Travis Thein
Project 1
Tuesday, October 15, 2019
Worked with Hope Fowler

A. README:

- To run code, download and import the attached functions into MatLab. Run the "Proj1.m"to run the code. The output should depict a table with the iteration, evaluation result, objective, runtime, best-observed instance, best-estimated instance, and the number of neighbors.
- Best observed feasible point is offered.
- Decision tree and knn model info will appear
- Updates with the script
- For the perceptron model, the code outputs the accuracy and classification error. In the end, the classification error pertinent to each of the models is printed; though, because all the errors are relatively low they print out as 0.

B. Technical Report:

1. Description of data and research question.

The data is a CSV file with a list of counterfeit and real currency data. The currency(data) is characterized by six features: length, left width, right width, bottom margin, top margin, and diagonal length with each example also being labeled as "real" or "fake." The goal here is to analyze which model fits the data the best, essentially, determine which model can best identify the counterfeit currency. The examples below compare decision trees (dt), k-nearest neighbors (knn), and perceptron (pt) models.

2. Feature selection rationale, together with an explanation of results from any techniques such as cross-validation.

This code compares model error rates with and without each feature to dictate feature selection. By comparing model error rates with and without each feature, this code analyzes data for correlated features through linear regression on each pair of features. Overall, correlated features do not improve models. For example, linear regression models create a much larger variance between several numerically unstable solutions. Useless features are must be deleted in addition to a cross-validation process run on the independent set. Analyzing the effectiveness of machine learning models in the end will create more

dependable results. If data is limited, a resampling procedure will use the results of statistical analysis to generalize to an independent data set.

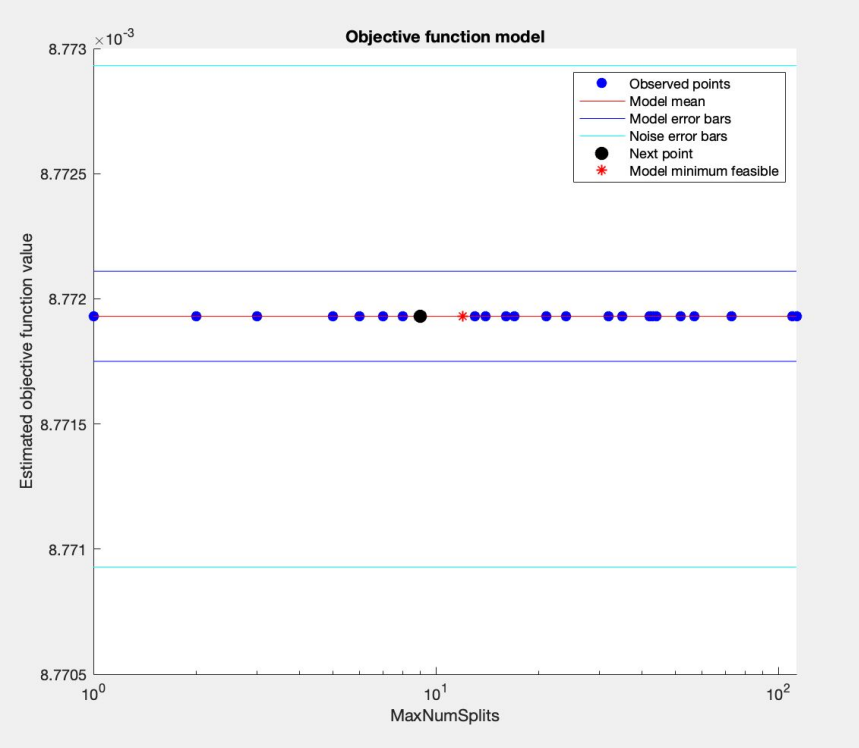
3. Description of classification techniques, the model choices you made (such as which k, decision tree purity measures, etc.), and why you made those choices.

When choosing features between models, depth was used for decision trees, k-values for k-nearest neighbors, and the iteration number for perceptron. The error was calculated for each parameter option(training and development), then stored in an array, from which the best error located. Loss functions were used to determine error for the decision tree and knn models, while the perceptron uses mean absolute error.

