ENGR 421/DASC 521: Introduction to Machine Learning
Homework 2: Discrimination by Regression
Deadline: November 7, 2022, 11:59 PM

In this homework, you will implement a discrimination by regression algorithm for multiclass classification using Python. Here are the steps you need to follow:

1. Your discrimination by regression algorithm will be developed using the following modifications to the linear discrimination algorithm with the softmax function we discussed in the lectures.

    a. Instead of the softmax function, you are going to use K sigmoid functions to $\hat{y}_{ic}$ generate values. Please note that, in such a case, the summation of $\hat{y}_{ic}$ values is not guaranteed to be 1. However, you are going to pick the largest value to predict the class label.

    b. Instead of minimizing the negative log-likelihood (i.e., $-\sum_{i=1}^{N}\sum_{c=1}^{K} y_{ic}\log(\hat{y}_{ic})$), you are going to use the sum squared errors as the error function to minimize (i.e., $0.5\sum_{i=1}^{N}\sum_{c=1}^{K}(y_{ic}-\hat{y}_{ic})^2$). Please note that you need to find the correct update equations for this modified model.

2. You are given a multivariate classification data set, which contains 15000 clothing images of size 28 pixels $\times$ 28 pixels (i.e., 784 pixels). These images are from ten distinct classes, namely, t-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. The figure below shows five sample figures from each class. You are given two data files:

    a. `hw02_data_points.csv`: clothing images,

    b. `hw02_class_labels.csv`: corresponding image labels (1: t-shirt/top, 2: trouser, 3: pullover, 4: dress, 5: coat, 6: sandal, 7: shirt, 8: sneaker, 9: bag, and 10: ankle boot).

3. Divide the data set into two parts by assigning the first 10000 images to the training set and the remaining 5000 images to the test set. (10 points)

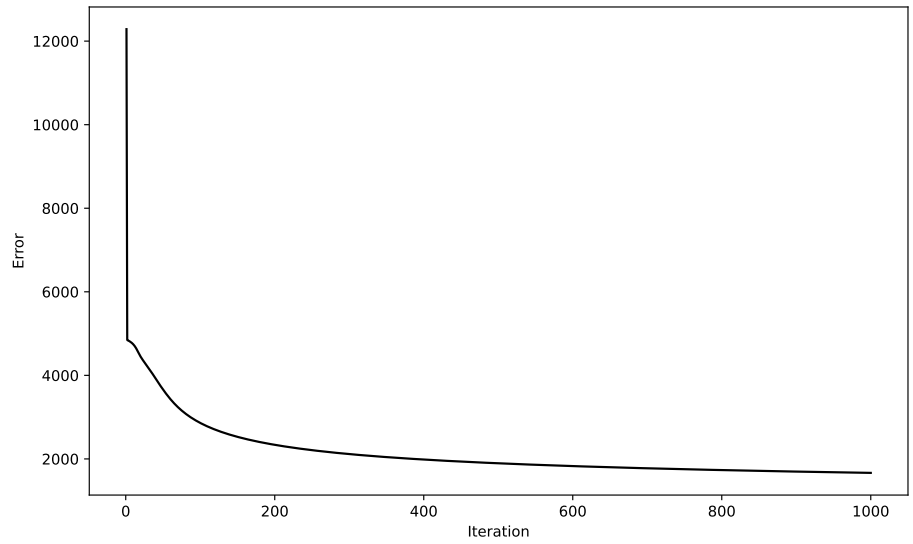
4. Learn a discrimination by regression algorithm for this multiclass classification problem using the following learning parameters.

```
eta = 0.00001
iteration_count = 1000
```

You should initialize the weight parameters to the values given in `hw02_W_initial.csv` and `hw02_w0_initial.csv` files before running your algorithm. Your parameter estimations should be like the following figures. (40 points)

```
print(W)
[[-0.01324798 -0.02823844 -0.00326868 ... -0.04877851  0.01212054
  -0.01143465]
 [-0.01183842 -0.03963016 -0.01295336 ... -0.04034705  0.00400381
  -0.02111644]
 [-0.01998825 -0.03633561 -0.00489235 ... -0.04108662  0.01387463
  -0.02484677]
 ...
 [-0.01341638 -0.0199712  -0.02071762 ... -0.03364511  0.00594223
  -0.01845717]
 [-0.00980747 -0.03218592 -0.02022552 ... -0.03211477  0.01611503
  -0.00845905]
 [-0.01977218 -0.02373074 -0.01468591 ... -0.04159601  0.01068509
  -0.02400039]]
print(w0)
[[-0.01287857 -0.02891159 -0.00873806 -0.03535891 -0.02597275 -0.06542254
  -0.01501564 -0.0451543   0.00689065 -0.01964791]]
```

5. Draw the objective function values throughout the iterations. Your figure should be like the following figure. (10 points)



6. Calculate the confusion matrix for the data points in your training set using the discrimination by regression algorithm you will develop using the estimated parameters. Your confusion matrix should be like the following matrix. (20 points)

```
print(confusion_train)
y_truth   1    2    3    4    5    6    7    8    9    10
y_pred
1        838    3   14   44    5    0  220    0    1    0
2          4  908    1   17    4    0    3    0    2    0
3         14   12  645   14   69    0  133    0    9    1
4         89   37    5  870   41    2   55    0   12    2
5          2    6  172   28  763    0  123    0    6    0
6         16    3   26    5    8  841   30   89   25   33
7         27    3   80   28   79    0  423    0   17    0
8          0    0    0    0    0  115    1  862    8   43
9         18    0    8    5   10   13   34    2  892    0
10         1    0    0    1    0   37    0   84    2  957
```

7. Calculate the confusion matrix for the data points in your test set using the discrimination by regression algorithm you will develop using the estimated parameters. Your confusion matrix should be like the following matrix. (20 points)

```
print(confusion_test)
y_truth   1    2    3    4    5    6    7    8    9    10
y_pred
1        397    1   11   22    0    1  129    0    0    0
2          4  459    0    8    3    0    1    0    0    0
3          7   10  320    6   49    1   52    0    5    0
4         51   15    3  443   19    1   31    0   10    0
5          2    2   83   14  382    0   58    0    1    0
6          9    2   13    2    4  405   12   39   14   18
7         14    1   50   11   47    0  218    0    9    0
8          1    0    0    0    0   47    0  387    7   27
9         16    0   10    1    2    7   18    1  459    0
10         1    0    0    0    0   28    0   44    3  472
```

What to submit: You need to submit your source code in a single file (`.py` file) named as `STUDENTID.py`, where `STUDENTID` should be replaced with your 7-digit student number.

How to submit: Submit the file you created to Blackboard. Please follow the exact style mentioned and do not send a file named as `STUDENTID.py`. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.
