Machine Learning: Project 1

Kai-yu Chen (GTID: 903233101)

Abstract

Objectives: This study aims to explore two classification questions in two dataset from UCI Machine Learning Repository: Breast Cancer Diagnosis (BCD) and White Wine Quality Assessment (WWQA). Machine learning (ML) methods were conducted to predict the breast cancer types (benign vs. malignant) of cell nuclei morphology and the quality of white wine of added manufacturing elements.

Methods: Learning Vector Quantization method was employed to select best features for outcome prediction; the selected attributes were worst parameter, worst radius, worst area, worst concave points, and mean concave points in the BCD dataset, whereas alcohol, free sulfur dioxide, citric acid, pH, total sulfur dioxide, and residual sugar attributes in the WWQA dataset. Prior to classifier training, BCD dataset was normalized via Min-Max scaling method, and the WWQA via z-normalization. Five algorithms were then introduced to build a learner and predict classification results: decision tree (DT), neural networks (ANN), boosting, support vector machine (SVM), and k-nearest neighbors (kNN). Each model was validated via 10-folds cross validation method. The error rate and training time were further measured and compared between models. Lastly, as a brief exploratory analysis, additional insight on how each algorithms works to build a learner according to classification questions and features will be included.

Results: In BCD dataset, the calculated accuracy for DT, ANN, boosting, SVM and kNN was 93%, 92.9%, 92.9%, 96.7% and 95% respectively. In the WWQA dataset, boosting algorithm outperformed the DT, ANN, SVM and kNN with accuracy 62%, 54.3%, 57%, 55% and 53.6% respectively

Discussion and Conclusions: The optimal Breast cancer classifier in all algorithms predicted outcomes with accuracy above 90%; DT algorithm is favorable because it is a more efficient way as compared to other algorithms while giving the similar accuracy. The WWQA classifier reached to the highest 62% predicting accuracy by applying boosting algorithms, indicating the robustness of boosting and its good ability.

Introduction

Breast Cancer Diagnosis

Cancer is among the leading causes of mortality and morbidity worldwide. One of the limitations in developing an effective cancer treatment results from tumor heterogeneity. Although there are increasingly innovative medicine been developed, the efficacy and safety in clinical application are uncertain. Successful classification between benign and malignant tumor type enables physicians to provide patients with better curative strategies, and thus early diagnosis of a cancer type are important in clinical treatment and cancer research. Because of the tumor heterogeneity and significance of personalized medicine, machine learning (ML) techniques have been introduced to detect key features from complex datasets and model the tumor progression and treatment. Here, ML tools are introduced to show its power in classifying tumor types based on the Breast Cancer Diagnosis (BCD) dataset from UCI Machine Learning Repository. The dataset is composed of 569 instances with 30 features computed from a digital image of a fine needle aspirate (FNA) of a breast mass. The characterizations of cell nuclei are described based on the image.

White Wine Quality Evaluation

Over the last decade, wine has been increasingly enjoyed by consumers worldwide. To boost its growth further, new technologies are explored for wine brewing and wine marketing by the wine industry. Wine certification and quality assessment play important roles in this context. Thus, wine quality assessment becomes crucial in the wine industry. Quality evaluation is part of the certification process and can be utilized to stratify wine brands. Accessing the effects of the physicochemical tests in the white wine quality is useful for improving the production process and also allowing manufacturers to have an insight in the white wine market. Such understanding is valuable not only for wine certification but also wine manufacturers and consumers. In this study, white wine quality classification will be focused. Models to predict white wine quality will be constructed by five different learners via taste preferences based on 11 manufacturing elements that are available to measure at the wine certification process based on the White Wine Quality Assessment (WWQA). Taste preference scores ranged from 1-10 based on subject's satisfaction

Methods

Feature Selection: Feature selection has been done in two datasets via Learning Vector Quantization (LVQ) method. The BCD dataset consists of five key features: worst parameter, worst radius, worst area, worst concave points, mean concave points; the WWQA dataset consists of alcohol, free sulfur dioxide, citric acid, pH, total sulfur dioxide, and residual sugar attributes.

Data Normalization: BCD dataset was normalized via Min-Max scaling method, and WWQA dataset was normalized via z-normalization. Min-Max scaling method was only used in white wine neural network learner to avoid prediction error after rounding the prediction results. 10-folds cross validation was applied and repeated 10 times in both dataset for cross validation, and were randomly split into training dataset and testing dataset with a default ratio of 70:30.

Model Performance Evaluation: 10 times 10-folds cross validation was utilized to measure model performance due to its stability in error estimation and popularity in litertures. In addition to accuracy, kappa value was used to evaluate agreement between the models' predictions and the actual values.

