Use decision tree and ensemble learning to predict the number of abalone rings

Jock Li z5240840

Abstract

This report majors in creating a decision tree model and ensemble learning in abalone classification problem. The goal is to estimate the number of rings of abalone by using features that are convenient to measure such as length and weight. This article will use the CART method to create model. Use the pruning method to optimize the model. The method to evaluate the accuracy of the model is cross-validation. Finally, it compares the effect of SGD and Adam in this project.

1 Introduction

Decision tree is a kind of multifunctional machine learning, which has been widely used in data mining, computer vision, nature language processing, autonomous driving, and artificial intelligence. It can implement classification and regression tasks and can fit complex data sets. Decision tree is also the basic part of random forest, which is one of the most powerful machine learning algorithms at present.

The major challenge for decision tree is because of any small changes in the training data will have an impact on the decision tree, it is easy to cause overfitting. Even using the same training data may get completely different models. The common solution is to set some hyperparameters to reduce the degree of freedom and pruning of the tree. Commonly used algorithms include ID3, C4.5, C0.5, and CATR.

The age of abalone can be estimated by the number of abalone rings. However, counting the number of rings is inefficient and inconvenient. Hence, it is meaningful to create a tree model to predict the number of abalone rings by the features which are easy to measure. The main steps include data processing, data visualization, modeling, data analyze and future direction.

2 Data

2.1 The Dataset

The dataset is offered by W. J Nash(1994) et, al. It includes 8 attributes and 4177 instances. The features are named and data types include nominal, continuous and integer. There is no missing value in the dataset.

2.2 Pre-processing

In the dataset, the sex of abalone is divided into three classes, M, F, I. To efficiently process data, replace three letters with three numbers 0, 1, -1. The According to the number of abalone rings, the abalones are divided into four categories: case 1 is 0-7 years; case to is 8-

10 years; case 3 is 11-15 years; case 4 is greater than 15 years. Replace these four cases with 1, 2, 3, 4 in the dataset.

2.3 Data visualization

2.3.1 Heatmap

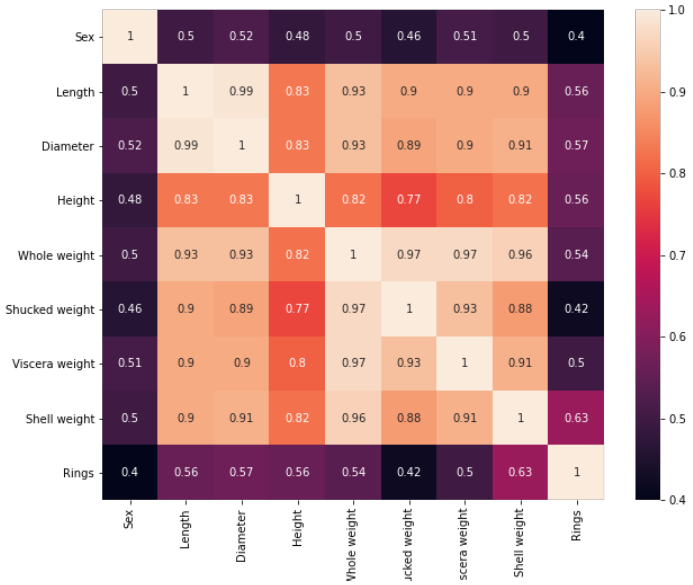

Figure 1 Correlation Heatmap

The sex is uncorrelated, but others are correlated. The height has the most correlated.

2.3.2 Scatter

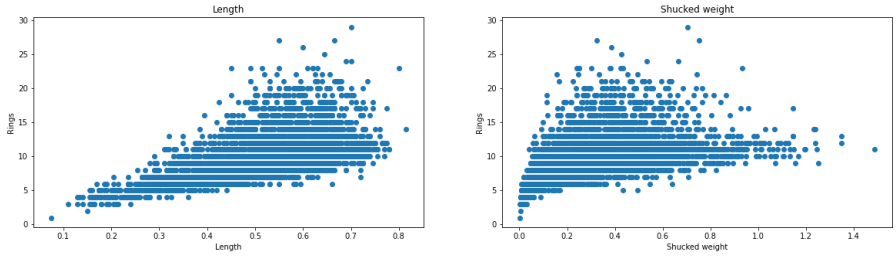

Figure 2 Length-Rings and Shucked weight Rings Scatter

Length has a significant positive correlation with ring-age.

