[Project Code: ABNN]

Abalone Age Prediction from Physical Measurements using Artificial Neural Networks

Project Duration: 25-Feb-2024 ~~ 16-Mar-2024
Submission Information: (via) CSE-Moodle

Objective:

Predicting 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).

Tasks to be done:

1. Starter code provided does this using python, can be used: Randomly divide the dataset into 80% training set and the rest as test set. Choose the important features from the dataset by modifying relevant parts of the starter code. Choose a mini-batch size to divide the dataset into batches.
2. Build the ANN model.
    a. Build the MLP classifiers by identifying the number of input and output nodes required for the problem, and specifying the number of hidden layers as:
        i. 0 hidden layers
        ii. 1 hidden layer with 32 nodes
        iii. 1 hidden layer with 64 nodes
    b. Use Sigmoid or ReLU activation function for the input and hidden layers. Use ReLU activation for the output layer.
    c. Define the forward and backward operations for your network. They are required for inference and weight updation of your model.
    d. Define the training function to train the model using a forward and a backward pass. Define the prediction function for obtaining the outputs from the network.
    e. Compare the implementation of your model compared to that using the Pytorch library, on the same dataset (code snippet provided).
3. Hyper-parameter tuning.
    a. For each of the architectures, vary the learning rates in the order of 0.1, 0.01, 0.001, 0.0001, 0.00001. Plot graph for the results with respect to accuracy and loss. (Learning rate vs accuracy/loss for each model).
    b. Report test set accuracy for all the learning rates in a tabular form and identify the best model.
4. Classification Report
    a. Create a classification report for comparing the performance of your algorithm, for your best performing algorithm in terms of accuracy.

b. You need to calculate precision, recall, f1-score and accuracy of the model.

5. You can use any number of training epochs. Any additional analysis or findings from the dataset, as a result of your task, is well appreciated.

Note: *The program can be written in C / C++ / Java / Python programming language from scratch. No machine learning /data science /statistics package / library should be used for model creation.*

Relevant information:

Source: https://archive.ics.uci.edu/dataset/1/abalone

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Name / Data Type / Measurement Unit / Description

Sex / nominal / -- / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer / -- / +1.5 gives the age in years

Submission Details: (to be submitted in CSE-Moodle, **by one representative of the group**)

1. ZIPPED folder containing code (with comments) and the dataset files
2. Report (in pdf format)

Submission Guidelines:

1. You may use one of the following languages: C/C++/Java/Python.
2. Your Programs should run on a Linux Environment.
3. You are **not** allowed to use any library apart from these (Also explore all these libraries if doing in Python, or equivalent of these):

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import operator
from math import log
from collections import Counter
from statistics import mean
```

Your program should be standalone and should **not** use any *special purpose* library for Machine Learning. Numpy and Pandas may be used. And, you can use libraries for other purposes, such as generation and formatting of data.

0. You should submit the program file and README file and not the output/input file.
0. You should name your file as <GroupNo_ProjectCode.extension>
   (e.g., Group1_WONN.pdf or Group1_WONN.zip).
0. The submitted program file *should* have the following header comments:
   # Group Number
   # Roll Numbers : Names of members (listed line wise)
   # Project Number
   # Project Title
0. Submit through CSE-MOODLE only.

Link to course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=561

You should not use any code available on the Web. Submissions found to be plagiarized or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TA:
Sumanta Dey (Email: sumanta.dey@iitkgp.ac.in)