Supervised Learning: Regression and Classification

by

Yuchen Li (s192234)

Baixun Wang (s192179)

12 November 2019

Project 2, Introduction to Machine Learning and Data Mining E19

Technical University of Denmark

Students contributions

Section	Contribution
Report	Yuchen: 50%, Baixun: 50%
Analysis	Yuchen: 70%, Baixun: 30%
Code	Yuchen: 70%, Baixun: 30%
Revision	Yuchen: 30%, Baixun: 70%

Supervised Learning: Regression and Classification

1. Introduction

This project follows the previous project "*Abalone Data Set: Feature Selection and Data Visualization*". Hence, the same data set is from UCI Machine Learning Repository.[1] This dataset contains 9 attributes (sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and rings) and 4177 observations. The detailed attributes are shown in Table 1. For the nominal attribute "sex", it is substituted with "F", "I", and "M" by one-of-*K* coding.

Attributes	Measurement Unit	Data type	Description
Sex	---	Nominal, discrete	M, F and I (infant)
Length	Millimeter	Ratio, continuous	Longest shell measurement
Diameter	Millimeter	Ratio, continuous	Perpendicular to length
Height	Millimeter	Ratio, continuous	With meat in shell
Whole weight	Grams	Ratio, continuous	Whole abalone
Shucked weight	Grams	Ratio, continuous	Weight of meat
Viscera weight	Grams	Ratio, continuous	Gut weight (after bleeding)
Shell weight	Grams	Ratio, continuous	After being dried
Rings	---	Interval, discrete	+1.5 gives the age in years

Table 1. Attributes of abalone data set and their types, descriptions.

2. Regression

In regression section, elementary linear regression is used to predict the rings of abalone. Later, performance of different regression models, linear regression, artificial neural network, and baseline, will be evaluated by two-layer cross-validation and statistical test.

2.1.Regularized linear regression

Abalone data set is basically to predict the age of abalone from their physical measurements. Hence, the attribute "Rings" is selected as predicted variable based on other variables. In this section, an elementary linear regression model will be constructed and used to predict the age of abalone. The data matrix X is normalized before generating the regression model.

In order to penalize large weight and prevent overfitting, a L2 regularization term with parameter λ is introduced to the cost function. The parameter ω of regularized linear regression is:

$$\omega^* = \left(\tilde{X}^T\tilde{X} + \lambda I\right)^{-1}\left(\tilde{X}^T y\right)$$

Meanwhile, to choose a reasonable range of values of λ, a K-fold (K=10) cross-validation is applied to estimate the generalization error for each λ.

[1] Data set is from Nash et al., 1995. https://archive.ics.uci.edu/ml/datasets/Abalone

