Miles Clark
February 5, 2015

Machine Learning Problem Set 1

1. When determining the optimal parameter of which to train our classifier, it is necessary to separate the training data into this validation set or else we risk over fitting the data. If we are tuning our parameters by testing our classifier against only the training data, then we can expect that classifier does very well on the training data, but we have no idea how it performs on unseen data, that it did not use to train. In order to get an idea of how the classifier performs on unseen, we separate the data into a validation set, and tune the training parameters used on the training by analyzing how different parameters perform on the validation set. This allows us to see how the classifier performs on unseen data, and gives a better estimate to how it will perform on the test data.

2. This transformation is done in my code through creating a class called “BinaryVectorizor”. This class creates and holds the vocabulary and also provides methods for transforming entire document datasets, as well as single document string examples into feature vectors based on the vocabulary.

3. In my code, I create a class called Perceptron that hold the training data, it's corresponding labels, and a weight vector (initialized to an empty array). The class provides methods for training called “perceptron_train” and testing called “perceptron_test”.

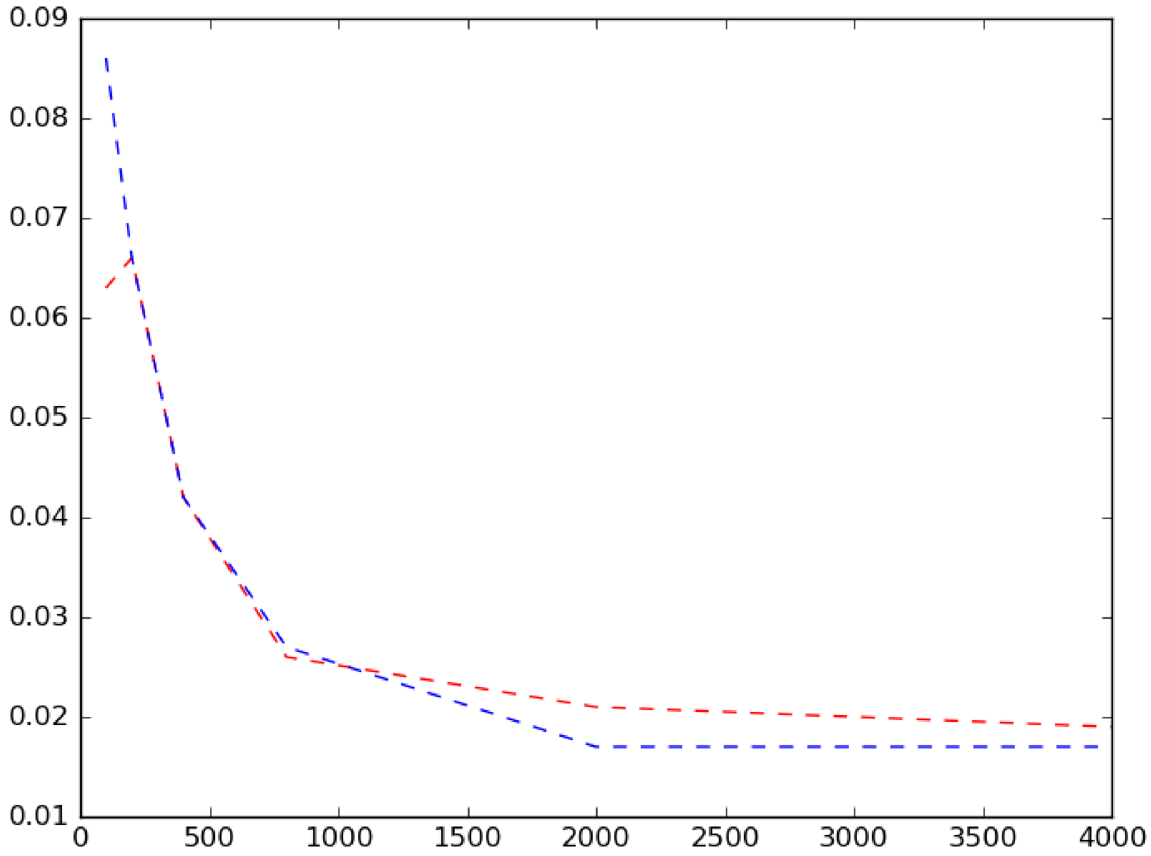
4. In training the perceptron_algorithm with the training data, it makes **430** mistakes before terminating, as displayed in the program. It also makes **11** passes through the data. I confirm that the training convergence condition was satisfied by running “perceptron_test” on the training data and displaying an error of 0. After vectorizing the validation dataset and testing the trained classifier on the validation set, I display that the validation error is **0.01900**.

