Lab 9: Classification

Breast cancer remains one of the most prevalent and impactful forms of cancer worldwide, affecting millions of individuals and their families. Early detection and accurate diagnosis are critical components in improving survival rates and outcomes for those diagnosed with this disease. Traditionally, the classification of breast tissue samples as benign or malignant has relied heavily on the expertise of pathologists. However, this manual approach can be time-consuming and is subject to variability based on the individual pathologist's experience and expertise. This is where Computer-Aided Diagnosis (CAD) systems come into play, offering a significant advantage by providing support tools that enhance the diagnostic accuracy and efficiency of medical professionals. The development and implementation of CAD systems for breast cancer classification leverage advanced machine learning algorithms, such as Logistic Regression and Random Forest, to analyze and interpret complex data from tissue samples. By training models on vast datasets of known outcomes, these systems can learn to recognize patterns and indicators of malignancy with a high degree of precision.

You need to use "wdbc.data.csv" and "wdbc.names.csv' for this lab. More information about this dataset can be found here.

Part I

Data Description

1 Dataset Overview

[**Write-up 1**] Describe the two CSV files associated with this lab, including the type of information each contains. Detail the specific data each file presents and its relevance to the lab objectives.

2 Loading the Dataset

Load the "wdbc.data.csv" file. Considering our goal is to classify the malignancy of breast tissue:

- [**Write-up 2**] Identify the target variable in this CSV file. What does this variable represent?
- [**Write-up 3**] Analyze and report the distribution of the target variable. Based on this distribution, discuss whether there is a class imbalance problem in this dataset.

3 Feature Analysis

- [**Write-up 4**] Determine and report the total number of predictive features included in the dataset.
- [**Write-up 5**] What is the total number of instances (rows) in the dataset?
- [**Write-up 6**] Evaluate whether the dataset faces a 'curse of dimensionality' issue, considering the number of features relative to the number of instances.

4 Missing Values and Outlier Detection

- [**Write-up 7**] Investigate the dataset for any missing values. Provide a summary of any findings.
- [**Write-up 8**] Select one outlier detection method discussed in this course and apply it to the dataset. Discuss whether any outliers were identified. If you find any potential outliers, explain their possible impact on the classification task.

Part II
Data Splitting and Preprocessing

5 Data Splitting Strategies

- [**Write-up 9**] Discussion on Splitting Strategies: For any classification task, describe how you would split your data using the hold-out method. Compare and contrast the hold-out method with the k-fold cross-validation method. Specifically, under what circumstances would we prefer to use k-fold cross-validation over the hold-out method?

6 Implementing 5-Fold Cross Validation

- [**Write-up 10**] Description and Implementation: For this specific dataset, explain how you would apply 5-fold cross-validation to split your data. After describing the process, implement this data splitting strategy in code. Ensure your explanation covers how the dataset is divided into folds, how each fold is used for training and validation across the iterations.

7 Feature Normalization and Model Implementation

- [**Write-up 11**] Normalization Necessity: Discuss whether feature normalization is necessary when using logistic regression and random forest classifiers for this dataset. Explain your reasoning.

- [**Write-up 12**] Assume normalization is required, discuss whether it should be performed before or after data splitting and why. How does the timing of normalization affect the training and validation processes?
- Implementation in Code: Based on your decision regarding feature normalization, implement the appropriate preprocessing steps in code. If normalization is performed, ensure it is applied correctly within the context of 5-fold cross-validation.

Part III

Model Implementation, Evaluation, and Analysis

8 Implementing Models with 5-Fold Cross-Validation

- Model Implementation: Implement logistic regression and random forest classifiers using the dataset prepared in Part 2. Employ 5-fold cross-validation to train and validate your models, ensuring each fold serves once as the validation set. What are the adjustable settings or parameters in the code for logistic regression and random forest models?
- [**Write-up 13**] Recording Performance Metrics: Record accuracy, sensitivity, specificity, and AUC (area under the curve) for both models for each fold.

9 Statistical Analysis of Model Performance

- Comparative Analysis: Calculate the mean and standard deviation of accuracy, sensitivity, and AUC for both models across all folds.
- [**Write-up 14**] Statistical Testing: Choose an appropriate statistical test to compare the two models. Discuss your choice of test and explain why it is suitable for this analysis. Is there a significant difference in performance (focusing on accuracy, sensitivity, and AUC) between the models? Interpret the results.

Part IV

Feature Importance Analysis

10 Analyzing Feature Importance

- [**Write-up 15**] Discuss strategies to extract the top three most important features from the random forest and logistic regression models. Implement your strategies in code. For both models, do they agree on the top three most important features?

Part V

Submission Instructions

- The grading rubric for this lab will be available in Canvas.
- Please submit both your source Jupyter Notebook file(s) and the write-up PDF file. **Failure to submit either of these two files will result in a grade of 0% for the assignment.**
- Late submission policy: Late submissions will incur a penalty, deducting from your earned points as follows: 5% for one day late, 10% for two days late, and 15% for three days late. **Submissions will not be accepted after three days.**