Proposal for Machine Learning Project

Hao Heng, Shuang Ma, Ziyu Huang

June 13th, 2020

1. Motivation

Nowadays, abalone market is extending in the past decade years in China and even get wider recently. Abalone is a shellfish considered as a delicacy not only in China but also in many parts of the world. Based on information from Wikipedia, abalone is an excellent source of pantothenic acid and iron, and a nutritious food resource, and 100 grams of abalone can offer more than twenty percent recommended daily intake of these nutrients. The abalone farms are usually in America, East Asia and Australia. The economic value of abalone is positively correlated with the age of it. Hence, the goal to detect the age of abalone is important for farmers and customers to evaluate the price. However, farmers usually determined abalone age by cutting the shell through the cone, staining it, then counting the number of rings through a microscope, and these processes are boring and time consuming. Our goal is to find out the best indicators to determine the rings in order to estimate the age of abalones

2. Project Plan

To do the task, first we need to acquire data from UCI Machine Learning Repository. Then we should determine the attributes we want to use in the machine learning algorithm. A last we should find the best machine learning approach to give us the results.

2.1: Data Acquisition

We downloaded the data from the following website. There are total 4177 samples with nine attributes.

- Download data on UCI (http://archive.ics.uci.edu/)

2.2: Attributes Determination

There are total eight attributes that may correlated with the age of abalone. Apparently that the Whole Weight and Shell Weight, Length and Diameter probably have the same influence on the age of abalone.

2.3: Machine Learning Approach

We plan to use PyCharm because it is open source such as Scikit-Learn package to implement Neural Network first to check the result. After that we will try Random Forest Classifier to see if there are any improvement of the outcomes. Then we will try Logistic Regression that we are about to learn in class to see how the outcomes look like. Finally, we provide a basic Linear Regression to compare with the previous three models. We will use 70% of the data as training set and the rest 30% as test set. We also plan to perform cross validation to avoid overfitting problem.

3. Evaluate Performance

We are going to use metrics like confusion matrix, accuracy between target and predicted rings, classification report and probably MSE to judge the performance of the network.

4. Schedule

Data Cleaning	15th – 17th Jun, 2020
Feature Engineering	18th – 21th Jun, 2020
Feature Selection	22th – 23th Jun, 2020
Training model and parameter optimization (If needed)	23th - 25th Jun, 2020
Model comparison	26th - 29th Jun, 2020
Presentation and submit report	30th Jun, 2020