CSCI4390-6390 Assign4

Assign4: Logistic Regression

Due Date: Oct 16, before midnight (11:59:59PM, Alofi Time; GMT-11)

You will use the Appliances energy prediction data set. You should ignore the first attribute, which is a date-time variable, and you should also remove the last attribute, which is a duplicate of the previous one. For Logistic Regression, use the first attribute (after removing the date-time variable), which denotes the **Appliances Energy Use**, as the response variables, with the remaining attributes as predictor variables. However, you have to discretize the response variable differently for the two sections as noted below, since CSCI4390 will implement binary logistic regression, whereas CSCI6390 will implement multiclass logistic regression.

CSCI4390: Binary Logistic Regression

You will implement the binary logistic regression algorithm as described in Algorithm 24.1 (Chapter 24, page 628).

Note that the **Appliances Energy Use** attribute takes values in the range $[10, 1080]$. However, for binary regression, we need only two values, so for the purpose of this assignment you should consider energy use less than or equal to 50 as the positive class (1), and energy use higher than 50 as negative class (0). You need to do this conversion to create the binary response variable, before you select the train (70%) and test (30%) subsets.

CSCI6390: Multiclass Logistic Regression

You will implement the binary logistic regression algorithm as described in Algorithm 24.2 (Chapter 24, page 634).

Note that the **Appliances Energy Use** attribute takes values in the range $[10, 1080]$. However, for multiclass regression, we will convert these into four classes as follows: energy use less than or equal to 30 is class c_1, energy use greater than 30 but less than or equal to 50 is class c_2, energy use greater than 50 but less than or equal to 100 is class c_3, and finally energy use higher than 100 is class c_4. You need to do this conversion to create the categorical response variable, before you select the train (70%) and test (30%) subsets.

What to submit

- Write a scripy named as **Assign4.py**, which will be run as **Assign4.py FILENAME ETA EPS MAXITER**. FILENAME is the datafile name, ETA is the step size η, EPS is the convergence threshold ϵ, and MAXITER is the upper bound on the number of iterations when learning the weights (i.e., terminate after MAXITER even if the EPS threshold has not been reached).

Your script should print out the weight vector(s), and also the final accuracy value on the test data, where you use 70% of the data for training and 30% for testing.

Note that accuracy is defined as the fraction of correct class label predictions. So for each test point, you should predict its class as the one that has the highest probability, and then you should count how many test points are correctly predicted. Divide that number of the test data size to get the accuracy.

You should try different η and ϵ values and report the best results.

- Submit a PDF file named Assign4.pdf that should include your answers to each of the questions (just cut and paste the output from python). **Failure the submit the PDF will result in lost points.**

- Submit the scripts and pdf file via submitty

Policy on Academic Honesty

You are free to discuss how to tackle the assignment, but all coding must be your own. Please do not copy or modify code from anyone else, including code on the web. Any students caught violating the academic honesty principle will get an automatic F grade on the course and will be referred to the dean of students for disciplinary action.