BIG DATA SCIENCE

Feature Selection and Data Classification Assignment

Spring 2024

Objective

The primary objective of this task is to apply data mining techniques to the Surveillance, Epidemiology, and End Results (SEER) Public-Use Data to predict the survivability rate of breast cancer patients. This involves developing a predictive model that can accurately classify patients into survivability outcomes based on a set of features derived from the SEER database. This assignment will provide you with hands-on experience in using feature selection and classification algorithms for predictive analytics. By the end of this assignment, you should be able to -

1. Perform data cleaning, validating and preprocessing
2. Apply Feature Selection and Feature Ranking
3. Benchmark various algorithms
4. Perform hyperparameter searches
5. Maintain clean and understandable code

For your reference, we have uploaded a research paper. Please go through it.

Step 1: Preprocessing [15]

Look through your data for outliers, perform standardization/normalization and handle missing values. Use dimensionality reduction if your dataset has a lot of features.+

Step 2: Modeling [15] – For this step you can use tools and/or libraries

Apply the Feature Selection and Feature Ranking Techniques we covered in class and/or a combination of both approaches.

Train the following algorithms on your dataset (feel free to experiment with more!) -

1. KNN (this should be implemented from scratch, do NOT use in-built libraries)

2. Naïve Bayes
3. C4.5 Decision Tree
4. Random Forest
5. Gradient Boosting

You can experiment with neural networks too and see if you achieve better performance. NOTE: For each model used, be sure to include a 1-2 line summary as well as the pros and cons of each algorithm and list out its main hyperparameters.

Step 3: Hyperparameter Tuning [15]

Pick any 2 of the above algorithms that contain at least 2 hyperparameters and perform a hyperparameter search using either Grid or Random search. Display the performance metrics and conclude which set of hyperparameters worked the best.

Step 4: Results [5]

Display your results using a table and explain whether you were able to answer your initial question or not. (Note: Points will not be deducted for poor results as long as the processes followed were sound).

If the models you have used allow it, present what were the most important features used in the classification.

Submission Instructions

1. This assignment can be done either in Python or Weka, the choice is yours.
2. Upload your code to a GitHub repository and make it public. Make sure it is clean and well documented
3. Create a report which includes -
   (a) Feature engineering and preprocessing done
   (b) Results from all the models and the feature selection/ranking in the form of a table
   (c) Results from the hyperparameter search
   (d) Conclusions

The report should not exceed 4-5 pages.

Helpful Links

- Hyperparameter Search
- KNN Overview
- Intro to Decision Trees
- Intro to Random Forests
- Intro to Gradient Boosting
- Feature Selection in ML
- Feature Selection Algorithms
- Intro to Feature Ranking
- Feature Selection and Ranking
- Weka Installation
- Getting Started with Weka
- Run your Classifier in Weka
- Data Mining in Weka
- Weka Tutorial