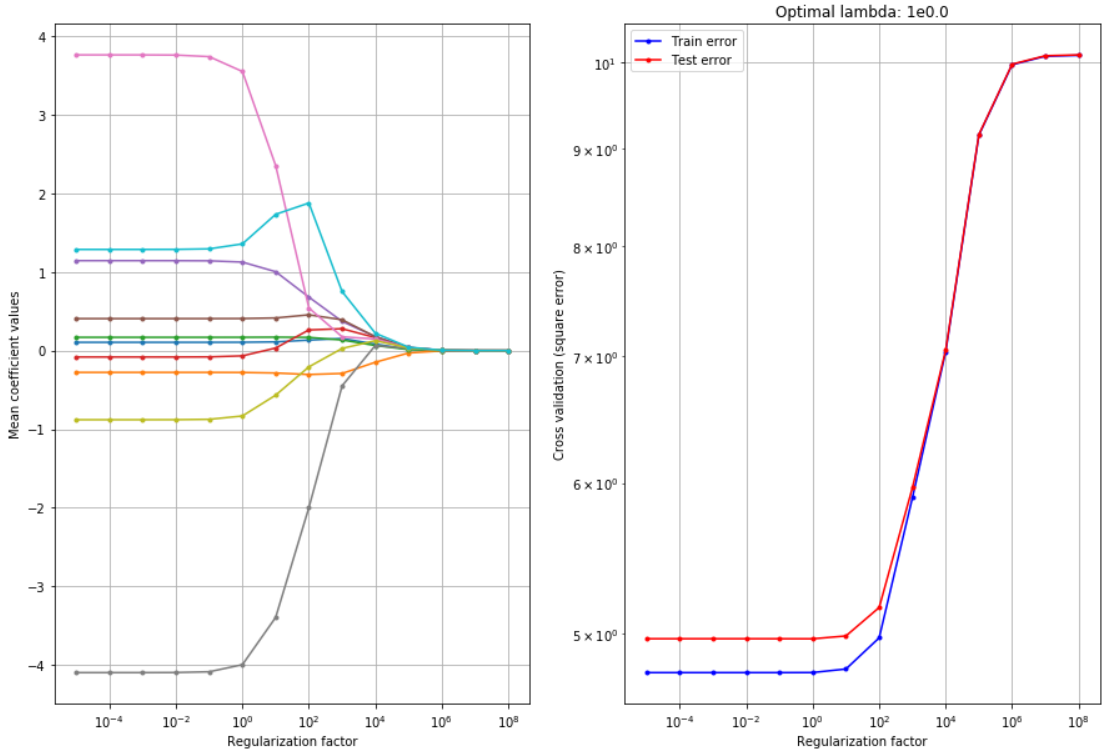

Fig 1. Mean coefficient values (left) and squared error (right) versus regularization factor

In the K=10 cross-validation, the mean coefficient values and the square error versus regularization factor is shown in Fig 1. The coefficient values except offset converge to 0 as regularization factor increases. If regularization factor is large enough, the model will become baseline model. In order to find the optimal regularization factor, the square error of train sets and test sets were calculated. The optimal regularization error is 1 in this case because the square error is the minimum value ($E^{\text{train}} = 4.76792,\ E^{\text{test}} = 4.96713$). An appropriate regularization factor can prevent overfitting when it is small enough. The coefficient term is:

$$w^* = \begin{bmatrix} \text{Offset} & 9.9399 \\ \text{F} & 0.1051 \\ \text{I} & -0.279 \\ \text{M} & 0.1694 \\ \text{Length} & -0.0643 \\ \text{Diameter} & 1.1446 \\ \text{Height} & 0.3671 \\ \text{Whole weight} & 3.5655 \\ \text{Shucked weight} & -4.0057 \\ \text{Viscera weight} & -0.8269 \\ \text{Shell weight} & 1.3611 \end{bmatrix}$$

A linear regression model was generated by adding regularization term. It uses linear regression equation:

$$\hat{y} = \tilde{X}^T w^* + \epsilon$$

to predict the rings of abalone based on test data set. The comparison between predicted rings and true rings are shown in Fig 2. Intuitively, the predicted rings follow a similar trend as ideal results (the green line where predicted rings equal to true ring). The trend line (orange line) of real results is slightly biased. To obtain a detailed evaluation of predicting performance, a histogram of residuals

(the difference between estimated rings with true rings) was plotted. The distribution is not centered at 0. It is obvious to tell the predicted rings are usually larger than the true rings in most cases.