The relation between shucked weight and rings is not significant as length.

2.3.3 Histogram

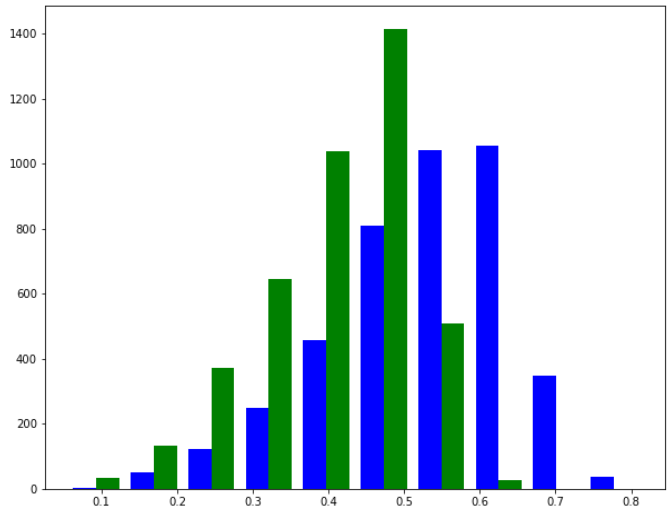

Figure 3 Features Frequency Distribution Histogram

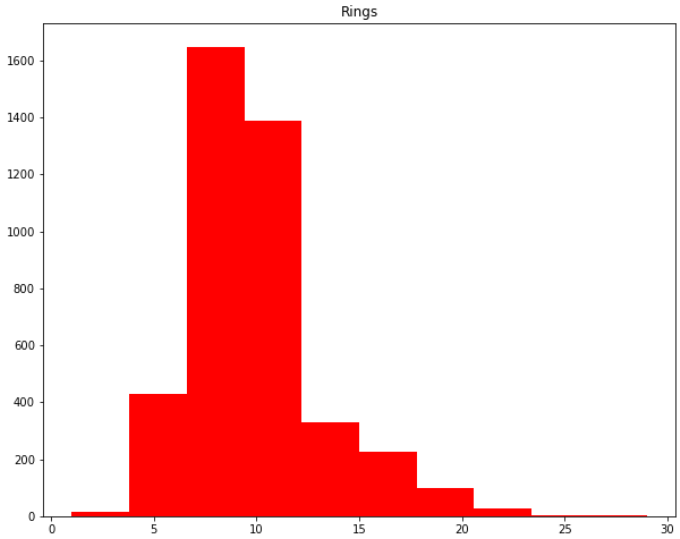

Figure 4 Rings Frequency Distribution Histogram

Most of data of features is distributed in the middle, while the data of rings are not which is a left-side distribution.

3 Modeling

3.1 Decision tree

Decision tree is machine learning algorithm which is usually used for classification problem and regression problem. It is also the basic function of random forest. Randomly divide the data set into 60/40% training set and test set. To make the experimental results more accurate, the data set is divided several times, and different training sets and test sets are obtained for

experimentation. The algorithm used to train decision tree in Scikit-Learn is CART (Classification And Regression Tree), its idea is to select a single feature and threshold to divide the train set and repeat the subsets. The maximum depth is controlled by hyperparameter(max_depth) or until cannot be divided.

