Uppsala University

Department of Information Technology

Introduction to Machine Learning

Forest Cover Type Classification Project

January, 2023

In this project, you will tackle the Covertype dataset [1]. The dataset includes a collection of attributes related to the Roosevelt National Forest in Colorado. There are both categorical and continuous variables, such as elevation, slope, aspect, soil type, and wilderness area. The goal of the project is to predict the forest cover type (one of seven classes) based on these features.

Your primary objective is to explore the effectiveness of various machine learning algorithms and preprocessing techniques in classifying the forest cover types accurately. This project emphasizes the significance of feature extraction, preprocessing, and their impact on model performance. You are expected to use the knowledge that you have acquired in the course about classification algorithms, to come up with one model that you think is suited for this problem and which you decide to put 'in production'. You are evaluated on your final report. The report should be about 10 pages. In the report you should describe and interpret your results (include lots of graphs and plots). Larger groups are expected to do more work than smaller groups.

You will not work with the original Covertype entire dataset but a random selection of instances from this one. This data file can be found in the GitHub repository for the assignments under the name of `data.csv`.

How to structure your report

1. **Introduction**

   Briefly introduce the Covertype dataset and its relevance. Highlight the classification problem of predicting forest cover types.

2. **Feature Extraction and Preprocessing**

   What preprocessing steps have you done and what features did you use? Emphasize the pivotal role of feature extraction and preprocessing in enhancing model performance. Also discuss about feature scaling and why it is mandatory for algorithms such as k-NN.

3. **Imbalanced Data Consideration**

   Acknowledge and address potential class imbalances in the dataset. Suggested techniques like oversampling, undersampling, or using synthetic data to mitigate this issue. Note that these techniques may not be suitable for this problem and therefore not necessary.

4. **Algorithm Exploration**

   Implement different machine learning algorithms introduced in the course or not and see how well they compare in this problem. Remember that you must do your own training and test split.

   If you picked any algorithm that was not described in the course, you must describe this briefly. Even if you do not implement the algorithm and only use some library you still have to explain the algorithm if it was not covered in the course.

5. **Hyperparameter Tuning**

   Encourage experimentation with hyperparameter tuning to optimize algorithm performance by exploring techniques such as grid search or randomized search for parameter optimization. Recall that these techniques use cross validation to select the hyperparameters. Explain why this is a good approach.

6. **Evaluation Metrics**

   Emphasize the importance of selecting appropriate evaluation metrics for classification tasks. Suggest metrics like accuracy, precision, recall, and F1-score for a comprehensive assessment and model comparison.

7. **Conclusion**

   This project focuses not only on machine learning algorithms but also on the crucial steps of feature extraction and preprocessing. By comparing multiple algorithms and trying various preprocessing techniques, you will gain valuable insights into the interplay between data preparation and model performance. This approach aims to provide a deeper understanding of the nuances involved in real-world machine learning tasks.

Each group will be assigned an assistant is there to help with ideas, and to keep you on track.

References

[1] J. Blackard, "Covertype." UCI Machine Learning Repository, 1998. https://doi.org/10.24432/C50K5N.