Machine Learning Engineer Nanodegree

Capstone Proposal

Aman Aggarwal
April 4th, 2017

Domain Background

Abalone is a shellfish considered a delicacy in many parts of the world. The abalone shell and the meat is of value. The abalone dataset from UCI Machine Learning Archives comes with the goal of attempting to predict abalone age given various descriptive attributes of the abalone. You can find the link to this dataset here. The dataset comes from a 1994 study "The Population Biology of Abalone" in Tasmania. It was donated to the UCI Machine Learning Repository in 1995 by Sam Waugh from the Department of Computer Science at the University of Tasmania (Australia).

In this project, we are going to build a regression model that will predict the age of abalones based on their physical measurements. Since this a regression problem with dependent variable already given to us, the machine learning technique used here will be a supervised learning model.

References :

1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.2321&rep=rep1&type=pdf – Helped me to understand the data mining concepts that'll be used in this project like treating missing values, conversion of continuous variable to categorical variable, attributes correlation.
2. http://www.cs.toronto.edu/~delve/data/abalone/abaloneDetail.html- This reference material helped me in understanding the data description and also the benchmark models that have been analyzed in the past.

Problem Statement

Abalone dataset consists of physical information about the abalones like their Weight, Height etc. which can be used to predict the number of rings on the abalone which, in turn, helps us to find out their age. The age of abalone is determined by cutting the shell through the cone and counting the number of rings through a microscope.

We will use Linear Regression to make our predictive model to solve this problem. The process involves cleaning/modification of the continuous data into Boolean or discrete data. Then dividing the data into training and testing set and building a Decision Tree regressor over it. Then we will try to optimize our model for more accuracy.

Datasets and Inputs

The Abalone dataset consists of approximately 4000 records and 9 variables that consists of both categorical and continuous variables.

The 9 variables description is as follows:

Attribute	Domain	Variable type
Sex	{M, F, I}	Categorical
Length	[0.075, 0.815]	Continuous
Diameter	[0.055, 0.65]	Continuous
Height	[0.0, 1.13]	Continuous
Whole_weight	[0.0020, 2.8255]	Continuous
Shucked_weight	[0.0010, 1.488]	Continuous
Viscera_weight	[0.0005, 0.76]	Continuous
Shell_weight	[0.0015, 1.005]	Continuous
Rings	{15, 7, 9, 10, 8, 20, 16, 19, 14, 11, 12, 18, 13, 5, 4, 6, 21, 17, 22, 1, 3, 26, 23, 29, 2, 27, 25, 24}	Continuous

Solution Statement

The solution to this problem is to increase our prediction accuracy by reducing the **Root Mean Square Error (RMSE)** of our model to the minimum possible unit. We will use DecisionTreeRegressor here as we need to estimate a continuous variable instead of a discrete one. It is a basic regression technique and it is used in the initial phases of predictive modelling to start off with the solution and furthermore, we can choose to optimize it in the future. Next, we plot the learning curves like scatterplot to understand the root mean square error found in both the training and testing data. If required, we can optimize our model by using RandomForestRegressor to reduce root mean square error.

Benchmark Model

Many attempts have been made on improving the accuracy of this dataset. Previous works have shown that on using:

a. Polynomial Regression technique with RMSE = 2.5
b. Lasso Regression technique with RMSE = 2.1

We can see that the prediction models are quite successful with such a small dataset. We will try to reduce the RMSE with our prediction model and understand the underlying complexity of this dataset empirically.

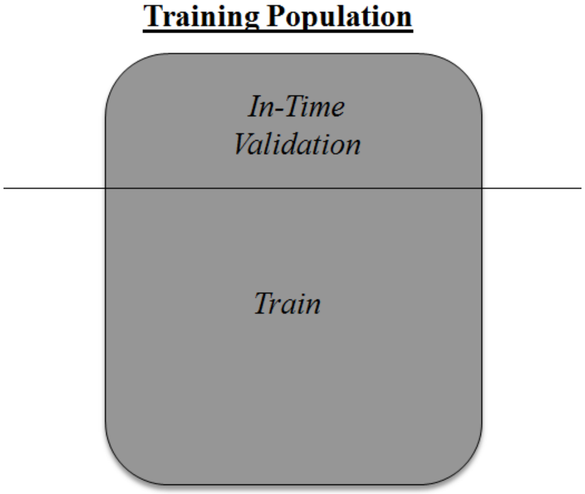
This problem has also been seen as a classification problem and has been solved with the techniques such as SVM, Logistic Regression but the accuracy was quite low. Here, we'll only take our dig at Linear regression to make a better estimator of the rings.

Evaluation Metrics

1. **RMSE** is the most popular evaluation metric used in regression problems. It follows an assumption that error are unbiased and follow a normal distribution.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$ Where, N is Total Number of Observations.

2. **Cross Validation** - Cross Validation is one of the most important concepts in any type of data modelling. It simply says, try to leave a sample on which you do not train the model and test the model on this sample before finalizing it.



Project Design

In this project, we will use Python libraries like **URLretrieve** to download the dataset from the UCI repository website, **numpy** and **pandas** to shapen and describe the data, **matplotlib** to plot the graphs and scatterplots, **sklearn** to make use of regression and evaluate performance metrics and cross-validate the data.

The project design is as follows:

1. Download the dataset and explore it. Describe the data using statistical tools like mean, standard deviation, range etc.
2. Data wrangling – Conversion of continuous variables to categorical/Boolean variables for analysis.
3. Check for attribute uselfulness and correlation between the attributes by plotting graphs.
4. Split data into training (75%) and testing set(25%).
5. Build a prediction model using decision tree regressor and explore the results on scatterplot and check for RSME.
6. Use cross validation and run the prediction model several times on the cross-validated datasets.
7. Plot learning curves to check if there is a scope of improvement. If it’s there, we can, maybe, use advanced methods like RandomForestRegressor and optimize our model.