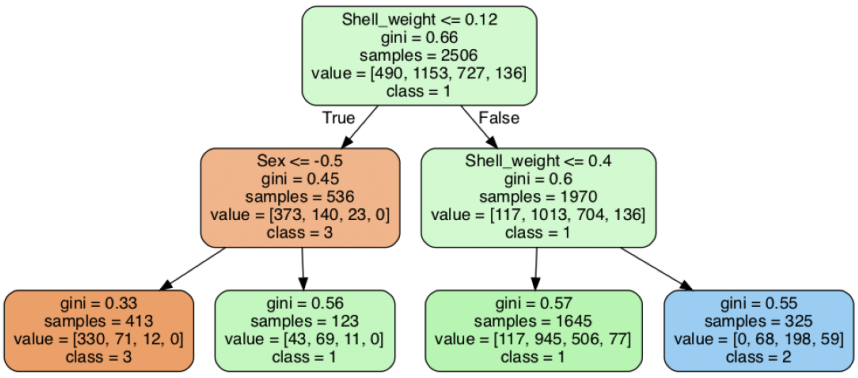

Figure 5 Decision Tree 1

The If-Then rule in this decision tree is:

If Shell_weight<=0.12 then
    If Sex<=0.5 then
        Class = 3
    Else class =1
Else:
    If Shell_weight<=0.4 then
        Class = 1
    Else class =2

Change the random state and obtain new couples of training set and test set. Use these new data sets can create different decision trees. Following is a decision tree with random state 10:

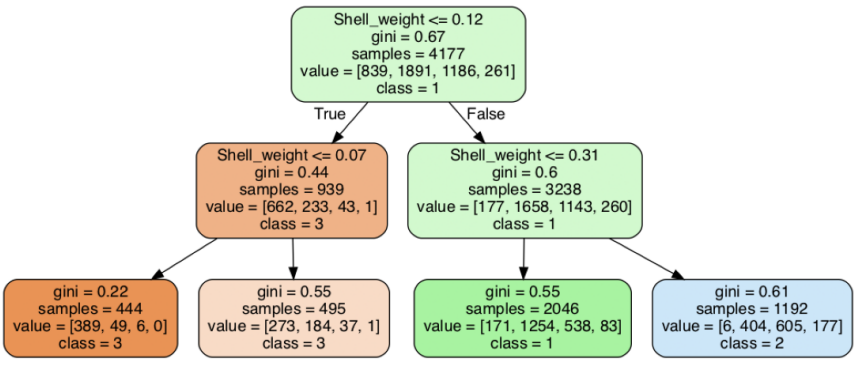

Figure 6 Decision Tree 2

The above figure is a simple decision tree. There are only three results in the tree while the correct results should be four classes. Hence, there should be more layers and nodes. However, if the number of depth is over large, it probably overfitting which should also be avoided. To explore suitable features, it is necessary to prune the tree. First, build a decision tree without any restrictions. Second, deleted nodes that (statistically) did not significantly improve the accuracy of the prediction. For example, if all the child nodes of a node are leaf nodes, and the split of the node only provides a small information gain (not statistically

significant). Usually standard statistical tests, such as the chi-square test, are used to determine whether the split is significant. If it is not statistically significant, delete the split and its children. Generally, pruning starts at a lower level and continues upwards until all unnecessary nodes are pruned. In the figure below, alpha is the threshold of the loss function. The figure shows the depth, the number of nodes and the accuracy of the new tree after the nodes whose value of loss function is smaller than alpha are pruned.

