CMSC726: Assignment #2 — Logistic Regression

Soheil Behnezhad

September 15, 2017

1. Generally, if the step size is too high or too low, the accuracy goes down. The initial step size of 0.1 seems to be a good choice.

2. Overall, increasing the number of passes increases the accuracy over both the test data and the training set. The following is the output for different number of passes:

(1 Pass) Update 1001 TP -172.011640 HP -36.470991 TA 0.946429 HA 0.887218

(5 Passes) Update 5001 TP -39.564084 HP -19.136738 TA 0.999060 HA 0.939850

(10 Passes) Update 10001 TP -24.141017 HP -18.779809 TA 0.999060 HA 0.947368

(15 Passes) Update 15001 TP -17.565676 HP -19.431405 TA 1.000000 HA 0.947368

(20 Passes) Update 20001 TP -13.782915 HP -19.250564 TA 1.000000 HA 0.939850

(25 Passes) Update 25006 TP -11.489173 HP -19.642891 TA 1.000000 HA 0.939850

(30 Passes) Update 30001 TP -9.844607 HP -20.098815 TA 1.000000 HA 0.939850

(35 Passes) Update 35001 TP -8.628585 HP -20.261032 TA 1.000000 HA 0.939850

(40 Passes) Update 40001 TP -7.687112 HP -20.531101 TA 1.000000 HA 0.939850

(45 Passes) Update 45001 TP -6.949706 HP -20.883358 TA 1.000000 HA 0.932331

(50 Passes) Update 50001 TP -6.338188 HP -21.043273 TA 1.000000 HA 0.932331

(55 Passes) Update 55001 TP -5.834498 HP -21.328470 TA 1.000000 HA 0.932331

Roughly after 10 passes HP converges and after 50 passes TP's change becomes very slow.

3. Words with higher (resp. lower) biases are the better predicators for class 1 (resp. 0). Therefore by sorting the words based on their bias we can determine the best predicators for the classes.

   The first 5 words with the lowest bias are as follows: [(-1.8519, 'hockey'), (-1.2709, 'playoffs'), (-0.9959, 'pick'), (-0.9045, 'playoff'), (-0.9012, 'points')]

   The first 5 words with the highest bias are as follows: [(1.2073, 'hit'), (1.1462, 'runs'), (0.9299, 'bat'), (0.9003, 'saves'), (0.8407, 'pitching')]

4. Based on the logistic regression's formula, the closer the bias of a feature is to zero, the poorest it is in predicating the class. The 5 words that have the closest bias to 0 are as follows: [(0.0, 'everywhere'), (0.0, 'blasted'), (0.0, 'intermissions'), (0.0, 'bloody'), (0.0, 'broad')]