Decision Tree (DT): DT classifier was constructed by package *caret* and package *rpart* in R-Studio. Classification and Regression Trees (CART) algorithm was used to build tree model. The optimal complex parameters (cp) were then selected by caret package to achieve minimal classification error rate. 10-folds cross validation was done 10 times to assess model prediction. To improve model performance, pruning tree had been done to reduce tree size and classification error by tuning parameters in cp value (i.e. 0 – 0.002), maximum tree depth (i.e. 5- 100), minimal data in a node (i.e.1-20), and minimal split per node (i.e. 2-20). Accuracy and kappa values were then achieved to evaluate model performance via comparing predictive model with test dataset.

Neural Network (ANN): Backpropagation method was used to build a neural network classifier. Training model was built based on default learning rate (i.e. 0.01 – 0.5), threshold (i.e. activation function, 0.01-0.5), maximum iterations (500000), number of hidden layers (0-3), and number of nodes in hidden layers (0-4). Then, 10-folds cross validation was done 10 times to assess training model prediction. Predictive model performance was finally evaluated with test dataset via correlation coefficient/accuracy.

Boosting: Tree learner was firstly created by C5.0 algorithm via package *C50*. The performance of the learner was validated by test dataset. Adaptive boosting was utilized to train and combine tree learners by different boosting iterations (i.e. 10 - 60). The boosting model performance was measured by accuracy.

Support Vector Machine (SVM): Linear kernel, sigmoid kernel, and Gaussian RBF kernel function were used to train a support vector machine learner by package *kernlab*. The performance of the model was then evaluated via accuracy by applying predictive model onto test dataset.

k-Nearest Neighbors (kNN): Package *caret* was used to build k-nearest neighbor leaner. The optimal k value was achieved by systemically selecting largest training model accuracy. The model was evaluated by 10-folds cross validation 10 times. Different k value was then implemented to derive optimal model candidates, ranging from 5 – 60. The optimal model was selected based on largest accuracy value and kappa value.

Results

Decision Tree (DT). In the BCD tree classifier, the optimal tree learner could be built from the cp value ranged approximately 0 - 0.001, with maximum tree depth ranged between 5-100, number of data per node ranged between 1-20, and minimum split per node ranged from 1-20. The optimal model accuracy reached to 93%, with kappa value of 0.85. The result pointed out the good agreement between model prediction and actual values; whereas in the WWQA tree classifier, the optimal tree model could be constructed from the cp value approximated 0.002, with maximum tree depth ranged between 5-10, number of data per node and minimum split per node of 3. The accuracy in the optimal tree model could be reached to 54.3%, with the kappa value of 0.28, indicating that there is fair agreement between model prediction and actual values.

Neural Network (ANN). In the breast cancer diagnosis neural network classifier, the optimal model can be achieved with 4 nodes in the first hidden layer and 1 node in the second hidden layer. The training

iterations were 251550, with learning rate of 0.01 and activation threshold in 0.01. The error was 4.97, with accuracy of 92.9%. (Table 2). The result also showed the trend of decreasing in training error along with the increasing in training interactions and decreasing in learning rate. In the white wine quality classifier, the lower training error rate could be achieved when three hidden layers were applied (with learning rate in 0.1 and 0.5 respectively). Lower learning rate is accompanied with high training iterations to achieve similar accuracy (0.57 and 0.556).

Boosting. To classify breast cancer types, the decision tree model was first built via C5.0 algorithm. The model falsely predicted 13 of the 399 training instances for an error rate of only 3.3%. The predictive accuracy was 92.9%. An adaptive boosting tree classifier was built by combinations of these trees (weak learner) with boosting iterations from 10, 20, and 30 to reduce error rates and to improve the accuracy of the classifier (learning time: 0.0 sec). The training error rate decreased from 0.5% to 0.0%. The testing error rate stayed around 7% (7%, 7.6%, 7%). The accuracy of the final model stayed around 92.9%. Again, adaptive boosting algorithm was applied to predict white wine quality. C5.0 algorithm was used to construct a decision tree model. The model correctly classified all but 454 of the 3431 training instances for an error rate of 13.2%. The accuracy of the model was 56.7%. To lower model prediction error rate and strengthen model's prediction robustness, these tree models were combined to form a boosting tree from 10 to 60 iterations (learning time: 0.6 sec - 2.7 sec). The training error rate decreased from 0.3 % to 0.0 %, while the boosted model accuracy slightly increased from 60.9% to 62% after validating with testing dataset.

