**Machine Learning PS1: Perceptron**

Template Write-Up

**NOTES:**

**X** = E-mail threshold for a word to be considered a feature (word has to show up in X e-mails)

**N** = The size of your training set (it’s how many e-mails in *spam\_train.txt* you’ll use to train)

You should start with **X = 20** and **N = 4000**. Some questions ask you to vary these numbers and then explain/show how your program’s results change.

Explain in a couple of sentences why measuring the performance of your perceptron classifier would be difficult if you do not create a validation set (size = **5000 – N**) from your training data.

|  |
| --- |
|  |

How many passes of the data did your implementation make before it converged? How many total mistakes did the algorithm make before convergence?

|  |
| --- |
|  |

What is your error rate with the validation set? (You just need to enter a percentage)

|  |
| --- |
|  |

Which 12 words are most correlated with spam? That is, which 12 “features” have the most *positive* weights? Provide a screenshot or copy/paste of your program output.

|  |
| --- |
|  |

Which 12 words are least correlated with spam? That is, which 12 “features” have the most *negative* weights? Provide a screenshot or copy/paste of your program output.

|  |
| --- |
|  |

Vary N = 100, 200, 400, 800, 2000, 4000. Provide a plot of the validation error percentage as a function of N.

|  |
| --- |
|  |

Vary N = 100, 200, 400, 800, 2000, 4000. Provide a plot of the number of perceptron algorithm passes as a function of N.

|  |
| --- |
|  |

Keep N=4000 constant, and vary the value of X. Try X = 30, X = 40, and any other values of X you’d like. What do you think is the optimal value of X for your perceptron configuration? (Optimal here means lowest validation set error percentage)

|  |
| --- |
|  |

Now use your best configuration on all of *spam\_train.txt* (N=5000, X=whatever you found to work well). Train with it, and report your error percentage on *spam\_test.txt*

|  |
| --- |
|  |

Try setting X=1200. How many features do you have in your model? What happens when the perceptron runs? Can you explain what’s going on, and why it happens?

|  |
| --- |
|  |

Why do we need a training set, validation set, and test set? One or two sentences on the purpose of each is fine.

|  |
| --- |
|  |