CS 532 (Sec. 4) – Project Update 1

Alex Guo

Due Date: November 17, 2020

1 Progress

Link to repo: https://github.com/ag262/CS532.git

My dataset contains 15,000 images of chinese characters (the number of classes for the data is 15). I wrote "preprocess_data.py" to preprocess my data, where I centered, rescaled, and then applied Otsu thresholding to bring out the foreground of the images.

I also implemented my first algorithm, kNN, in "kNN.py" (see GitHub repo). I used 10,000 training data images and 5,000 testing data images. For parameters/decisions, I used $k = 500$ (which I decided based off of 5% of the number of training data) and the L2-norm for distance calculation (decided randomly). Without preprocessing, I got a misclassification percent error of $e = 93.1$; with preprocessing, I got $e = 48.9\%$. By random guessing, e should be $100 * (1 - 1/15) = 93.3\%$. My results make sense, since images are very unstructured and preprocessing helps to make them slightly more structured.

2 Plan ahead

I need to finish up the investigation of kNN by: 1) using a validation dataset to decide the best k, and 2) test the impact of using the L1-norm for distance calculation. I will then need to start coding up my other two algorithms (SVM and NN).

3 Project timeline (revised)

11/18 - 11/30: code up the other two algorithms

12/1 (milestone): second update due

12/2 - 12/11: do last-minute corrections/testing and write up the final report

12/12 (milestone): final report due