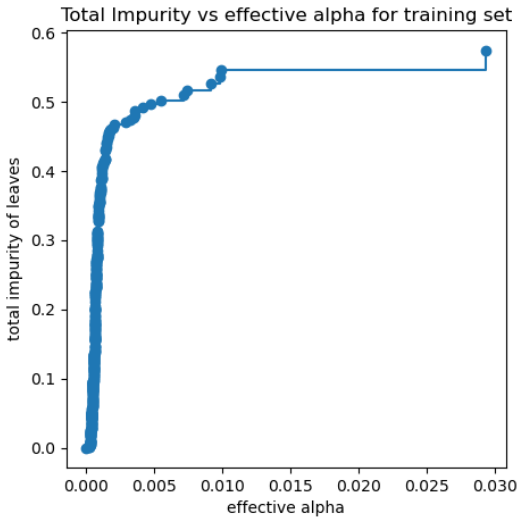

Figure 7 Impurity-Effective Alpha

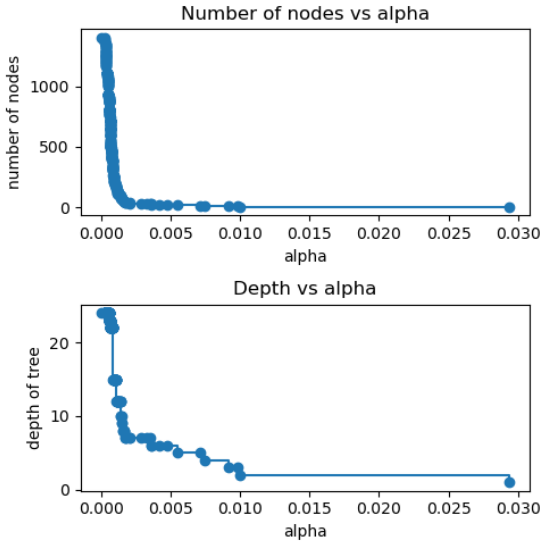

Figure 8 Node-Alpha

The minimum cost complexity pruning recursively finds the node with the "weakest link". The weakest link is characterized by effective alpha, in which the node with the smallest effective alpha is pruned first. Then, use the most effective alpha to train the decision tree.

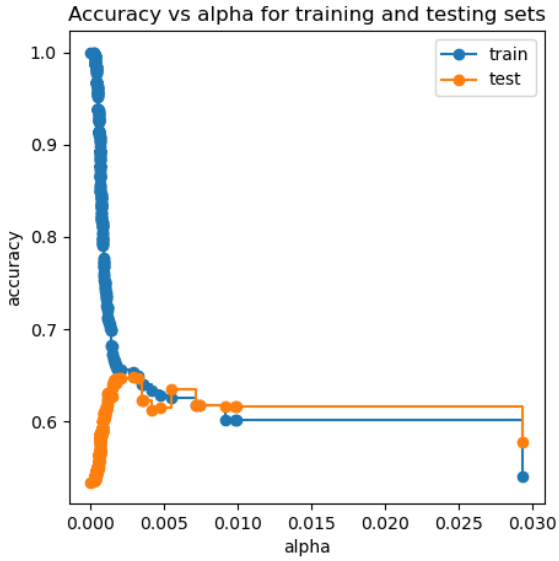

Figure 9 Accuracy-Alpha

From the above figure, under the initial condition, the accuracy of train set is 1.0 which means overfitting while the accuracy of test set is around 0.1 which is underfitting. With the increasing of alpha, they reach a more suitable position when the value of alpha is around 0.003, while the depth of tree is 5. Hence, the layers over the fifth should be pruned.

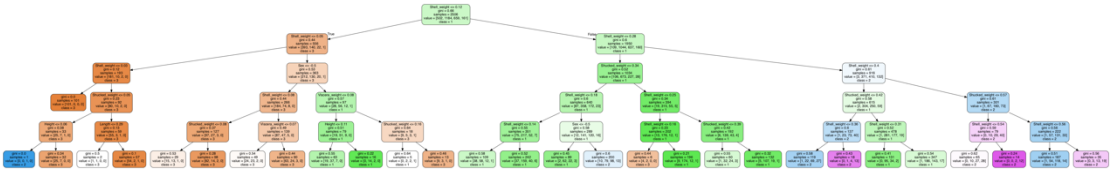

Figure 10 Depth 5 Decision Tree

3.2 Bagging of Trees via Random Forests

There is an important problem for decision tree model which is overfitting. A method is random forest technology. The idea is to create several decision trees and combine their output together. The method to get random decision trees is to select a subset by uniformly sampling and replacing the training data set, which is called bagging algorithm. The method to evaluate the model is cross-validation.