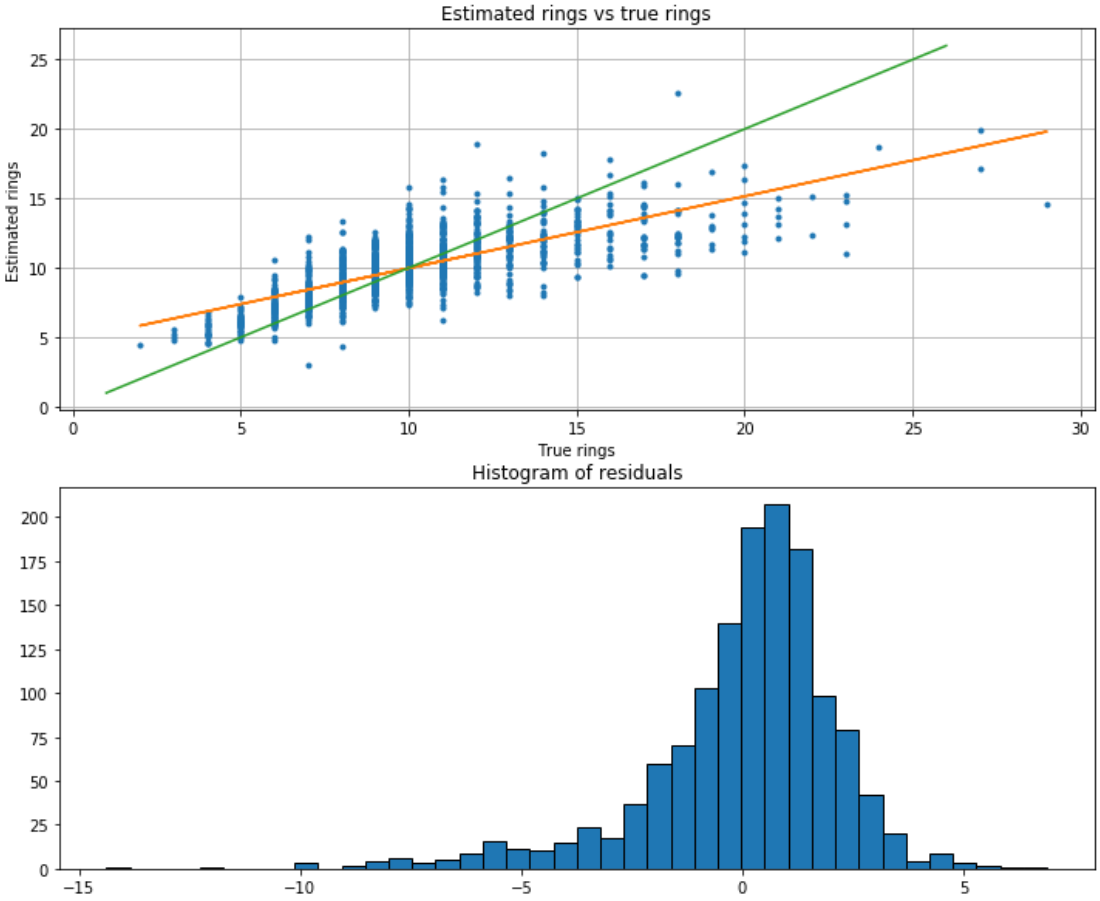

Fig 2. The upper figure is estimated rings versus true rings. The green line is ideal result and the orange line is actual result. The lower figure is a histogram of residuals (estimated value subtracted by true value).

A possible cause of this result is that the relations between the rings and most selected attributes are nonlinear. The distributions of rings versus different attributes are shown in Fig 3. Length and diameter ($w_5 = 1.1446$) have a similar distribution pattern and a nonlinear relation to the rings. The rings look like a log function to whole weight ($w_7 = 3.5655$), shucked weight ($w_8 = -4.0057$), viscera weight, and shell weight ($w_{10} = 1.3611$). A further modified linear regression may give more precise predicted value.

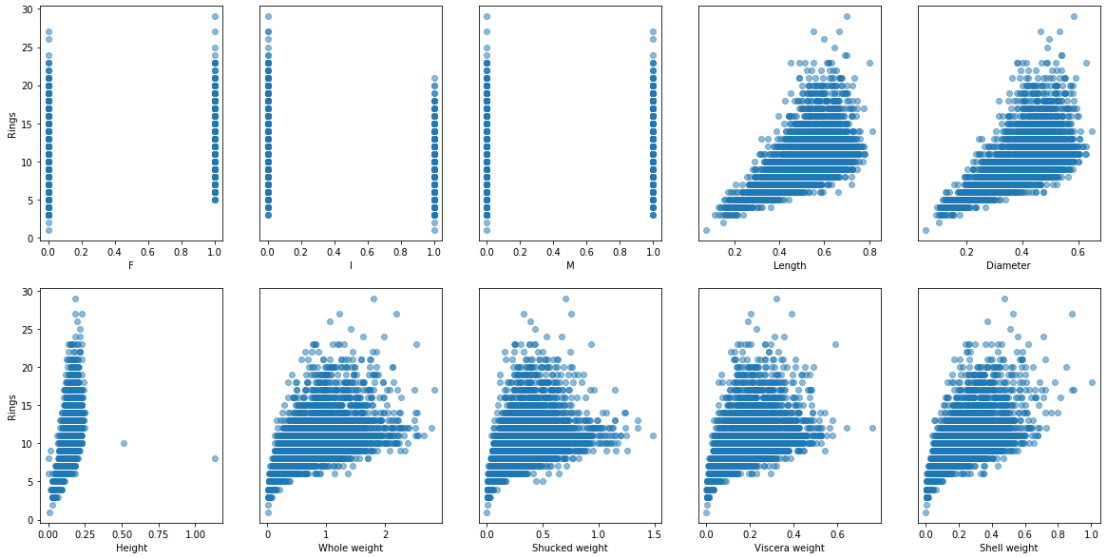

Fig 3. These figures are scattered plot of rings versus different attributes.

