CSCI6380 Data Mining: Assignment #4

Adnan Kivanc Corut

University of Georgia — October 29, 2020

Model Development

1 Introduction

In this assignment, I try to develop 3 distinct models for each of the two datasets which includes Wisconsin breast cancer data set and a dataset with NAM weather forecasts with observed solar radiation. The target variable for the breast cancer dataset is diagnosis which makes it as a classification problem. On the other hand, the target variable for the second dataset is the solar radiation 3 hours ahead of the reference time which makes it a regression problem. For this reason, for the first dataset, I will be developing three binary classifiers by using scikit-learn. For the second dataset, I will be developing 3 regressors by again using scikit-learn. I will be testing different parameters, and perform features selection to improve the accuracy of each model. Below I first explain my approach such as how the data were preprocessed, partitioned, which parameters were used and how they were selected. Then, I will present the accuracy results of different models and also present hypothesis testing results.

2 Approach

2.1 Data Processing

First, datasets were imported using `read_csv` from pandas. All rows of both datasets was checked against NA values but none was detected. Last colums of each datasets ('diagnosis' and 'SOLARRADIATION_0003' respectively) were treated as target variables. Both datasets was split into a 75/25 train/test splits.

2.2 Parameter Tunning

Three binary classifiers and three regressors were chosen based on previous familiarity with the models (Table 1). In order to evaluate the effects of hyperparameters, which are parameters that are not directly learnt within estimators, in each classifier and to assess their impact on the models, grid search was performed using `sklearn.model_selection.GridSearchCV`. GridSearchCV comprehensively assesses all parameter combinations from a given parameter space.

All hyperparameters that were tested in this study for each estimators can be found in the supplementary jupyter notebook. I will be presenting here only the best parameter settings that were identified as a result of these individual grid searches. Alongside, parameter tuning I also performed scaling and dimensionality reduction using PCA. Scaling method was chosen for each model independently as part of grid search. Three method was evaluated (`StandardScaler()`, `RobustScaler()`, `MinMaxScaler()`) in grid search. Number of commponents to keep (`n_components`) in PCA for each method was also chosen based on grid search. The best parameters for each model was identified as a result of grid search. Table 2

Binary Classifiers	**Regressors**
Gaussian Naive Bayes	Linear Regression
Decision Tree Classifier	Decision Tree Regressor
Random Forest Classifier	Random Forest Regressor

Table 1: Models that were chosen for this assignment.

Binary Classifiers	**Best Parameters**
Gaussian Naive Bayes	*{'reduce_dim': PCA(n_components=3), 'reduce_dim__n_components': 3, 'scaler': RobustScaler()}*
Decision Tree Classifier	*{'classifier__criterion': 'gini', 'classifier__max_depth': 12, 'reduce_dim': PCA(n_components=5), 'reduce_dim__n_components': 5, 'scaler': MinMaxScaler()}*
Random Forest Classifier	*{'classifier__bootstrap': True, 'classifier__max_depth': 25, 'classifier__max_features': 5, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 10, 'reduce_dim': PCA(n_components=10), 'reduce_dim__n_components': 10, 'scaler': MinMaxScaler()}*
Regressors	**Best Parameters**
Linear Regression	*{'reduce_dim': PCA(n_components=30), 'reduce_dim__n_components': 30, 'scaler': StandardScaler()}*
Decision Tree Regressor	*{'reduce_dim': PCA(n_components=6), 'reduce_dim__n_components': 6, 'regressor__criterion': 'mse', 'regressor__max_depth': 10, 'regressor__min_samples_leaf': 20, 'scaler': StandardScaler()}*
Random Forest Regressor	*{'reduce_dim': PCA(n_components=21), 'reduce_dim__n_components': 21, 'regressor__max_depth': 50, 'regressor__max_features': 10, 'regressor__min_samples_split': 5, 'regressor__n_estimators': 100, 'scaler': MinMaxScaler()}*

Table 2: Best parameters for each classifier identified by GridSearchCV.

shows best parameters for each model. Using this parameters in each model, 10 fold cross validation was performed to compare the models. Binary classifiers were used on Breast Cancer data, while regressors were used on solar data. After 10 fold cross validation, top 2 performed model (based on accuracy) was further evaluated by using `paired_ttest_5x2cv` function of `mlxtend` library to see if these two models were significantly different form each other.

3 Results

3.1 Breast Cancer Data

Below results show mean accuracy of each binary classifier based on 10 fold cross validation results (Table 3, Figure 1). Based on accuracy results, random forest classifier performed best however t-test results showed that random forest did not significantly performed better than second best classifier decision tree. T-test results can be seen below:

```
Random Forest vs. Decision Tree
P-value: 0.748, t-Statistic: 0.340
Algorithms probably have the same performance
```

Binary Classifiers	Mean Accuracy (Standard Deviation)
Gaussian Naive Bayes	*0.924499 (0.035927)*
Decision Tree Classifier	*0.947306 (0.035924)*
Random Forest Classifier	*0.949060 (0.024078)*

Table 3: Results for binary classifiers performed on Breast Cancer data.

