# Assignment 3

Garcia, Jorge A.
*Department of Physics*
*New Mexico State University*
(Dated: October 25 2019)

**Course:** C S 508

**Instructor:** Dr. Tuan Le

## I. K-NN CLASSIFIER AND 5-FOLD CROSS VALIDATION

The script "prob1_knn_kfold.py" loads the "iris.csv" data set, shuffles and splits it into 5 folds for model cross validation. K-Nearest Neighbors is trained and tested on all 5 fold of the data set, and repeated using 1 through 50 k-Neighbors for classifying. The resulting average accuracy for both training and testing sets as the number of neighbors is varied is shown in Figure 1. The overall trend indicates that a larger number of neighbors being considered results in a worse model. The model with the highest average testing accuracy and smallest standard deviation occurs when considering $k = 11$ neighbors, resulting in the most optimal model for this data set.

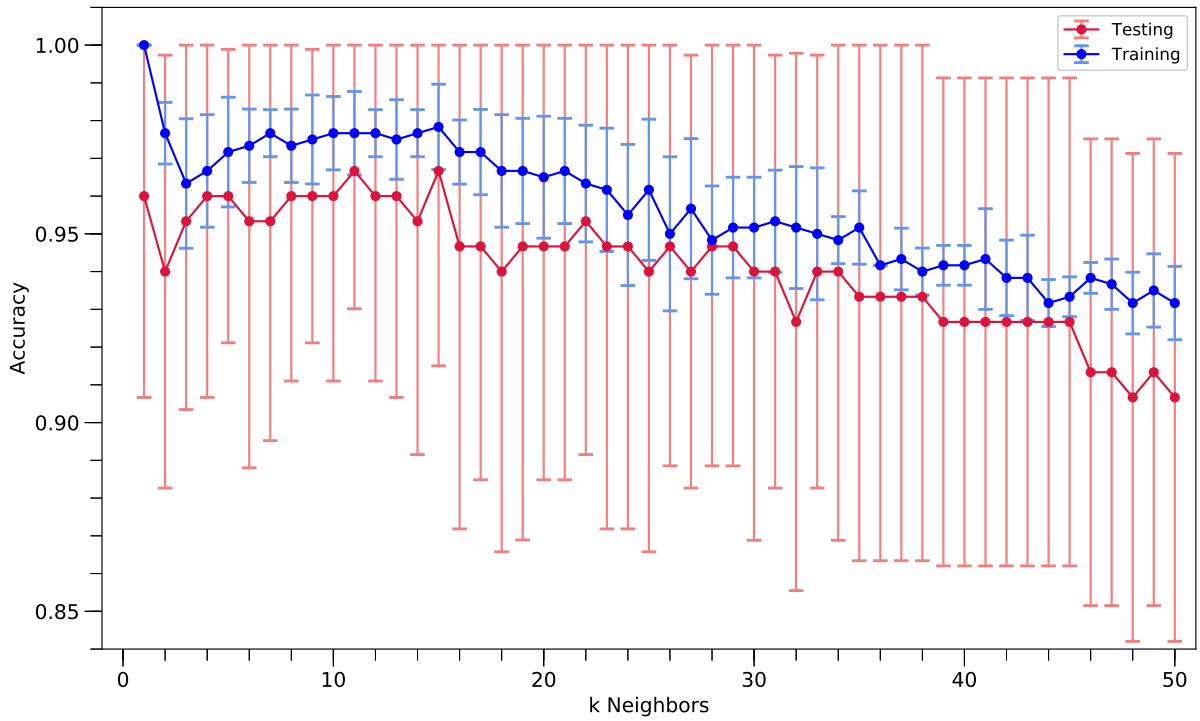


FIG. 1. Average model accuracy and its standard deviation as k-neighbors are considered.

## II. ROC CURVE

The script "prob2_logreg_roc.py" loads the "iris.csv" data set, remaps the classes to Setosa (1) and Non-Setosa (10), shuffles and splits the data into 80% training and 20% testing sets. Logistic Regression is used to fit a model, and the class probabilities are predicted. Using the probabilities for Setosa class, the ROC curve is evaluated and plotted in Figure 2. The curve shown indicates that the model perfectly predicted the testing data, which at first I was skeptical about. Appendix A shows the scatter plot matrix for the data set, and the attributes are all linearly separable from each other. Thus, the perfect performance obtained from Logistic Regression is reasonable.

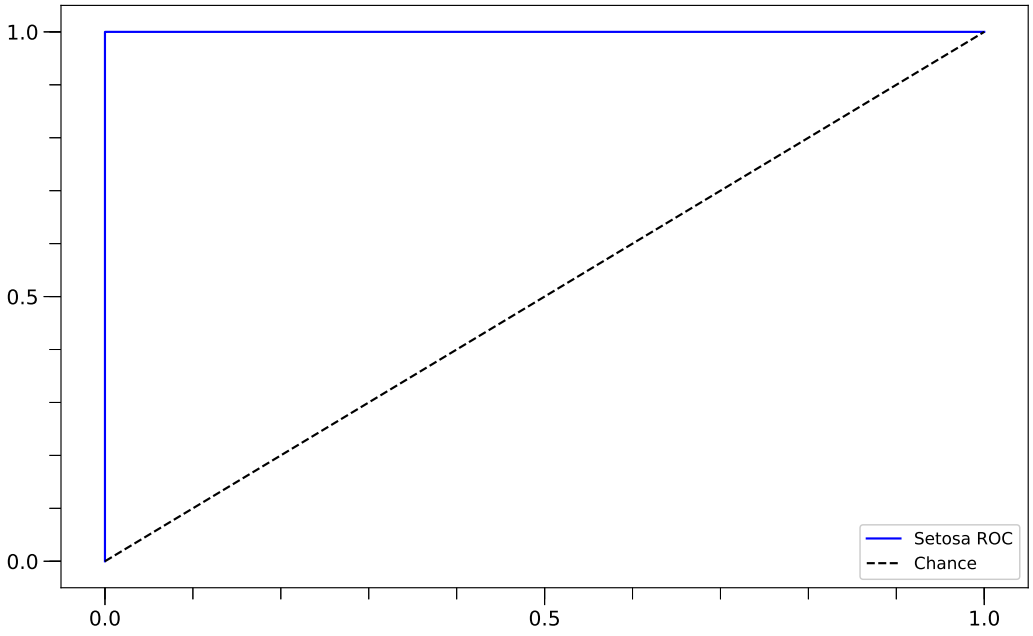


FIG. 2. ROC curve for prediction of type Setosa.

**Appendix A: Setosa/Non-setosa Scatter Plot Matrix**