2.2.Performance evaluation

In this section, linear regression, artificial neural network (ANN) regression, and baseline model are selected to evaluate performance by using two-level cross-validation and statistical test.

Before evaluating different regression models, a few test-runs were designed to obtain a range of parameters. The complexity-controlling parameter h of ANN has a range from 1 to 7. Intuitively, the increasing of number of hidden units leads to lower the error, in Fig 4. Nevertheless however, a large range of numbers requires expensive computation, and the loss decrease slowly when h is large. Based on previous tests, the regularization factor of linear regression ranges from 10^{-5} to 10^{8}.

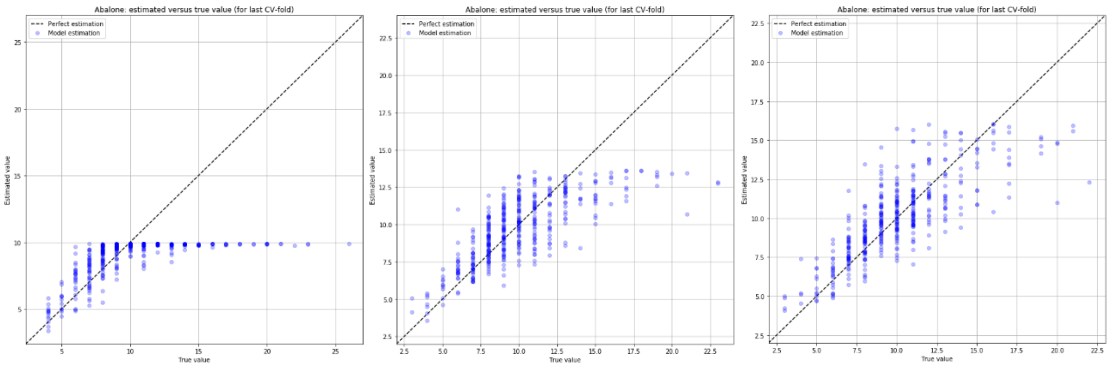

Fig 4. Figures of predicted value versus true value with different complexity-controlling parameter h. (From left to right, $h = 1, 4, 7$)

Firstly, for the two-layer cross-validation ($K_1 = K_2 = 10$), generalization errors E_i^{test} are estimated on $\mathcal{D}_i^{\text{test}}$ and measured by squared loss:

$$E = \frac{1}{N^{\text{test}}} \sum_{i=1}^{\text{test}} (y_i - \hat{y}_i)^2$$

The results are shown in Table 2. The selected number of hidden units h^* of the ANN range from

4 to 7. The optimal regularization factors λ^* are basically concentrated in 0.1 and 1 which is the same value to the previous section. By comparing the generalization error of each regression model, ANN has a slightly better performance than linear regression. For a few outer folds, the generalization errors are lower than the those corresponded in ANN. However, it is clear that both of ANN and linear regression have better performance the baseline model.

Outer fold	ANN		Linear regression		Baseline
	h^*	E_i^{test}	λ^*	E_i^{test}	E_i^{test}
1	5	4.9995	1	5.9605	9.8181
2	5	4.8630	1	4.5917	10.074
3	5	4.1628	0.1	4.4074	9.8157
4	7	5.0466	1	5.1309	9.9761
5	7	4.9970	0.1	5.0659	10.072
6	4	4.7966	0.1	4.8854	9.7895
7	4	4.9168	0.1	5.0395	10.106
8	6	4.7647	1	4.6026	9.8369
9	5	4.3920	1	4.1383	9.8273
10	6	4.8587	0.1	5.1331	10.010

Table 2. Two-layer cross-validation ($K_1 = K_2 = 10$) of ANN, linear regression, and baseline with optimal complexity-controlling parameter from inner cross-validation.

In addition, statistical methods, that is hypothesis test and confidence interval, were applied to evaluate the performance difference pairwise. The "**Setup I**: the training set is fixed" was chosen and pair t-test was used. In these cases, null hypothesis H_0 is "Model A and model B have the same performance in regression". From a selected cross-validation, the hypothesis test and confidence interval were evaluated by the difference of square loss:

$$z_i = z_i^A - z_i^B$$

The null hypothesis could be transformed to $Z = 0$. Under $\alpha = 0.05$, p-value of ANN and linear regression in 6.2279×10^{-6}, which indicates the null hypothesis is rejected, i.e., ANN and linear regression have different performance in this regression task. However, the p-value of ANN and baseline is strange, 1. The result basically states the no evidence against null hypothesis and it cannot be rejected. Comparing linear regression and baseline gave a 2.787×10^{-64} of p-value. This also indicates the linear regression and baseline model have different performance.