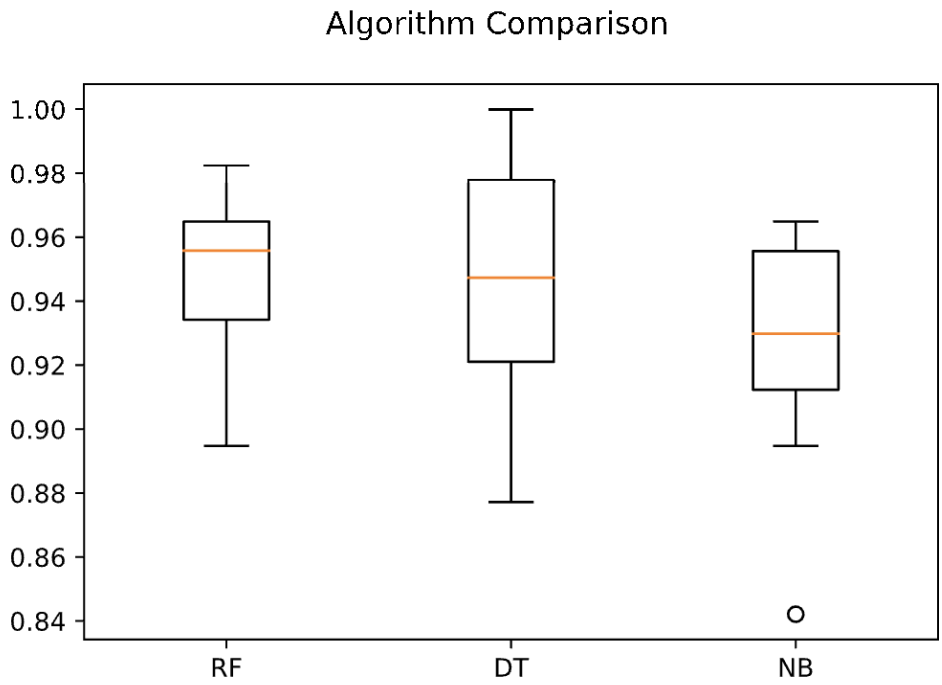

Figure 1: Comparison of binary classifiers performed on Breast Cancer data.

3.2 Solar Data

Below results show r2 scores of each regressor based on 10 fold cross validation results (Table 4, Figure 2). Based on r2 scores, linear regression performed best however t-test results showed that linear regression did not significantly performed better than second best regressor random forest. T-test results can be seen below:

```
Random Forest vs. Linear Regression
P-value: 0.211, t-Statistic: 1.433
Algorithms probably have the same performance
```

On the other hand, t-test results showed that linear regression performed significantly better than decision tree regressor. T-test results can be seen below:

```
Decision Tree vs. Linear Regression
P-value: 0.014, t-Statistic: -3.719
Difference between mean performance is probably real
```

Regressors	R2 (Standard Deviation)
Linear Regression	*0.862257 (0.037557)*
Decision Tree Regressor	*0.792085 (0.074192)*
Random Forest Regressor	*0.852844 (0.041029)*

Table 4: Results for regressors performed on Solar data.

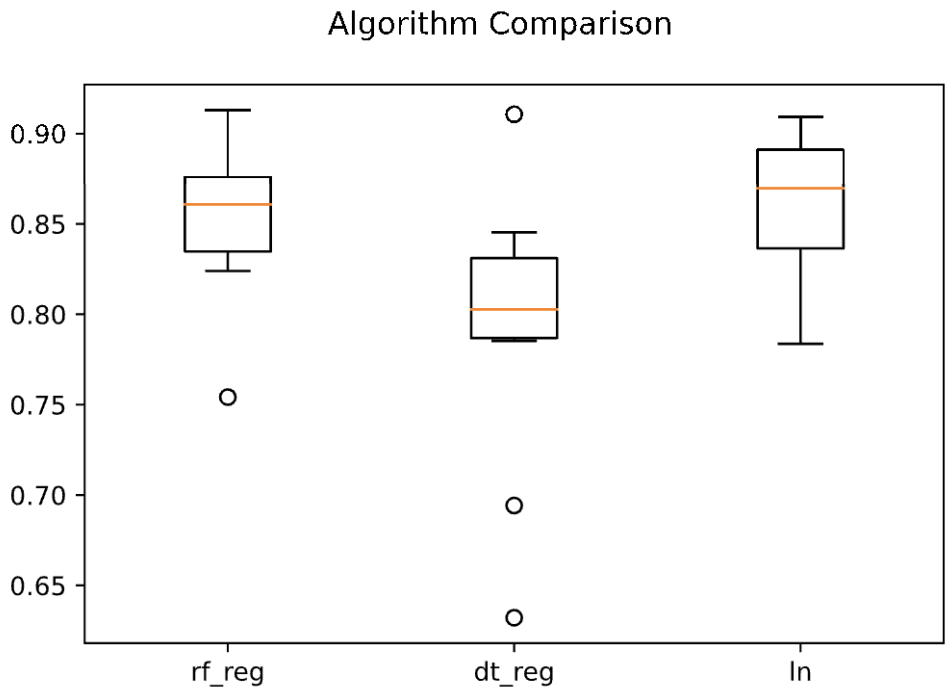

Figure 2: Comparison of regressors performed on Solar data.

4 Discussion

Performances of the models used for each datasets can be further improved by possibly using different feature selection methods, such as SelectKBest, RFE or feature_importances_ instead of dimensionality reduction method used in this study. It can be also speculated that performances could be further improved by using more effective grid search with the most important parameters and more fine tuning. Another way to investigate performence improvement would be testing different classifiers and regressors on each dataset.