CSCI316 Big Data Mining Implementation and Techniques

Laboratory 7

The deliverables are assessed. This assessment component constitutes 5% of final marks.

This lab involves group work. You can chose the grouping. But group sizes are limited to 2-3 students (not smaller than 2 not larger than 3).

Please check the submission link on the Moodle site for the submission deadline. Late submissions result in 25% of mark deduction per day including weekends.

Task Specification

Dataset
The Abalone Dataset from https://archive.ics.uci.edu/ml/datasets/Abalone and available on Moodle.

Objective
In this task, you will implement feedforward networks (i.e., MLPs) in both plain TensorFlow and TensorFlow's Keras API. Your models are trained with records from the above Abalone Dataset and predict the target variable *numerically*. Thus, this problem is a ***regression*** problem.

Requirements

(1) The implementation task includes **two** MLPs. One MLP is a self-implemented model by using the plain TensorFlow. The other MLP is trained and evaluated by using the Keras API in TensorFlow.
(2) In your self-implemented model, you can (but not must) use the autodiff and optimizer (e.g., `GradientDescentOptimizer` and others) in TensorFlow.
(3) Your MLP models contain at least one hidden layer.
(4) The target variable is the number of rings of abalones. All other attributes are used to predict this target variable.
(5) Represent the nominal attribute(s) of the data by using one-hot encoding.
(6) Use the holdout evaluation method by randomly selecting (approximately) 70% records as training data and (approximately) 30% as test data.
(7) Use either Mean-Square-Error (MSE) or Cross-Entropy to evaluate the performance of your models.
(8) Use L2 the regularisation term in the model training.
(9) Clearly explain the your model architecture (e.g., how many neurons in each layer and the selection of activation functions at each layer) and implementation ideas.
(10) No other machine learning library (e.g., Scikit-Learn and Spark MLlib) is allowed in this task.

Deliverables

- ***One Jupyter Notebook source file (named lab7.ipybn***) which include the implementation of the task. This source file must also contain some clear English explanation of your implementation ideas and methodologies (in the markdown format and/or as comments of your Python codes).
- ***One PDF file (named lab7.pdf)*** which are generated from the two Jupyter Notebook source files. (You can use the Web browser's PDF printing function to create the PDF files.)

Ensure that the header of the your submitted files must contain the name and student numbers of all group members. Each group is expected to make just one submission.

Submission:

Submit your work via the correct submission link provided on the Moodle site.

Assessing criteria

The lab environment defines the standard environment of the tasks. Correctness of the implementation (*with respect to the requirements*) will be assessed. The Jupyter Notebook source files must be ***executed*** successfully. All results of your implementation must be produced from your submitted Jupyter notebook. Your submission must document clear and accurate explanation of your implementation and results. A poor presentation of results will gain less marks in this task.