Support Vector machine (SVM). Two kernel functions were applied to breast cancer dataset: linear kernel function and Gaussian RBF kernel function. With linear kernel function, the SVM showed 0.052 in training error and 95.6% in accuracy. Whereas with using Gaussian RBF kernel function, the SVM model showed 0.052 in training error and 96.7% in accuracy. The performance of SVM models could be improved via tuning cost and sigma values, yet in this section the optimal sigma value was tuned by the package along with the cost value. The optimal linear kernel function model could be built with cost value in 1 and sigma value in 1.31. The model accuracy could be reached to 93%. (Table 3) In addition, the optimal Gaussian RBF kernel function model was built when cost value equals approximately to 1 and sigma values approximated to 2. In this model, the accuracy could be reached to around 92%. On the other hand, three kernel functions were applied to white wine quality dataset to expand model candidates and chose the best model: linear kernel function, Gaussian RBF kernel function, and sigmoid kernel function. All of three functions showed similar results, with around 0.43 in training error and 54% in accuracy. Cost value was adjusted to derive optimal model, ranging from 0.01 – 10. The table showed that the largest accuracy could be achieved with the cost value equals to 10 in all three models.

k-Nearest Neighbors (kNN). In the breast cancer classifier, largest accuracy value was a baseline to select the optimal training model. Overall, the accuracy and kappa value of the model were both very high in this model (Table 4), among all, highest accuracy can be achieved at 20 or 40 near neighbors. All kappa values ranged from 0.84 – 0.91, indicating good agreement between the classifier and the true values. In the white wine quality classifier, different k values were plugged into classifier to derive a better model accuracy. Among all, the final k value, 55, was achieved via 10 times 10-fold cross validation. All the kappa values ranged between 0.24 and 0.26, indicating fair agreement between the model's predictions and the true values.

Discussions and Conclusion

Based on the results, the optimal breast cancer classifier in all algorithms can predict outcomes with above 90 % accuracy, whereas the optimal white wine quality classifier reached to highest accuracy of 62% via adaptive boosting algorithm. This observation may explain the intrinsic property of each dataset.

Decision Tree. By Using DT algorithm, all breast cancer classifier could satisfy above 90% accuracy. Large tree was grown first followed by pruning to avoid the possibility of ill performing in stop criteria. Among observations from the results, reducing tree size did not significantly affect the accuracy, yet trees with larger size equip risk in overfitting[1]. Thus, the smaller tree would be selected as a better model. In the white wine quality decision tree classifier, the maximum model accuracy could reach to 54%. Surprisingly, the lowest cp value (cp=0) in this model did not lead to largest model accuracy. This could be explained by the fact that the lower cp value might result in overfitting of the data. Cross validation could be employed to eliminate this bias.

Neural Network. By applying ANN algorithm, both classifiers showed several interesting trends when they were learning from the data. The training error would reduce with the increased iterations. This can be explained by the property of backpropagation in neural network that the neural system optimizes the weights of each node from the previous layers according to cost function in each iteration step[2,3]. The observed higher learning rate with higher errors may because that with a high learning rate, the system contains too much kinetic energy and the parameter vector bounces around chaotically, unable to settle down into deeper, but narrower parts of the loss function. In general, it is better to use fewest hidden nodes that result in adequate performance. Although more complex network connections allow the learning of more complex problem, this runs a risk of overfitting. Future work will be emphasized on the tradeoff between numbers of hidden layer nodes, threshold function, and learning rate.

Boosting. The accuracy of the breast cancer boosting classifier did not boost after combing all the weak tree learners, indicating that the boosting algorithm do not need to be applied into BCD dataset to build a strong classifier. Given the C5.0 tree learner alone already showed high model accuracy, there was not much space to improve its performance. On the contrary, the accuracy of white wine quality was boosted after applying adaptive boosting algorithm. The accuracy even increased with increased boosting iterations. The white wine quality classifier reached to the highest predicting accuracy (accuracy=62%) by applying this algorithm compared to other four algorithms, indicating the robustness of boosting and its good ability in handling high dimensional spaces as well as large number of training examples[4,5].

Support Vector Machine. The SVM algorithm was applied to both classification questions starting from smaller cost value to larger one. The result showed that the optimal classifier could be achieved when larger cost value was applied. The cost value is the parameter for the soft margin cost function and controls the effects of each support vector as well as involves trading error penalty for stability. The finding could be annotated by the fact that the support vector machine algorithm has low bias and high variance [6,7]. Yet the trade-off can be tuned by increasing the cost value that influences the number

of violations of the margin allowed in the training data, which thus increases the bias but decreases the variance. There was no reliable rule for matching a kernel to a particular learning task in this study. More combinations of kernels and model parameters (C, gamma) can be tested to obtain a better model.

k-Nearest Neighbors. Generally, the optimal k value is set equals to the square root of the number of training instances, yet this practice is hard to see from the result of this study. There is no significant difference between the accuracy by using wide range of k values in both datasets. The high accuracy and kappa value of the BCD kNN classifier may be explained by the intrinsic property of the dataset. The dataset was well-classified with less noise and clear distinction among the hyperparameters. In addition, the data size smaller and suitable to be trained by kNN algorithm. Yet, the poor performance of the white wine quality classifier could be explained by the fact that classification types (sensory score rank) in the dataset are similar and tend to be fairly homogeneous. The algorithm is not well-suited for identifying the boundary.