Estimator	50	100	150	200	250	300	350	400	450	500
Accuracy(train)	0.6313	0.6432	0.6472	0.6389	0.6401	0.6389	0.6385	0.6377	0.6409	0.6381
Accuracy(test)	0.6254	0.6362	0.6397	0.6409	0.6397	0.6361	0.6373	0.6415	0.6397	0.6403

Table 1 Accuracy of Train/Test in Different Estimator

Estimator	mean	std	Confident interval
50	0.617	0.023	(0.571, 0.663)
100	0.620	0.021	(0.579, 0.661)
150	0.619	0.022	(0.575, 0.663)
200	0.619	0.024	(0.572, 0.665)
250	0.620	0.022	(0.577, 0.663)
300	0.621	0.021	(0.580, 0.663)

350	0.621	0.021	(0.580, 0.662)
400	0.621	0.022	(0.577, 0.665)
450	0.620	0.023	(0.575, 0.665)
500	0.620	0.020	(0.583, 0.663)

Table 2 Bagging

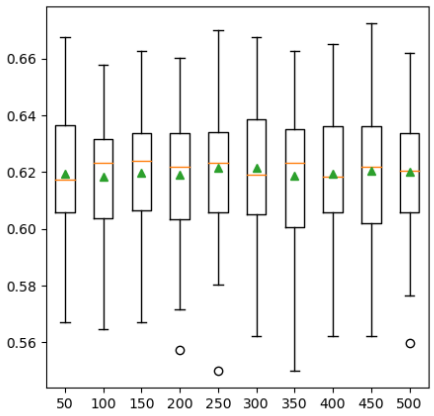

Figure 11 Bagging

3.4 Adam and SGD

For a given number of hidden neurons in a single hidden layer, compare Adam with SGD and report the percentage of correct classification as the performance of model. From the below figure, SGD has a better mean percentage of correct classification while Adam is more stable because its standard deviation is smaller. In this case, the learning rate is 0.02, the number of hidden nodes is 15, the number of layers is 3.

	mean	std
SGD	0.652	0.003
Adam	0.639	0.008

Table 3 comparison of SGD and Adam

The mean of SGD is higher than Adam while the std is lower. It means that compared to Adam SGD has higher accuracy and stability.

The confusion matrix is as following:

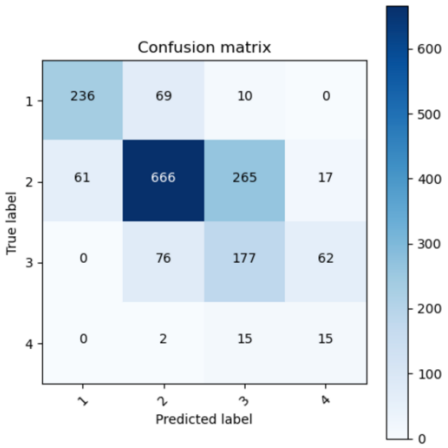

Figure 12 Confusion Function

A method to visualize the performance of multi-class classification problems is AUC and ROC curve which is the area under the curve (AUC) and receiver operating characteristic (ROC) curves. It is also written as AUROC.

