IST707 Data Analytics

HW4: Statistical Learning, Deep Learning and A Complete Comparison

Due: 11:59pm, Sunday, Apr 26th, 2020

Problem Set

In this homework assignment, you are going to use multiple machine learning algorithms, including logistic regression, multiple artificial neural networks and deep learning models with different architectures to build disease diagnosis models using the same dataset in HW3. It is a binary classification problem to predict whether or not a patient has a certain unspecified disease. You also need to combine the findings with HW3 results to come up with a ranking of performance for various machine learning models on this dataset.

Dataset

Attributes' information about the dataset (*Disease Prediction Training.csv*):

- Age: in years
- Gender: male/female
- Height: in unit of cm
- Weight: in unit of kg
- Low Blood Pressure: lower bound of blood pressure measurement
- High Blood Pressure: higher bound of blood pressure measurement
- Cholesterol: three cholesteral levels
- Glucose: three glucose levels
- Smoke: 1/0 regarding if the patient smokes
- Alcohol: 1/0 regarding if the patient drinks alcohol
- Exercise: 1/0 regarding if the patient exercises regularly
- **Disease: The binary target variable. Does the patient have the disease?**

Analysis and Report

Organize the main body of your report using the following structure (with the section breakdown and grading rubrics):

Section 1: Data preparation (10%)

- Discuss the potential data quality issues you identify about the dataset and how you apply various data preprecessing techniques to cope with those issues and perform Exploratory Data Analysis (EDA).
- Specifically discuss the type of techniques you carry out in order to prepare the dataset for the machine learning algorithms you use in the next section.
- Whenever appropriate, enhance your EDA with the effective data visualization.

Section 2: Build, tune and evaluate various machine learning algorithms (50%)

- Apply a list of machine learning algorithms most recently covered in the course (Logistic Regression, Artificial Neural Network/Deep Learning) to the training data and construct models. Perform extensive model experiments with hyper-parameters' tuning. Elaborate your strategy on issues regarding: regularization, optimization (loss function, gradient descent strategy, epoch, batch size, etc.), hidden node and output node properties (e.g. activation functions), etc. Discuss your choice of hyper-parameters for each algorithm and produce tables summarizing the best performing models and their corresponding model specifications (i.e. the combination of hyper-parameters). Also explain your choice of model performance evauation methods and metrics in order to produce unbiased and low variance estimates.

- For ANN/DL models, please consider three architectures with 0, 1 and 2 hidden layers (ANN0, ANN1 and ANN2, respectively). For each architecture, experiment with different numbers of hidden nodes

wherever applicable. You are encouraged, but not required, to experiment with deep learning models with more than 2 layers to see if doing so could improve the model performance. (Hint: compare training accuracy with validation accuracy of models with different complexity for signs of overfitting)

- Compare the performance specifically about the following three learning algorithms: linear SVM, logistic regression, and single layer perceptron (with ZERO hidden layer). Do they have similar performance? Please provide some discussions about the theoretical comparison of these three learning algorithms.

- Wherever applicable that the above machine learning classification algorithms could predict the probability of the target class label, generate Receiver Operating Characteristics (ROC) curve and calculate Area Under Curve (AUC) metric for the identified best performing models.

- Include detailed explanation of your modeling process and interpretation of the results in your analysis writeup (with markdown language) and structure such writeup in an easy-to-follow layout. Please limit your program output only to the most relevant part which is used to support your analysis. Excessive amount of less relevant outputs (e.g. display the whole dataset) in your report will have a negative effect on the grade.

Section 3: Combination and Comparison of Multiple Machine Learning Algorithms (20%)

- Apply decision tree learning algorithm and fine tune the model on the disease dataset. Display the best performing model and analyze/interpret it to see if the model is consistent with findings by other learning algorithms performed in HW3, such as Logistic Regression (check the model specification such as coefficients and significance) and Ensemble Learning methods (Random Forest and Gradient Boosting Machine) and the relative importance of features derived from those methods.

- Generate a model performance master table that includes all the algorithms you tried in both HW3 and HW4 (i.e. all the supervised learning algorithms we cover in this course). Columns of the table should include: name of algorithm, key hyperparameters tuned, model performance with the metrics of your choice, ball-park estimate of time needed to run the model. The table should be sorted in the decreasing order of the model performance for the primary metric. Please make sure you use the consistent model performance evaluation method (e.g. 10 fold cross validation or hold out method) and justify your choice of the primary performance metric.

Section 4: Prediction and interpretation (20%)

- After building the classification models, apply them to the test dataset (*Disease Prediction Testing.csv*) provided to predict if each person in the testing dataset has the disease.
- Please submit your prediction results as a CSV file with SIX columns (ID, DT, LR, ANN0, ANN1, ANN2) for the classfication results out of the pre-specified machine learning algorithms respectively.

Guideline

- Report layout: as a rough guide, include at the very least the report title (with the author name and date information), an executive summary, introduction, the main body of your report (with analysis, results and interpretation for each required machine learning algorithm), and a conclusion section that provide a high level summary of your findings and any lessons learned.
- Report writing: use markdown language extensively to explain the purpose of the code chunks and summarize your data mining process, interpret the results and draw noteworthy insights and conclusions. The instructor will primarily grade based on your report and only occasionally refer to the codes/outputs for clarifications. Therefore do not just provide codes with outputs without explanation/interpretation and expect the instructor to figure out what you try to achieve.
- Reproducibility of the analysis results: set the random seeds whenever using any functions that require random number generation.
- Adopt the best coding practices as much as possible: for example, try to avoid hardcoding any values in your program; modularize the codes and functionalize the repetitive code snippets whenever possible; comment your codes wherever helpful, etc.

- Keep your analysis and report concise and relevant: filter your analyses and outputs to include only a small sample to make your points; choose to visualize only those features that are impactful to your analysis rather than a laundry list of all visualization techniques and all the features; etc.
- **MOST IMPORTANTLY: all the submitted work should be performed independently!**

Submission

Your submission package should include the following THREE files:

- A knitted report (exported in either PDF, html or word format)
- A rmarkdown or jupyter notebook document
- A prediction CSV file (please make sure your file has all the required headers and the exact number as the provided testing dataset)

Please submit your package as a ZIP file.