I 526/B659 Programming Assignment 2 - Due Tuesday October 14, 2014

Data set information: Recall that this data set is extracted from the UCI zoo data set (zoo-train and zoo-test). Note that there are 16 features (the first 16 columns) and the class labels are in the last column. There are 7 classes (numerically specified as class 1 to 7). All features are binary except for feature 13, which is a categorical variable with possible values 0,2,4,5,6,8. Note that to create binary split, please use the one-vs-rest approach.

1. Implement the logistic regression algorithm. Present the accuracy on the test set. Vary the learning rate (η) and report the results for 3 different learning rates. Report the confusion matrix on the test set. Since this requires a k-class logistic regression, let us simplify further. Your goal is to predict if the class is 1 or not. To achieve this, you have to create a new version of the data set where for all the examples where the class label is not 1, you assign a new class label (say 0). Thus now your binary task is predicting whether class 1 is true or not. Remember to do the same for the test set as well.

2. Now implement the counting based learning of Naive Bayes classifier. You will assume Laplacian correction. Again, treat the task as binary and present the results as a confusion matrix.

3. Why do you think one method does better than the other, speculate.

4. Run the modified data set through Weka for Logistic Regression and Naive Bayes. Present the results. How are they different from yours? Briefly explain.

The total points on this homework is 20.