Logistic Regression Analysis, CSCI 5622 Homework 2

Alex Gendreau

1. **What is the role of the learning rate?**

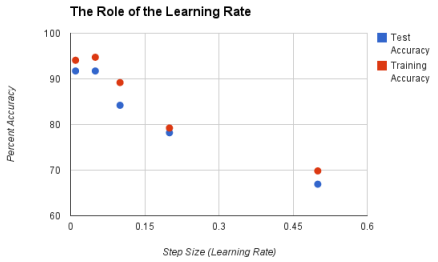

Figure 1: The percent accuracy of both the training and test sets based on varying the step size (the learning rate) for regularized logistic regression, $\mu = 0.01$.

2. **How many passes over the data do you need to complete?**

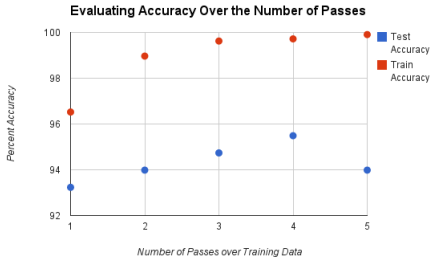

Figure 2: The percent accuracy of both the training and test sets based on varying the number of passes over the training set.

We conducted an experiment using the default settings (i.e. $\mu = 0$ and a learning rate of 0.1) while varying the number of passes and examining the accuracy of both the training set and the test set. In Figure 1, we see that the while the training set increase in accuracy for all passes (as expected since we are running multiple passes on the training set), the accuracy of the test set increases until the fourth pass but actually decrease on the fifth pass. One reason for this is that we are overfitting the training set with five passes. From the experiment, I would suggest 4 passes over the training set on the default settings for generating the best logistic function.

3. **What words are the best predictor of each class?**
Using the default settings, I found the best predictors for **baseball were: pitching, saves, bat, runs, and hit**. The best predictors for **hockey were: hockey, playoffs, pick, playoff, and points**. To find these values I looked at the features that had the largest coefficients (betas) for baseball and the smallest coefficients (betas) for hockey. I chose to use this as my metric for best features because features with large coefficients (either positive or negative) give greater weight to the logistic function. I accessed these words using the argsort function in python which returns a list of the indices of the of an array in sorted order. These indices correspond to feature numbers, which then allowed me to access the features.

4. **What words are the poorest predictors of classes?**
   Using the default settings, I found the worst predictors were: **racist, bloody, blasted, hooked, and intermissions**. To find these values I looked the feature coefficients with the smallest absolute value. These features would contribute least the logistic function and thus the classification process. I used a similar implementation to access the smallest in absolute value coefficients by first taking the absolute value of all the feature coefficients (betas).