Among all the algorithms, I will apply DT algorithm to train Breast cancer classifier because it returned the same result (above 90% accuracy) as compared to other four algorithms but employed with a more efficient way. DT only uses the most important features and the model can be easily interpreted to public without complicated mathematically background. In addition, the highly automatic learning process can handle numeric/nominal features, or missing data. As for training a WWQA classifier in this study, I will choose adaptive boosting algorithm for the data training since it retuned the highest accuracy in this dataset. Moreover, the flexibility in applying into different machine learning algorithm and the entity in combining weak leaners make it a powerful classifier. Neural network is so far the most accurate modeling approaches, yet it may costs a large amount of time in data training when it handling a complex network.

Besides the property of each algorithm, the property of the dataset will influences the prediction outcome as well. While BCD dataset consists only 569 instances and 6 key features, the features are all linearly separable. Its property makes it easier to be analyzed and interpreted by majority of the algorithms. Yet its accuracy could be improved to 97.5% by choosing different key features based on the lecture studies. Although the six features were selected by LVQ methods, the key features would be only worst area, worst smoothness, and mean texture. This could be explained from the intrinsic biological perspectives that comparing to benign tumors, malignant tumor tissues are more loosen. The shape of the cell nuclei is irregular and coarse[8,9] . On the other hand, WWQA consists 4898 instances, with 7 features and are not linearly separable based on prior knowledge. Moreover, the classification category was done based on individual sensory taste, which could be ranged widely according to personal preference and hard to distinguish between neighboring categories as well. For example, according to the LVQ result, there was nearly no difference in feature weights between the criteria of being “Very Good” and “Fantastic”. And it may be the reason why the model accuracy cannot be improved even after various parameters tuning process. It is challenged to distinguish a clear boundary between dependent variables and classify data into clear groups in this dataset. Therefore, in addition to interpreting result solely from modeling result, it is also important to collect and understand background information according to the study questions.

Table 1. Decision Tree Classifier

Breast Cancer Decision Tree Classifier						
cp	Max Depth	Min Split	Min Bucket	Method	Accuracy	Kappa
0	100	20	20	rpart	0.93	0.85
0	10	2	1	raprt	0.93	0.85
0.001	100	20	20	rpart	0.93	0.85
0.001	**5**	**2**	**1**	**rpart**	**0.93**	**0.84**
0.001	10	2	1	rpart	0.93	0.84
0.22	30	3	3	rpart	0.89	0.76
0.22	5	3	3	rpart	0.89	0.76
Wine Quality Decision Tree Classifier						
cp	Max Depth	Min Split	Min Bucket	Method	Accuracy	Kappa
0	5	3	3	rpart	0.531	0.288
0	10	3	3	rpart	0.531	0.288
0	30	2	2	rpart	0.531	0.288
0.001	20	5	5	rpart	0.537	0.28
0.001	5	3	3	rpart	0.537	0.28
0.002	10	3	3	rpart	0.54	0.28
0.002	5	3	3	rpart	0.543	0.28

Table2. Neural Network Classifier

Breast Cancer Diagnosis Neural Network Classifier							
Number Of Node			Learning Rate	Threshold	Iterations	Accuracy	Errors
Hidden Layer							
First	Second	Third					
4	**1**	**0**	**0.01**	**0.01**	**251550**	**0.929**	**4.97**
2	1	0	0.01	0.01	52	0.637	93.5
3	1	0	0.01	0.01	82364	0.91	6.89
1	0	0	0.01	0.01	10489	91.8	11.1
1	1	0	0.1	0.1	801	0.918	12.14
2	1	0	0.1	0.1	695	0.918	11.4
1	1	3	0.1	0.1	915	0.912	11.3
1	1	1	0.5	0.5	22	0.637	93.4
1	2	1	0.5	0.5	331	0.035	13.52
White Wine Quality Neural Network Classifier							
NumberOfNode			LearningRate	Threshold	Iterations	Accuracy	Errors
HiddenLayer							
First	Second	Thrid					
4	**2**	**0**	**0.5**	**0.5**	**2066**	**0.553**	**910**
1	1	0	0.5	0.5	411	0.47	1034.55
4	2	1	0.5	0.5	3710	0.556	897.64
4	3	2	0.1	0.1	138249	0.57	843.75
1	0	0	0.1	0.1	13288	0.47	1033.5
4	2	0	0.01	0.1	13288	0.47	1033.5
1	1	0	0.01	0.01	87732	0.47	1029.99

Table 3. SVM Classifier

Breast Cancer Diagnosis SVM classifier			
Kernel	Cost	Sigma	Accuracy
Linear	0.5	1.29	0.92
Linear	**1**	**1.31**	**0.93**
Linear	2	1.23	0.92
Linear	0.001	1.18	0.57
RBF	0.001	1.07	0.52
RBF	0.1	0.77	0.54
RBF	1	2.07	0.92
White Wine Quality SVM classifier			
Kernel	Cost	Sigma	Accuracy
Linear	0.001	0.17	0.45
Linear	0.01	0.17	0.45
Linear	1	0.18	0.53
Linear	**10**	**0.18**	**0.54**
RBF	0.01	0.16	0.45
RBF	0.1	0.16	0.5
RBF	1	0.18	0.53
RBF	**10**	**0.17**	**0.55**
Sigmoid	0.01	0.17	0.45
Sigmoid	0.1	0.17	0.5
Sigmoid	1	0.16	0.53
Sigmoid	**10**	**0.16**	**0.55**

