Statistical Machine Learning 2024

Final Project Specification

In the final project, you are required to choose any two algorithms from either the classification or clustering approaches and apply them to a given dataset. Follow these guidelines:

- Divide the dataset for training and testing/validation, ensuring that the testing dataset does not include samples from the training set. Consider the following points in preparing the report:

    1. Provide a summary of the two selected classification algorithms, identifying critical parameters. Compare the algorithms and discuss the advantages and disadvantages of each.
    2. Describe the data set you have selected. What type of data does it include? Identify the target/classes and the measurements/features within the data.
    3. Explain your preprocessing steps, including feature extraction, data cleaning, PCA, and normalization. Plot the data for each preprocessing steps.
    4. Optimize critical parameters discussed in "point 1" and document the optimization process.
    5. Perform proper validation and address issues of overfitting or underfitting.
    6. Analyse the results thoroughly, offering clear explanations. Provide information about the computational time required.
    7. Identify potential areas for further improvement in the project. What enhancements could be made to the methodology, data handling, or analysis techniques?
    8. Visualize the training process. Provide scatter plots or bar plots, a confusion matrix, and validation plots such as ROC curves and error plots. Explain the insights each type of visualization offers about the model's performance.
    9. Detail how the selected classification algorithms differ from others commonly used algorithms.
    10. Explain any assumptions to the algorithms selected. Discuss how these assumptions might affect the results.
    11. Identify which features were most important for the classification decision and discuss why these features play a critical role.
    12. Analyze how class imbalance affects the performance of the classifiers and describe strategies used to handle it, such as resampling techniques or specific algorithmic adjustments.

- The report may be a topic of discussion during the oral exam, and feedback might be provided before the exam begins. Therefore, please ensure that the report is well-executed.