COSC 425 — Introduction to Machine Learning — Fall 2019

Project 4: Back-Propagation

Due Nov 18, 2019

In this project you will apply multilayer artificial neural nets (ANNs) to spam email classification. The file is `spambase.data` in the Project 4 folder; the attributes are described in `spambase.names`, also in the folder. (For additional information, see https://archive.ics.uci.edu/ml/datasets/spambase.) Your task will be to classify sample instances as spam or not. As usual, your report should cover the relevant steps in the workflow, reporting what you did and the results. Implement your own ANN training and testing functions, except for basic numerical libraries.

1. Implement a function to use back-propagation to train an artificial neural network with a specified architecture (input dimension, output dimension, and specified number of neurons in each hidden layer). Your function should allow you to specify whether the output layer is linear, logistic sigmoid, or softmax. You should be able to specify a learning rate. You should be able to determine the length of training (e.g., stop after a specified number of epochs, error below a specified threshold, or change in error less than a threshold). For simplicity use online training.
2. During training, your function should display (print or plot) the error on the training and validation sets (so that you can stop training before overfitting).
3. Implement another function that applies a trained neural net to a dataset (e.g., validation or testing), produces a set of predictions, and displays the net's performance (confusion matrix, accuracy, and other metrics).
4. Experiment with various architectures and training regimens (learning rate, stopping condition) to see which gives you the best performance on the validation data. After you have determined the best hyperparameters, apply your trained net to the testing data.
5. **Extra credit**: Apply dimension reduction to see if you can improve the performance of your neural network.