Image Classification with CIFAR-10 Dataset

Tim Lieb tlieb@wisc.edu

22-October-2020

Project Dataset

For this project, I will be using the CIFAR-10 dataset. This dataset has 60,000 color images that are 32x32 pixels in size. The images fit into 10 different classes, with 6,000 images per class. Of the total images 50,000 are training images are 10,000 are test images, and the training data is split into 5 training batches.

The classification problem explores assigning each image to one of classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. These classes are mutually exclusive, meaning each image belongs to a single class with no overlap.

Instead of using the data as 5 training sets with 1 test set, I will use all 6 sets for cross validation. Therefore, which sets are uses for training and which are held out for testing will change depending on the iteration. The error rate and the squared error will be used to compare the different iterations to each other.

Dataset: https://www.cs.toronto.edu/~kriz/cifar.html

Classification Algorithms

For the classification problem, I will compare 3 different algorithms:

- Ridge Regression- this uses the least squares classifier along with a regularization part to help balance bias and variance. One area of exploration will be in choosing the regularization parameter.
    - I will test a set of lambdas and use cross validation to pick the one that minimizes the error rate
    - I will perform the analysis using different testing and hold out sets to see if there is any difference between the different trials
- K-Nearest Neighbors Clustering- this iterative approach assigns each element to one of k clusters until convergence, where k=10 in this case. One parameter to explore is the initialization of the clusters to avoid converging to local minimum.
    - NOTE—the K-means initially proposed does not work for supervised learning. Check to see if this approach is alright
- Neural Networks- we have not yet covered neural networks in class, but the general idea is that the classification is done through the mapping of nodes. More information on this implementation to come.

Since the data is already split into training data and test data, I will use each subset for their respective roles. The test data is a good way to evaluate the utility of each algorithm. By using the same subsets for training and testing across each algorithm, the 3 can then be compared fairly by looking at the rates of correct classification for each method.

Github
https://github.com/tlieb21/ECE532_FinalProj

Timeline
DONE- Oct 22: Project Proposal
DONE- Nov 8: Data loaded, formatted and usable in python, Ridge Regression started
Started- Nov 17: Project Update #1- K-Neighbors started
Nov 24: Neural Networks lectures started
Dec 1: Project Update #2- Neural network implementation started
Dec 12: Final Project and Report
Dec 17: Peer feedback