CISC 484/684
Machine Learning
Homework 1

Due Date: Friday March 13th, 2020

Q1 (Individual Work): Assume that I am able to find the odds of success (Y=1) for a given X (specified by air temperature, humidity, water temperature and chance of precipitation) as predicted by a logistic regression model. For a given instance, suppose the odds of success is 2. The only information I know about this instance is that the water temperature (one of the attributes) was 20 centigrade. I was also told that the odds of success assigned by the same logistic regression model is 4 for another instance. The only information I have about the new instance is that the water temperature was 21 centigrade and all other attributes in the two instances were identical.

What can you say about the value of the weight associated with the "water temperature" attribute.

Q2 (Individual Work). Consider the following two Boolean operators, F1 and F2 (inputs are assumed to be 0 or 1 only). Indicate for each of them whether a perceptron can be built to capture this function. If so, give the weights of a perceptron and if not, indicate why this might not be possible.

F1(x1,x2) is 1 iff x1=1 and x2=0
F2(x1,x2,x3) given by the following table

x1	x2	x3	F2 output
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Q3 (Group Work) Program the perceptron training algorithm using "perceptron update" rule. Use the stochastic method for updating (i.e., update after each instance). Assume the target value is one when $7.x_1 + 3.x_2 – 5.x_3 >0$. Assume an initial value of 0.1 for w_1, w_2 and w_3 and the threshold is 5 initially. What is the training error initially and what is the training error after each instance. The training data has two instances: first instance <1,2,3> and second instance is <3,2,1>. Your program should determine the error (sum of square error) on the training data.

Q4 (Group Work) This involves using the Pima Indian dataset (all the data can be found in files with extension .pkl) and the logistic regression program on scikit-l(see README-HW1-Q4).

a. Train on train_1 file and test on test_1 file. Which attribute/feature seems to have the greatest impact on the prediction?
b. Now train of train_2 file and test on test_2 file. This time, we have added a new feature and retained the others with the same values. Observe the accuracies of the model remain the same as in part a. Examine the data and the model parameters. What do you observe about the first (the new feature) feature's weight and that of others. Explain what you observe and why you obtain the same accuracy.
c. Repeat a similar exercise as (b) on train_3 and test_3. These two sets have a new feature added to data from train_1 and test_1. (We are no longer concerned about the new feature in (b) and train_2 and test_2 datasets.