Code manual

Table of content:

Method interpretation

1.Dataset

The mnist dataset was downloaded through a python deep learning API: keras. The dataset consists of 60000 training samples with labels and 10000 testing samples with labels.

The question ask for the regression on digit 0 and digit 7, so I extracted the samples with label “0” and label “7” from all samples. As a result, 12188 samples in training set and 2008 samples in testing set are selected from mnist. As Figure1 shows, I restored some samples to images, and displayed the labels and images together.

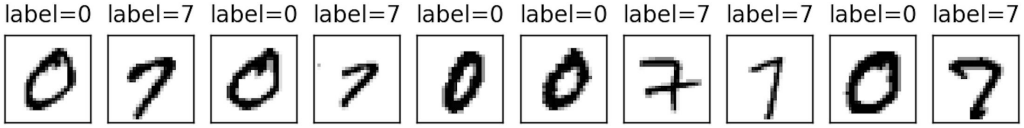

Figure 1: dataset visualization

2.data preprocess

The raw data are 28x28 pixel matrix. I reshaped it into 1x784 one-dimensional vector. After that I divided the training set into training set and validation set. The training set was used to train the model, the validation set was used to choose the best model.

Then, I defined a function which can randomly select 100 images from dataset to form a bag. The function also calculate the percentage of digit 0 and digit 7 in each bag as the bag-level label(similar to the genomic tumor purity in the given paper). The bags generated by this function will be used in the steps of model training and testing.

3. Model

I used the MIL model which was proposed by the given paper to predict the fraction of each bag. The model consists of feather extractor(18-layer residual network), MIL pooling filter(distribution pooling filter) and bag-level representation transformation module(a 3-layer neural network).

The code of model definition, training and testing are **mostly cited** from the given paper.

The link of original code: https://github.com/onermustafaumit/SRTPMs

4. Result

Due to time constraints, I have not completed the integration of the code, so the prediction results and corresponding evaluations are not given here.