: Under what assumptions do Linear Discriminant Analysis and Logistic Regression outperform others? Why are these assumptions critical to their performance?

Question 5: Suggest at least four techniques to address the problem of imbalanced datasets in classification. Explain each method thoroughly, including when and why it should be applied.

Question 6: What is the difference between one-vs-one and one-vs-all multiclass classification approaches in classifiers? Under what circumstances would you use one over the other?

Practical Assignment: Campus Recruitment Prediction

Objective

Develop a classification model to predict whether a student will be placed during campus recruitment, using academic, demographic, and institutional features. You are required to test multiple classification models, compare their performance, and identify the most effective approach.

Dataset Overview

The Campus Recruitment Dataset [Link] comprises records from placement drives, including various features such as:

- Academic data (e.g., percentage scores)
- Demographic information (e.g., gender, work experience)
- Institutional attributes
- Target variable: **status** (placement outcome)

Assignment Tasks

1. Data Preprocessing and Exploratory Data Analysis (EDA)

Clean the dataset. This process includes handling missing values, addressing outliers, finding meaningful relationships between features, analyzing feature importance, and more. At the end of the EDA section, you must demonstrate a deep understanding of the data in your report. Finish by normalizing the data and encoding categorical columns with appropriate encoding methods.

2. Model Development and Evaluation

Apply classification models on your data, which is already splitted into 80% train and 20% test, to predict placement.

Models to implement:

- Logistic Regression
- Naive Bayes
- Linear Discriminant Analysis (LDA)

You may test additional classification models, but these are the minimum requirements.

Evaluation Metrics:

- Accuracy
- Precision, Recall, F1-score

- ROC-AUC

Report each metric for both training and testing datasets for all models.

Interpretation & Visualization:

- Confusion Matrix
- ROC Curves
- Precision-Recall Curves

You must interpret the plots and provide insights from comparing the models' performances.

3. Performance Enhancement

- Choose the best-performing model.
- Apply techniques to improve performance. Possible approaches include:
  - Feature Engineering
  - Regularization (e.g., L1, L2)
  - Cross-Validation
- Justify your choices with clear explanations and measurable results.

Submission Guidelines

Your final submission must include:

1. A PDF with answers to theoretical questions.
2. A Jupyter Notebook with code and all outputs clearly presented.
3. A well-documented PDF report summarizing your methodology, findings, and conclusions.