Table 4. kNN Classifier

Breast Cancer kNN Classifier			
k	Tune Length	Accuracy	Kappa
5	20	0.936	0.859
10	20	0.929	0.846
20	**20**	**0.95**	**0.89**
30	20	0.94	0.86
40	**20**	**0.95**	**0.91**
White Wine Quality KNN Classifier			
k	Tune Length	Accuracy	Kappa
10	20	0.51	0.25
15	20	0.516	0.254
20	20	0.52	0.25
25	20	0.527	0.242
30	20	0.536	0.26
40	**20**	**0.534**	**0.26**
60	20	0.525	0.24

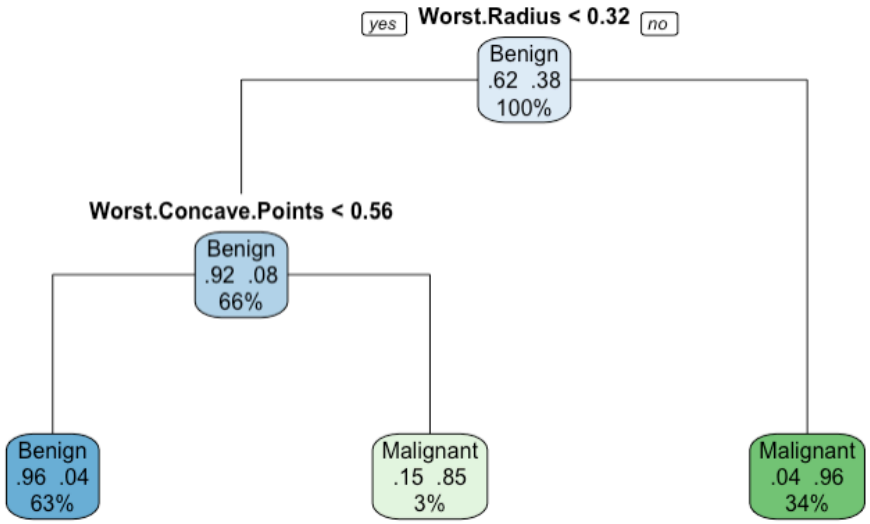

