Abstract

The data featuring physical measurements of abalone is obtained from UCI Machine Learning Repository and analyzed for pattern recognition to predict the age of abalone based on its physical characteristics. A Linear Fit multivariate model is constructed to predict the age. Predictions on test set are obtained with a mean of 9.98 and RMSE of 2.26. The accuracy of Linear Fit model is confirmed with the Stochastic Gradient Boosting model which shows only a slight improvement with RMSE of 2.22.

Aim

The aim of this project is to predict the age of abalone from its physical measurements using the data collected from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/machine-learning-databases/abalone).

Introduction

Data comes from an original (nonmachinelearning) study: Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 10343288).

The aim of the project is to predict the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem. From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).

Summary Statistics

Table 1: Overview of variables in the data

	n	mean	sd	median	min	max
Length	4177	0.52	0.12	0.55	0.08	0.82
Diameter	4177	0.41	0.1	0.42	0.06	0.65
Height	4177	0.14	0.04	0.14	0	1.13
Wholeweight	4177	0.83	0.49	0.8	0	2.83
Shuckedweight	4177	0.36	0.22	0.34	0	1.49
Visceraweight	4177	0.18	0.11	0.17	0	0.76
Shellweight	4177	0.24	0.14	0.23	0	1
Rings	4177	9.93	3.22	9	1	29

Exploratory data analysis

Figure 1: Pairwise relationships between variables for all abalones