To get a further comparison, the confidence interval between ANN and linear regression was computed. Its value is $(-0.768, -0.293)$, which means ANN are slightly better than linear regression and the performance difference may be 0.5 reflected in the generalization error. The confidence intervals between ANN and baseline, linear regression and baseline are $(-6.109, -5.012)$ and $(-5.584, -4.476)$, respectively. The confidence intervals show that ANN and linear regression have significant performance difference to baseline mode, which results in the same conclusion in the previous section.

3. Classification

In classification section, the predicted variable is different from the previous regression section. The sex is assigned to be the predicted variable, because sex is a discrete nominal attribute that may be easier to address in classification. Sex has three values, "F", "I", and "M", in abalone data set. Hence, it is a multiclass classification problem. Hence, the multinomial regression is substituted for logistic regression for multiclass. All the following logistic regression refers to multinomial regression.

Likewise, logistic regression, ANN, and baseline mode were selected for classification. For complexity-controlling parameter, number of hidden units of the ANN ranges from 1 to 7. The range of the regularization factor of logistic regression is from 10^{-6} to 10^{6}. The generalization error is measured by:

$$E = \frac{\{\text{Number of misclassified observations}\}}{N^{\text{test}}}$$

The two-layer cross-validation was computed and the results shown in Table 2. The generalization error of ANN and logistic regression have no significant difference. The complexity-controlling parameter of ANN is 4 and 7, and they are 1×10^{-6} and 1×10^{5}. The generalization errors of these two classifiers are greater than those of baseline model, which suggests the performance of ANN and logistic regression are even not good as baseline model.

Outer fold	ANN		Logistic regression		Baseline
	h^*	E_i^{test}	λ^*	E_i^{test}	E_i^{test}
1	7	0.4665	10000	0.4976	0.3923
2	4	0.4545	10000	0.4761	0.3708
3	7	0.4545	10000	0.4498	0.3612
4	4	0.4426	10000	0.4617	0.3852
5	7	0.4187	10000	0.4211	0.3804
6	4	0.4115	10000	0.4617	0.3493
7	7	0.4498	10000	0.4928	0.3493
8	4	0.4916	1e-05	0.4844	0.3741
9	7	0.4556	10000	0.4748	0.3405
10	4	0.4293	1e-05	0.4149	0.3765

Table 3. Two-layer cross-validation ($K_1 = K_2 = 10$) of ANN, logistic regression, and baseline with optimal complexity-controlling parameter from inner cross-validation.

Furthermore, statistical evaluation was computed on different classifier pairwise, and **Setup I** was selected. In the hypothesis test, the null hypothesis H_0 is two classifiers have the same performance. It is evaluated by McNemar's test, where n_{11} is the number of both models predicted correctly, n_{22} is the number of both models give wrong predictions. The computed p-value is shown in Table 3. With $\alpha = 0.05$, all these p-values are larger than 0.1, which means these three null hypotheses cannot be rejected, i.e., they can be considered to have the same performance in classification. The confidence interval of ANN and logistic regression is $(-0.0076, 0.0163)$. This basically suggests that there is almost the same performance between ANN and logistic regression. Meanwhile, ANN and baseline or logistic regression and baseline have similar value of confidence intervals, which

indicates they have similar performance and worse than the baseline model.

	p-value	Confidence interval
ANN – logistic regression	0.6749	$(-0.0076, 0.0163)$
Logistic regression – baseline	2.6532	$(0.1680, 0.2220)$
ANN - baseline	2.9139	$(0.1717, 0.2270)$

Table 4. p-value and confidence interval of comparing classifier pairwise.

The cause of performance of above classification may be the attributes of abalone data set. To have an intuitive recognition, two attributes were simply chosen to give an illustrative explanation, Fig 5. Basically, there are somewhat relations between these attributes. "I" class is always at the lower part of the figures because infant abalone usually have lower values of physical measurements. Nevertheless, it is difficult to draw class boundaries between these three classes. The data of scatter plot are mixed and stacked together. Hence, the misclassification rates with all attributes are usually slightly lower than half as shown in Table 3.