Figure 1. Optimal Breast Cancer Diagnosis Decision Tree Classifier

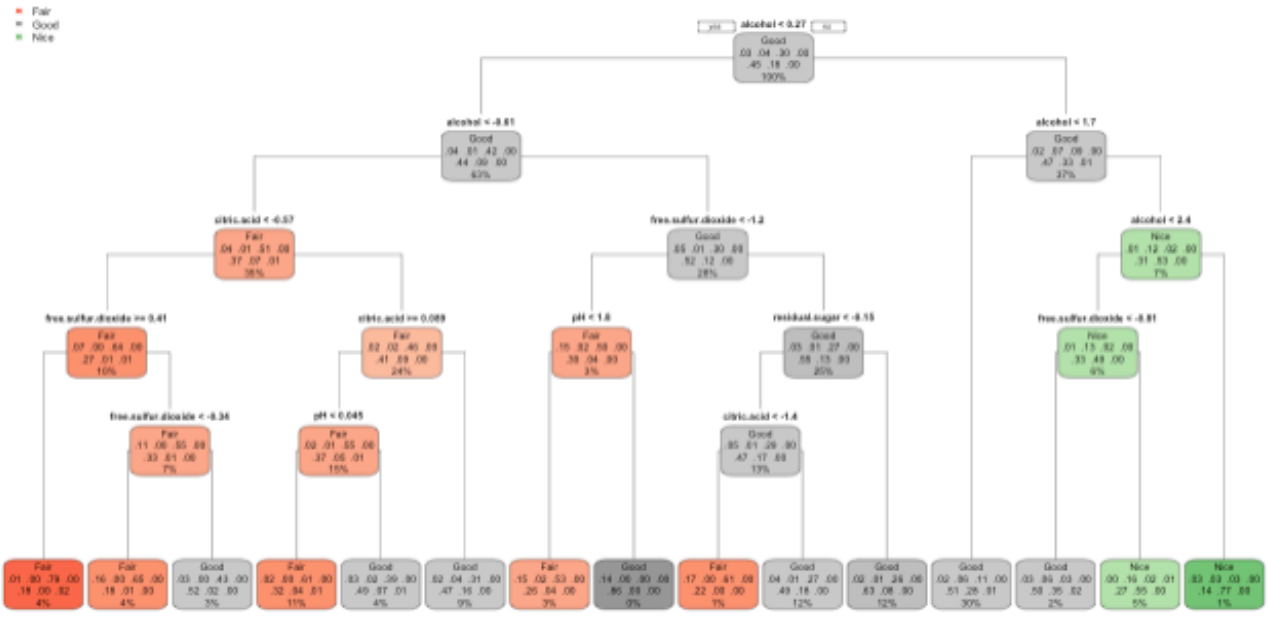

Figure 2. Optimal White Wine Quality Decision Tree Classifier

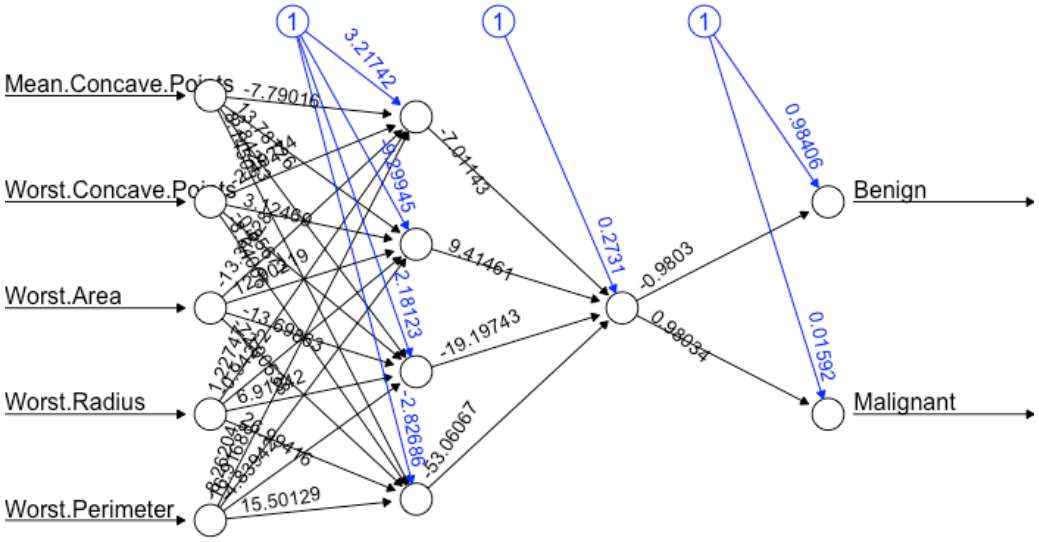

Figure 3. Optimal Breast Cancer Neural Network Classifier

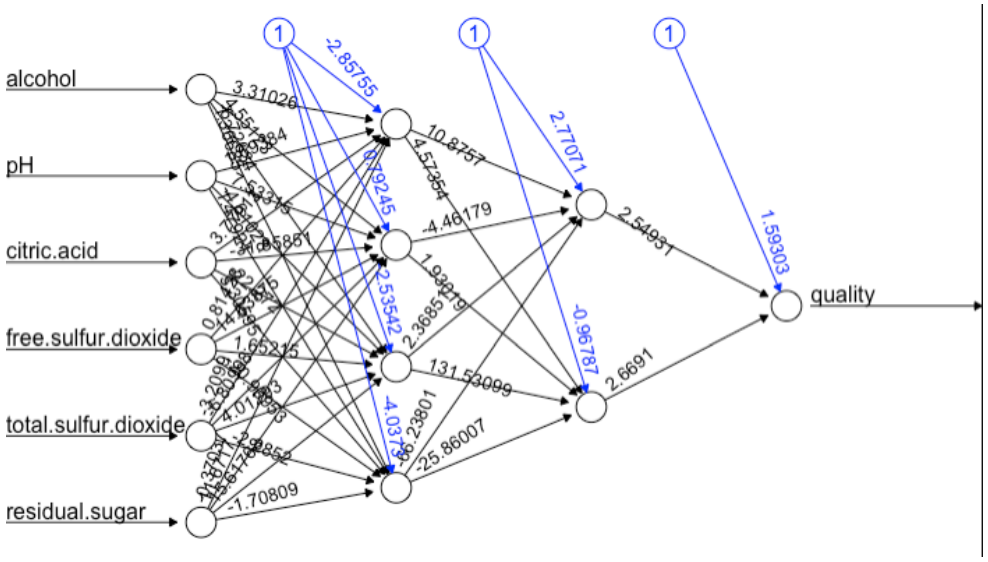

Figure 4. Optimal White Wine Quality Neural Network Classifier

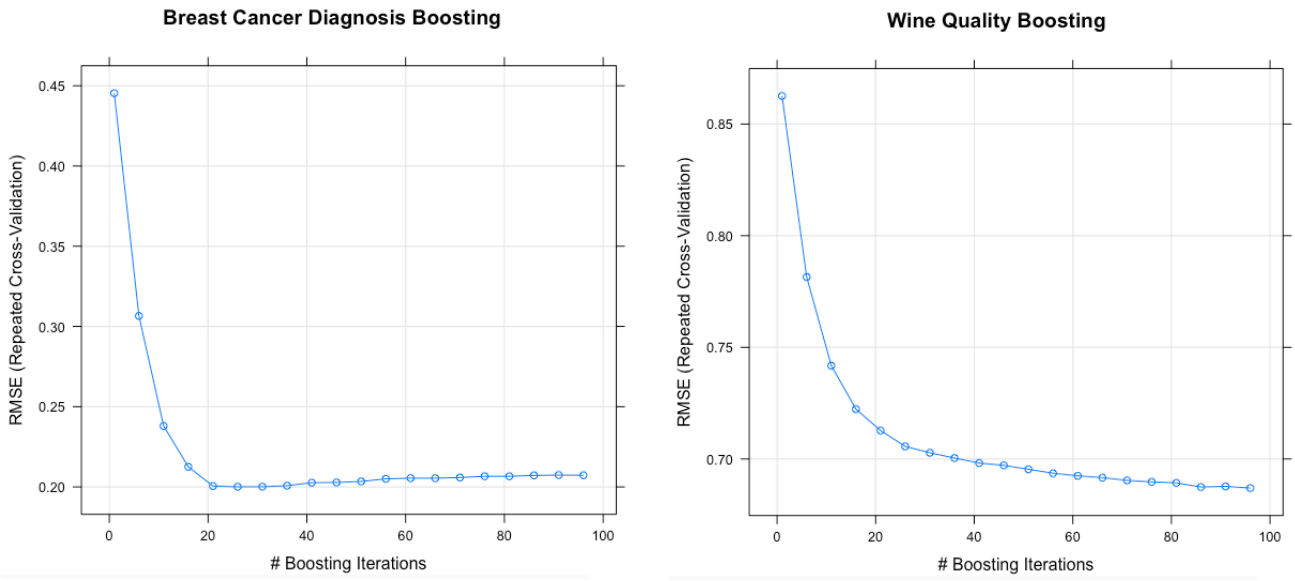

Figure 5. Boosting Iterations.

