STAT3612 Assignment 3: Handwritten Digit Classification with Keras

Date: April 24, 2018

Sumbit via Moodle on or before May 6, 2018.

The MNIST dataset consists of 70K small images (60K for training, and 10K for testing) of handwritten digits 0~9. It can be loaded from R:keras and used as an example of computer vision study, as we have discussed in Lecture 12. In this assignment, let us consider a reduced problem by picking the handwritten digits 4 and 9 for binary classification tasks.

Step 1.(20%) Load the MNIST dataset and select the subset samples corresponding to digits 4 and 9. Find the sample sizes for each digit in both training and testing data. For each digit, display 5 random images from the testing data.

Step 2.(40%) Reshape each image to a long vector and normalize the pixel values to [0,1]. Convert the labels to be 0 (for 4) and 1 (for 9), and perform one hot encoding for the binary classes. Then, we may perform logistic regression modeling by using `keras_model_sequential` with **only an output layer** and softmax activation. Fit the model with 20 epochs. Plot the history of model fitting. Evaluate the model with testing data, report the prediction accuracy, plot the ROC curve and calculate the AUC score.

Step 3.(40%) Add a hidden layer to the logistic regression model above with 100 nodes of ReLU activation type. Fit the model with 20 epochs. Plot the history of model fitting. Evaluate the model with testing data, report the prediction accuracy, plot the ROC curve and calculate the AUC score.

Step 4. (Bonus 20%) Given the above sub-MNIST dataset with binary labels and image pixels, if we use `glm` type of logistic regression for making classifcation, it would be very time

consuming and non-stable if we directly use all pixel values as covariates (i.e. independent variables, or predictors) like above. Can you come up with a reasonable approach for such image-based classifcation problem based on logistic regression? (Hint: feature extraction and selection).