Binary Logistic Regression

Dataset Description:

The CSV file contains 5172 rows, each row for each email. There are 3002 columns. The first column indicates the Email name. The name has been set with numbers, not recipients', to protect privacy. The last column has the labels for prediction: 1 for spam, and 0 for not spam. The remaining 3000 columns are the 3000 most common words in all the emails, *after excluding the non-alphabetical characters/words*.

Dataset Preparation:

1. Select proper columns.
2. Normalize the dataset.
3. Randomly split the dataset into TRAINING(80%) and TEST(20%) sets.

Train (update Θ):

1. for each sample, $X = [x_1, x_2, ..., x_n]$ in TRAINING set:

   Concatenate 1 and turn it into $X' = [x_1, x_2, ..., x_n, 1]$
2. randomly initialize $\theta = [\theta_1, \theta_2, ..., \theta_{n+1}]$ within 0 to 1 // $\theta_1, \theta_2, ..., \theta_n$:weights, θ_{n+1}:bias
3. max_iter = 500, lr = 0.01
4. history = list()
5. for itr in [1, max_iter]:

   $J = 0$ //total cost

   for each sample, X' in TRAINING set:

   $z = X'.\theta$ //use np.dot function

   $h = sigmoid(z)$ //sigmoid available in Python

   $L = -ylog(h) - (1 - y)log(h)$ //h = prediction label, y = true label

   $J \mathrel{+}= L$

   $dw = X'.(h - y)$ //$dim(dw) = n + 1$

   $\theta = \theta - dw * lr$ //$dim(\theta) = n + 1$

   J /= N_train //N_train = # of training samples

   append J into history

Validation:

1. correct = 0
2. For each sample in X' in the TEST set:

   $z = X'.\theta$

   $h = sigmoid(z)$

   if h >= 0.5: h=1

   else: h=0

   if h == y: correct += 1
3. *test_acc = correct*100/N_test* //N_test = # of testing samples

Instruction:

- Plot total cost vs iteration(History) for different learning rate, *lr* = 0.1, 0.01, 0.001 and 0.0001(*max_iter* = 500)
- Make a table with 2 columns: learning rate lr and *test_acc.*
- Submit a .ipynb file and a report .pdf file.
- Do not use libraries such as: SKlearn, Scikit learning for this assignment.