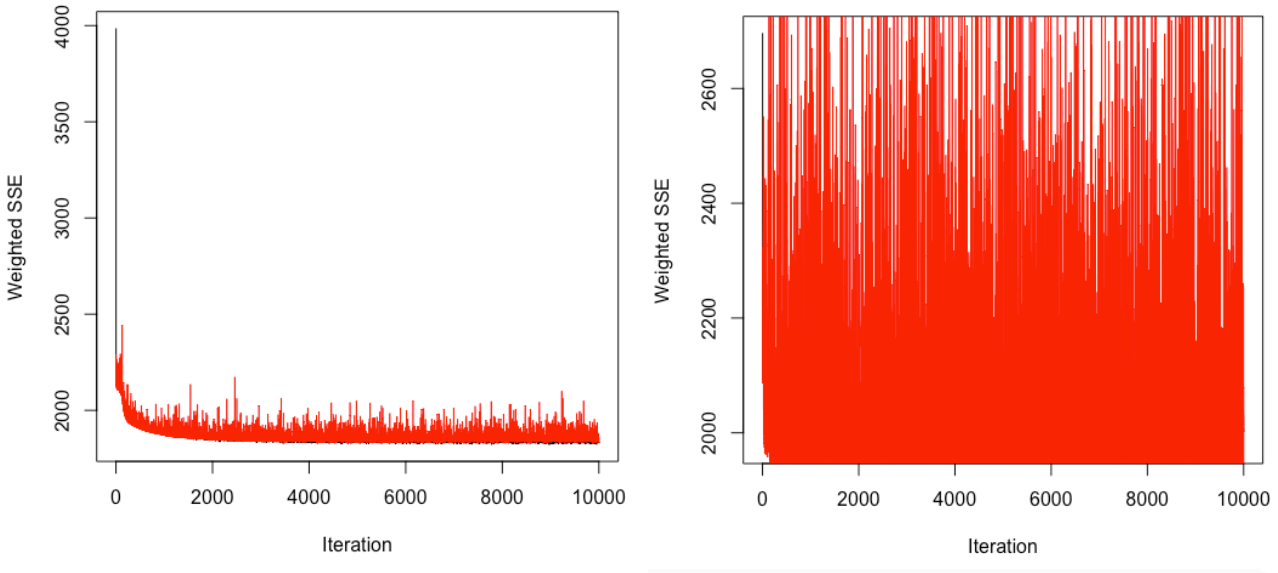

Figure 6. Iteration vs. Weighted SSE in Neural Network

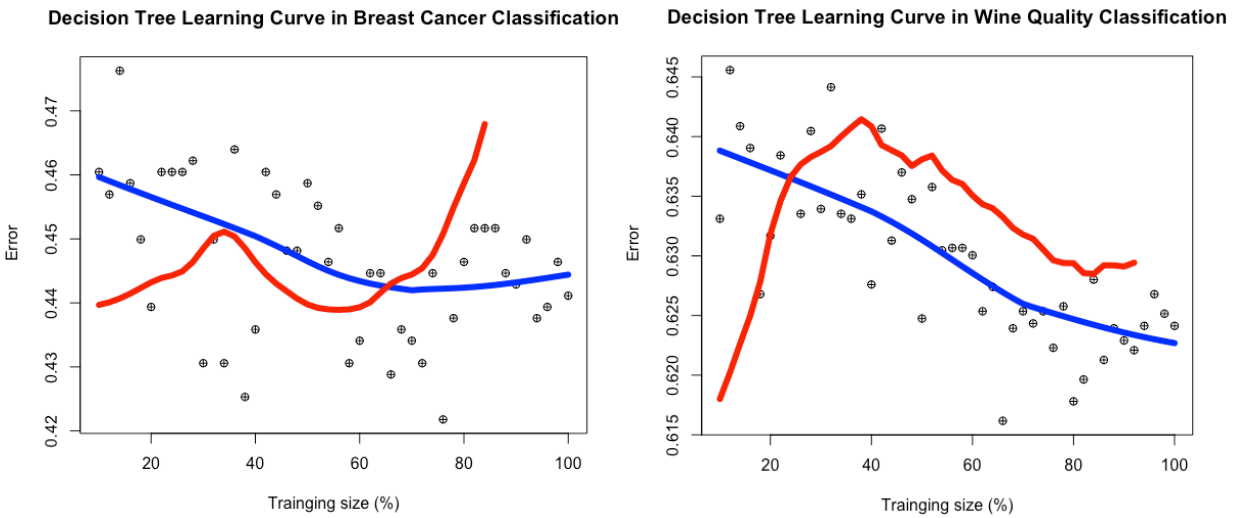

Figure 7. Decision Tree learning Curve.

Blue curve represents training error curve, while red curve represents testing error curve.

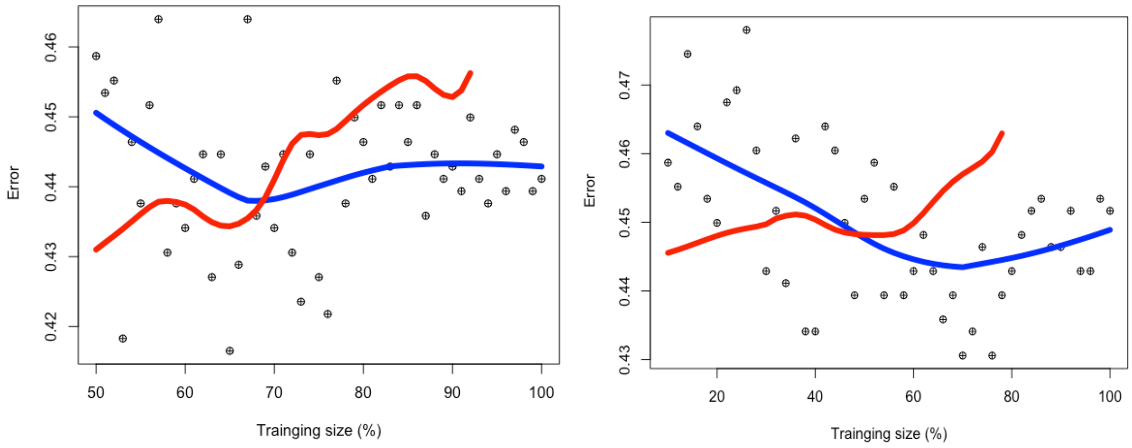

Figure 8. kNN learning Curve.

Blue curve represents training error curve, while red curve represents testing error curve.

Breast cancer diagnosis classifier is shown on the left, while white wine quality classifier is shown on the right.

Reference.

1. Song, Yan-yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry* 27.2 (2015): 130.

2. Leung, Henry, and Simon Haykin. "The complex backpropagation algorithm." *IEEE Transactions on Signal Processing* 39.9 (1991): 2101-2104.

3. Schiffmann, W., M. Joost, and R. Werner. "Optimization of the backpropagation algorithm for training multilayer perceptrons." *University of Koblenz: Institute of Physics* (1994).

4. Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." *icml*. Vol. 96. 1996.

5. Freund, Yoav. "Boosting a weak learning algorithm by majority." *COLT*. Vol. 90. 1990.

6. Grandvalet, Yves, and Stéphane Canu. "Adaptive scaling for feature selection in SVMs." *NIPS*. 2002.

7. Cao, Peng, Dazhe Zhao, and Osmar Zaiane. "An optimized cost-sensitive SVM for imbalanced data learning." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2013.

8. Abercrombie, M., and E. J. Ambrose. "The surface properties of cancer cells: a review." *Cancer Research* 22.5 Part 1 (1962): 525-548.

9. Albini, A., et al. "A rapid in vitro assay for quantitating the invasive potential of tumor cells." *Cancer research* 47.12 (1987): 3239-3245.