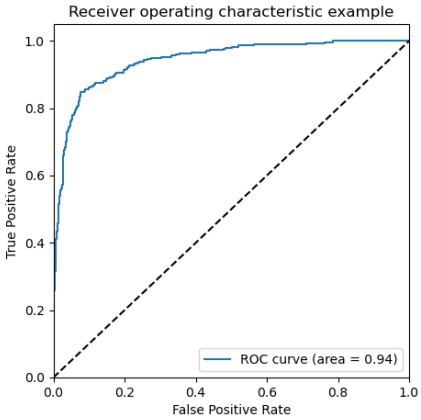

Figure 13 AUROC for class 1

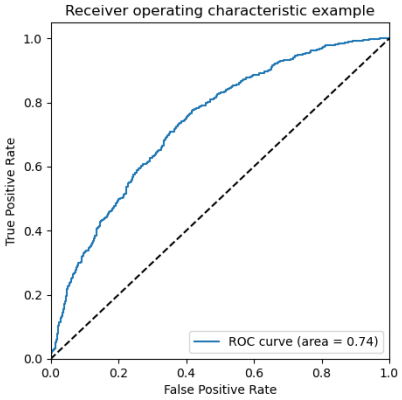

Figure 14 AUROC for class 2

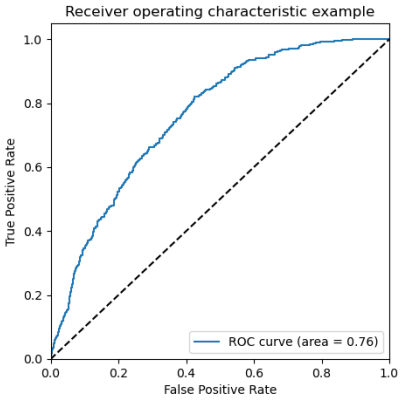

Figure 15 AUROC for class 3

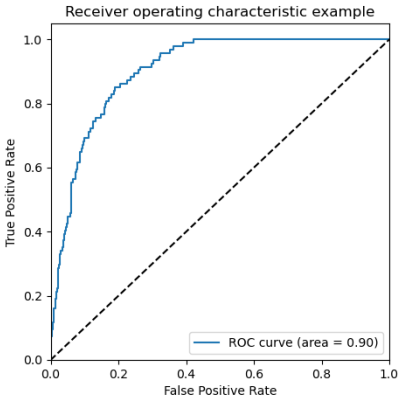

Figure 16 AUROC for class 4

Because of the area of class 1 is the largest, this model has the best simulation results for class 1.

4 Conclusion

In the problem, the best depth of the decision tree is 5 which can reach an accuracy rate of more than 60. The influence of the number of trees in a random forest is little. The best number of trees is around 400. In this case, SGD can reach higher accuracy mean and smaller standard deviation, which is 0.652 and 0.003 respectively. It has better effect compared to Adam while the accuracy mean is 0.639 and standard deviation is 0.008. Judging from the confusion matrix and ROC curve, the most accurate prediction in the model is the prediction of class 1.

5 Limitations and future directions

In the model, the accuracy of the model is only around 60 percent, the number did not meet expectation. There are three possible reasons: first, the size of dataset is small; second, the correlation between features and the number of rings is not obvious enough; third, unrecorded factors such as environmental factors have a greater impact on the ring number. In the future work, more information should be recorded and explored.

6 Reference

[1]"Post pruning decision trees with cost complexity pruning", scikit-learn, 2021. [Online]. Available: https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html. [Accessed: 17- Nov- 2021].

[2]DIANE M. DENNIS., *FOUNDATION OF PYTHON NETWORK PROGRAMMING*. [Place of publication not identified]: TRITECH, 2018.

[3]J. Goerzen, *FOUNDATION OF PYTHON NETWORK PROGRAMMING*. Beijing: Publishing House of Electronic Industry, 2007.

[4]"Ed — Digital Learning Platform", Edstem.org, 2021. [Online]. Available: https://edstem.org/au/courses/6212/lessons/13877/slides/100389. [Accessed: 19- Nov- 2021].

[7]"Introduction to the principle of decision tree pruning (cart pruning)", *Blog.csdn.net,* 2021. [Online]. Available: https://blog.csdn.net/zhengzhenxian/article/details/79083643. [Accessed: 19- Nov- 2021].

[8]"Visualize a Decision Tree w/ Python + Scikit-Learn", *Kaggle.com*, 2021. [Online]. Available: https://www.kaggle.com/willkoehrsen/visualize-a-decision-tree-w-python-scikit-learn. [Accessed: 19- Nov- 2021].

[9]"Ed — Digital Learning Platform", *Edstem.org*, 2021. [Online]. Available: https://edstem.org/au/courses/6212/lessons/13873/slides/100317. [Accessed: 19- Nov- 2021].

[10]"Ed — Digital Learning Platform", *Edstem.org,* 2021. [Online]. Available: https://edstem.org/au/courses/6212/lessons/13873/slides/100318. [Accessed: 19- Nov- 2021].