5. The 15 most positive words in the vocabulary are:
sight, our, click, market, dollarnumb, remov, deathtospamdeathtospamdeathtospam, most, pleas, yourself, below, present, guarante, your, longer

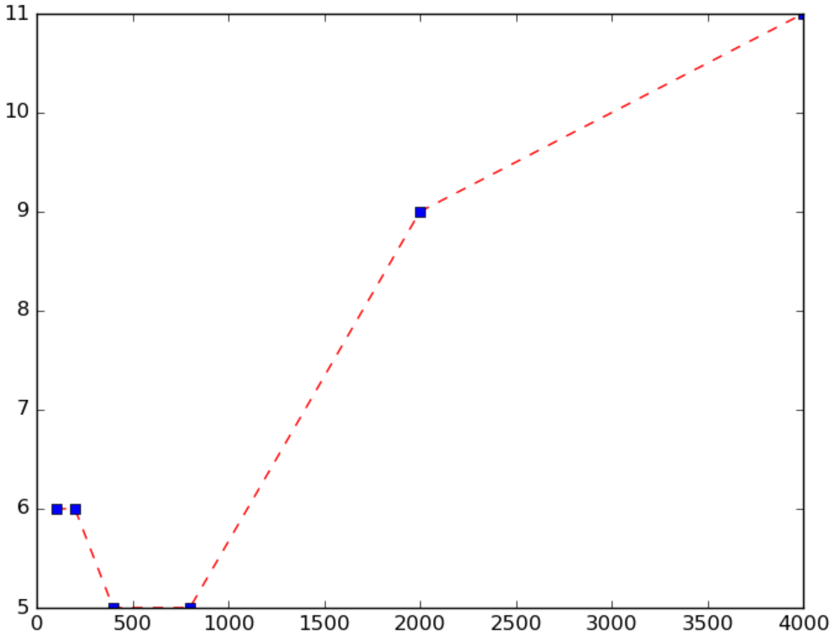
The 15 most negative words in the vocabulary are:
wrote, prefer, set, I, re, too, technolog, digit, post, support, until, team, still, recipi, develop

6. The Average Perceptron algorithm is implemented with an optional parameter in the previously stated perceptron_train method. Throughout my perceptron_train method, I increment a variable called w_avg with every attempted weight vector. Because vector addition and scalar multiplication is linear, if the “average” parameter is set to “True”, I return this w_avg vector divided by the total number of passes through the data times the number of examples in my dataset. That product would represent the total number of weight vectors attempted and the final returned vector would then be the average of all attempted weight vectors.

7. In the following plot, the red dashed line represents the Perceptron Algorithm, and the blue dashed line represents the Average Perceptron Algorithm, both plotting validation error as a function of the number of examples used in training:



8. In the figure below, the red dashed line represents the Perceptron Algorithm and the blue squares represent the Average Perceptron Algorithm, both plotting the number of iterations needed to terminate as a function of the number of samples used in the training:



9. The max_iter option is added to my code as a parameter, with a default value set as “None”. If this parameter is left as “None,” then max_iter is set to the largest integer value in Python. Otherwise, when the max_iter is set to N, the algorithm terminates either when the weight vector correctly classifies all of the examples or the data has been passed over N times (which ever comes first).

10. After training several configuration of the Perceptron Algorithm, and testing on the validation set, the configurations with the lowest validation error, that is, 0.017 were the Perceptron Average Algorithm without a maximum iteration limit given, the Perceptron Average Algorithm with a maximum iteration limit set to 5, and the Perceptron Average Algorithm with the maximum iteration limit set to 10. Because they all gave the same validation error, the first was used. This trained classifier gave test error of 0.019 on the “spam_test.txt” dataset.