Report

January 26, 2025

1 Homework 1 Report

1.1 Code Explanation

1. In my solution for this homework, I first pre-processed the data using PCA from sklearn. Following the guidelines, I used 32 components for the PCA and then fit a logistic regression model as my linear classifier and then MLP as my NN classifier both from sklearn. I used the layers sizes as mentioned in the guidelines and then got the accuracy scores of both models from the sklearn metrics. For the classifiers implemented in PyTorch, I implemented my solution as closely as possible from the guidelines by first converting the data into tensor datasets and then data loaders. My accuracy calculation function as well as train functions were done as per the guidelines. I used the same number of layers as well as the cross-entropy loss function as mentioned in the guidelines.
2. For external sources that were used was ChatGPT for any errors that I was not able to debug which was just an error on the count of classes for the final layer of the neural net.

1.2 Discussion

1. What is the shape and data type of each provided matrix?
   - The shape of the train features matrix is (15707, 426) and type is int16.
   - The shape of the train labels matrix is (15707,) and type is uint8.
   - The shape of the test features matrix is (1554, 426) and type is int16.
   - The shape of the test labels matrix is (1554,) and type is uint8.
2. What are the rows and columns of the matrices?
   - For the train and test features matrices, the columns are the values of the 426 spectral bands and the rows are the number of examples, which would be the number of trees in the dataset.
   - For the train and test labels matrices, the columns are the labels of the trees and the rows are the number of examples.
3. What are the ranges?
   - The range of the train features matrix is from 0 to 14998.
   - The range of the train labels matrix is from 0 to 7.
   - The range of the test features matrix is from 0 to 6908.
   - The range of the test labels matrix is from 0 to 7.
4. How many classes are there and what are the classes?
   - There are 8 classes in the dataset and they are seven dominant species as well as dead standing trees in a mixed-conifer forest in the Southern Sierra Nevada, California.
5. How many examples are provided of each class in the train and test splits?
   - The classes go 0-7.

- In the training set there are [2519, 821, 1575, 3980, 2640, 88, 852, 3232] examples.
- In the test set there are [389, 30, 278, 404, 100, 22, 43, 288] examples.

6. Compare the 2 PyTorch models in terms of test performance and overfitting.
   - For the test performance it seems that both the linear classifiers and the neural network classifiers have similar performance on generalizing to unseen data since both have 0.83 accuracy. It seems that the neural network fits the training data perfectly well since the training accuracy is 1.0 while the linear classifier has a training accuracy of 0.86. The linear classifier is less prone to overfitting while the neural network is more prone to overfitting since the gap between the accuracies is larger for the neural network.
7. Compare PyTorch results to scikit-learn results.
   - Both of these methods have nearly identical test accuracies since the linear classifier had a test accuracy of 0.83 while the neural network had a test accuracy of 0.84, which are both from using PyTorch. There could be a minor improvement in generalization from using PyTorch and the neural network classifier.