4. Results of evaluation on your test set of each classification technique (plots and graphs may be useful)

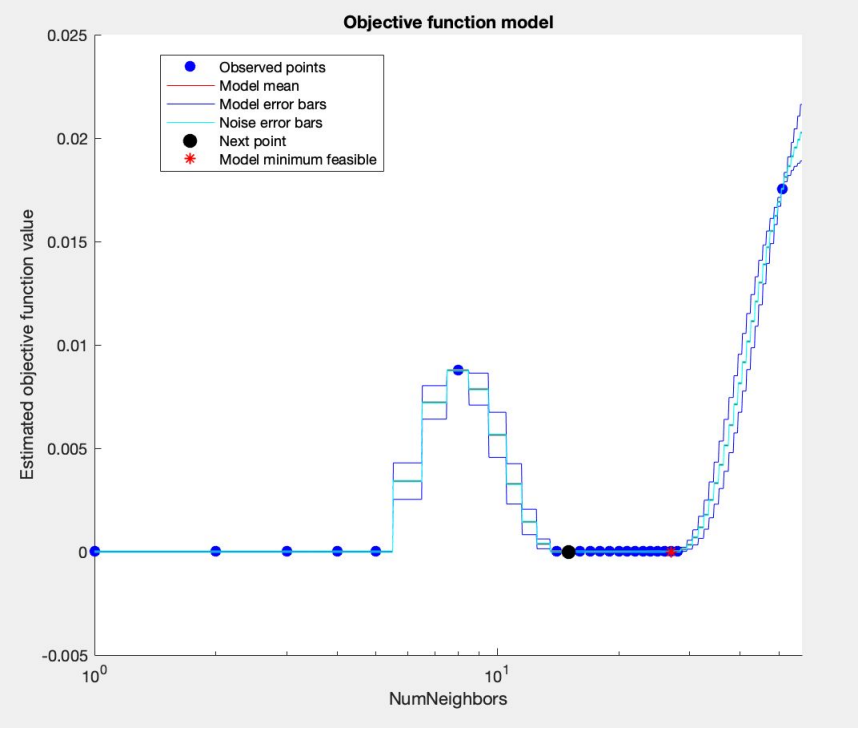
Evaluating the results of each classification technique on the test set exposed that the models best suited to interpret the data, in descending order, were perceptron, k-nearest neighbors, and decision trees.

DECISION TREES



Auto-generated by fitctree

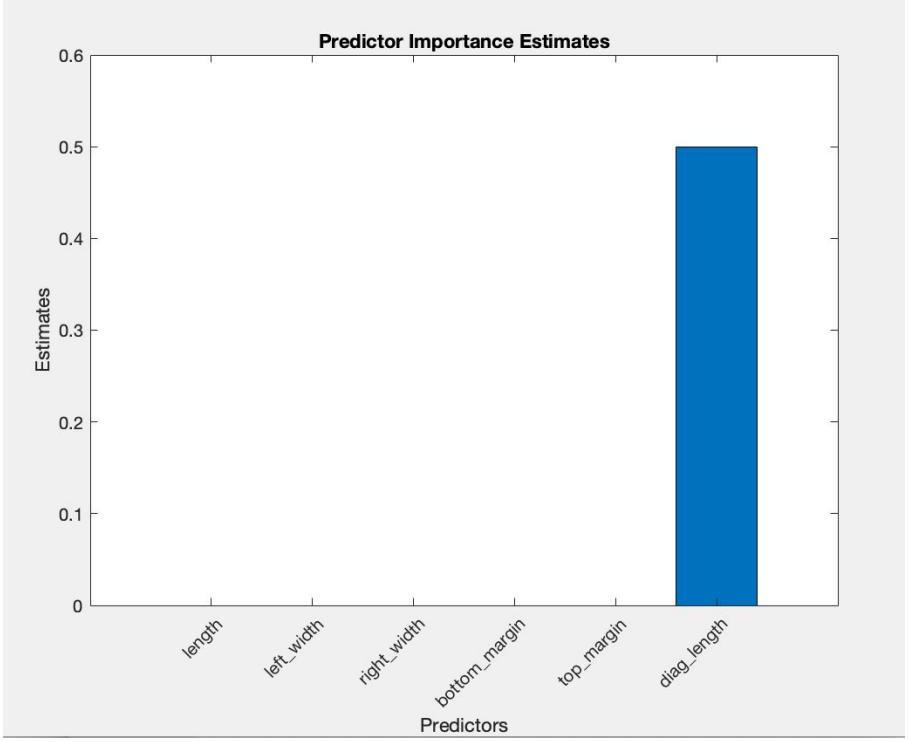
KNN



Auto-generated by fitcknn

5. Analysis of the results, discussion of errors made in training and testing and why, the relationship of feature selection to results, and any other details you found interesting about the results.

Determining feature importance for the decision tree model revealed that diag_length is the only significant feature of the dataset.



6. Future work you might be interested in performing to better understand this dataset (I will not hold you to this, don't worry.)

To better understand this dataset, future work may include an expansion of the features under consideration. The dataset provided only considers the physical dimensions of the bills but considerations such as color and design can be indicative of counterfeit currency as well. In addition, determining the opaqueness of the bill could also help determine the realness of the bill.