Visual Tool Box for Classification

Amit Meir, Yoni Fintzi

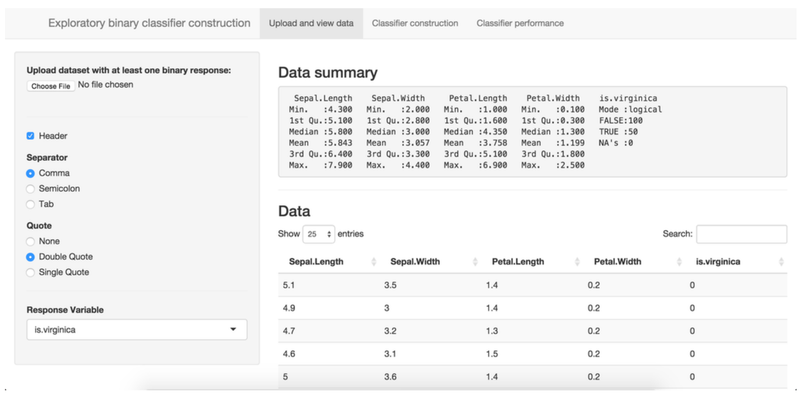
amitmeir@uw.edu, fintzij@uw.edu

Introduction

Classification is one of the most common tasks in statistics and machine learning. Constructing a classifier is highly nontrivial, as the data analyst must optimize predicting performance, often while preserving interpretability, parsimony, and scientific plausibility. Building a classifier is an iterative process, wherein the data analyst alternates between data exploration, variable selection, and model assessment.

The Visual Classification Toolbox was designed to assist an analyst in performing exploratory visualizations to facilitate classifier construction and to visually compare performance of various classifiers. Allowing the analyst to visually inspect the behaviour of the classifier with respect to different aspects of the data is central to this task.

Data Input



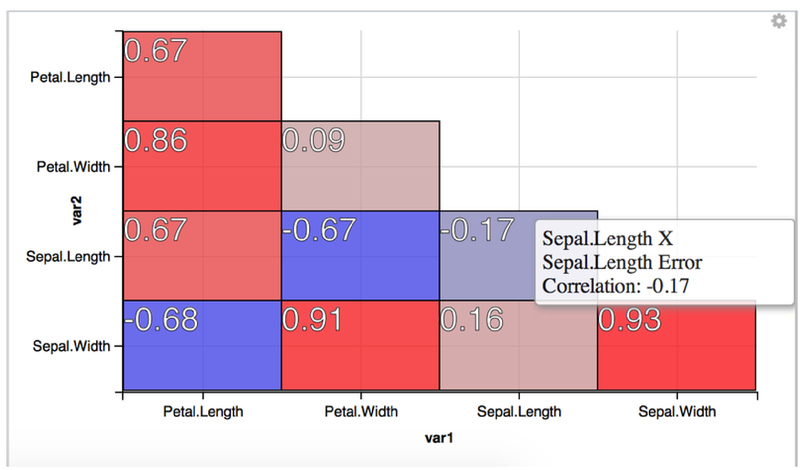
The first step in fitting a classifier to data is selecting a data set and specifying the response variable. The *Visual Classification Toolbox* allows the user to upload data sets in several formats. Once a data set has been uploaded univariate summary statistics are presented for all variables in the data set.

Model Construction

The *Visual Classification Toolbox* provides the user with two interfaces for selecting which variables to use for classification. The interfaces are designed to assist the user in quickly identifying features that have the most potential for aiding in the classification task.

As a measure for how "promising" a feature is, we compute the Spearman correlation of the variable with the errors produced by the model. In the main effect plot these correlations are double encoded via position and color.

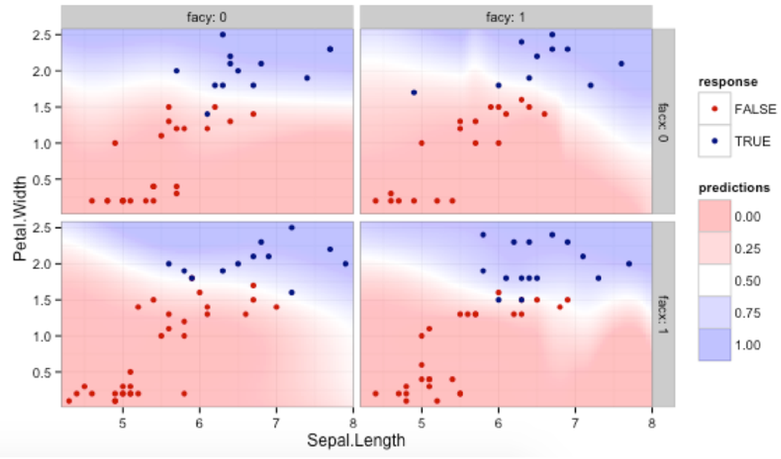
The interaction plot presents the error correlations for interaction and quadratic terms. The correlations are color encoded.



Additional Algorithms

In order to assist the user in analyzing the behavior of the classifier the *Visual Classification Toolbox* enables plotting scatter plots of any two variables and faceting according to any two other variables. The scatter plots are plotted with the decision regions of the algorithm.

Since we expect to work with high dimensional data the decision regions plotted are an interpolation of the actual decisions made on the plotted data points.



Results

An ROC and a cross validation plot are produced in order to assist the user in assessing the overall performance of the classifier.