CSE514 – Fall 2021 Programming Assignment 2

This assignment is to give you hands-on experience with the comparison/selection of different classification models for a given dataset/problem. It consists of a programming assignment (with optional extensions for bonus points) and a report.

Topic

Compare, analyze, and select a classification model for predicting breast cancer.

Programming work

A) Data preprocessing

In the last programming assignment, we (correctly) assumed that there were no missing values and that the last 130 samples had a similar distribution of values as the first 900. For this dataset you must remove null values, pick 10% of the data to act as a final validation set, and divide the remaining into subsets for 10-fold cross validation.

Optional extension 1 – Imputing missing values

There are few enough missing values that you can simply drop out the effected samples. Up to 10 bonus points if you choose to impute instead, and explain the process in your report.

B) Model fitting

For this project, you must consider the following classification models:

1. k-nearest neighbors
2. Decision tree
3. Random Forest
4. SVM using the polynomial kernel
5. SVM using the RBF kernel
6. Deep neural network with sigmoid activation
7. Deep neural network with ReLU activation

For each model, choose a hyperparameter to tune using 10-fold cross-validation. You are expected to test enough values to find a performance peak, or up to 10 values if that peak cannot be found.

Optional extension 2 – Tune more hyperparameters

For bonus points, tune more than just one hyperparameter per model.
3 bonus points for each additional hyperparameter, up to 15 bonus points total.

Optional extension 3 – Consider more classification models

For bonus points, suggest additional classification models to me. If I give the go-ahead, you may include one for 5 bonus points, or two for 10 bonus points.

IMPORTANT: You may use any packages/libraries/code-bases as you like for the project, however, there will be expectations in the report for you to have control over certain aspects of the model that may be black-boxed by default. For example, a package that trains a kNN classifier and internally optimizes the k value is not ideal, since you will need the cross-validation results of testing different k values for your report.

Data to be used

We will use the Breast Cancer Wisconsin (Original) dataset in the UCI repository at

UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set

(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29)

Note that the last column of the dataset is the response variable (i.e., y).

There are 699 instances in this dataset.

Report template – follow the instructions here to earn full points

(100pts total)

- Introduction (10pts)
  - (5pts) Your description of the problem and the motivation for trying to determine the "best" DM classifier.
  - (5pts) Your description of what factors should be considered in determining a classifier as the "best," e.g. computational complexity, validation accuracy, model interpretability, etc.
- Methods (10pts)
  - (5pts) A summary of how you preprocessed the data
    - Bonus points if this involved imputing missing values
  - (5pts) A summary of the code packages/libraries that you used in your project.
- Results (65pts)
  - For each classifier:
    - Brief description of the classifier and its general advantages/disadvantages
    - Figure: Graph the cross validation results over the range of hyperparameter values you tested
      - Bonus points for additional hyperparameters tuned
    - Any additional details needed to replicate your results
  - Bonus points for additional classifiers
- Discussion (15pts)
  - (5pts) Compare the performance of the different classifiers on the final validation set with either a table or a figure
  - (5pts) Compare the run time of the different classifiers for training and for predicting on the validation set
  - (5pts) Lessons learned: What model would you choose for this problem and why? What would you do differently if you were given this same task for a new dataset? Anything else about this project that made you think?

Due date

Wednesday, December 1 (midnight, STL time). Submission to Gradescope via course Canvas.