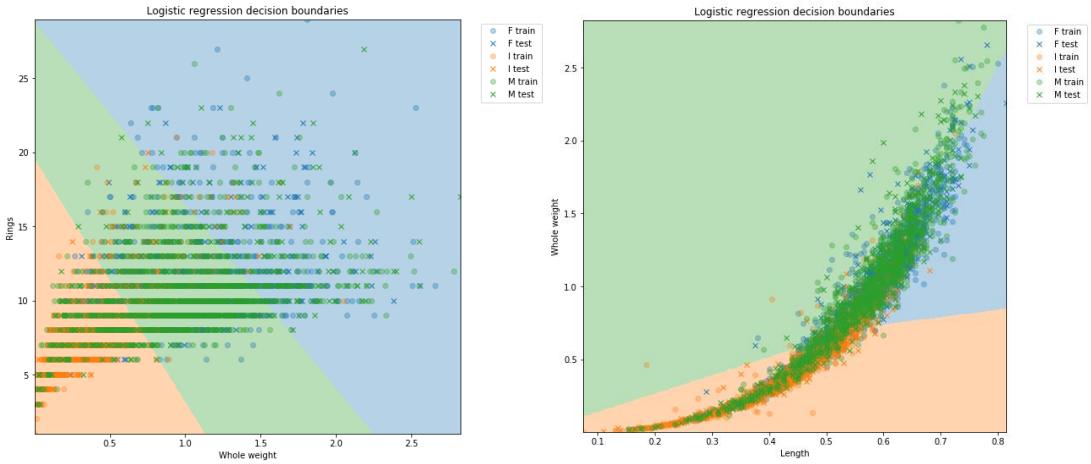

Fig 5. Logistic regression results of rings versus whole weight (left) and whole weight versus length (right). Orange is classified for infant, green is for male, and blue is for female.

For different regularization factor of L2 penalty term, it ranges from 1×10^{-6} to 1×10^{5}. For classification of binary class, logistic regression is generated from linear regression by applying the logistic transformation:

$$\hat{y}_i = \sigma(\tilde{x}_i^T w) = \frac{1}{1 + e^{-(\tilde{x}_i^T w)}}$$

The coefficient term w is evaluated by Bernoulli distribution. For classification of multinomial regression, the class is predicted by:

$$\hat{y}_i = \boldsymbol{W}\tilde{x}_i$$

In this case, the matrix of weight $\boldsymbol{W}$ is computed by 'lbfgs' solver with L2 penalty.

$$\boldsymbol{W} = \begin{bmatrix} \omega_1^T \\ \omega_2^T \\ \omega_3^T \end{bmatrix}, \quad w_1 = \begin{bmatrix} -0.8896 \\ 0.7080 \\ 0.2553 \\ 0.7982 \\ -0.3166 \\ 0.5392 \\ -0.2850 \\ 0.2365 \end{bmatrix}, \quad w_2 = \begin{bmatrix} 1.8503 \\ -1.0356 \\ -0.2182 \\ -1.2948 \\ -0.0395 \\ -0.9814 \\ 0.4049 \\ -0.4623 \end{bmatrix}, \quad w_3 = \begin{bmatrix} -0.9608 \\ 0.3276 \\ -0.0371 \\ 0.4967 \\ 0.3561 \\ 0.4423 \\ -0.1199 \\ 0.2258 \end{bmatrix}$$

This matrix of weight shows that length, diameter, whole weight, and viscera weight are more likely to be dominant. Comparing with previous relevant features of regression model (mainly diameter, whole weight, shucked weight, and shell weight), not all features are considered to be relevant to both of regression and classification. There are also overlaps between features of two supervised learning task.

4. Discussion

In regression section, the predictions of linear regression are fairly acceptable. However, a nonlinear regression may be more compatible to abalone data set because of the nonlinear relations between attributes and predicted variable. As a comprehensive linear regression, the ANN regression could address nonlinear relation more effectively than linear regression. Moreover, it is vital to choose appropriate complexity-controlling parameter to reach the balance between precision and computation. However, the results indicate this abalone data set may be not appropriate for the classification task.

All the models used in this project, i.e., linear regression, ANN, logistic regression (multinomial regression), and baseline, can be considered to be derivative of linear regression equation. By tuning parameters, they can be applied to both regression and classification tasks.