PSTAT 131 HW 4

Chris Meade

12/1/2016

1. Description of the dataset:
   (a) I chose to analyze the Pima Indians Diabetes Data Set from the UCI Machine Learning repository. The dataset has 768 observations of 9 variables. All attributes are numerical. Features include number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin, body mass index, diabetes pedigree function, age, and class label (diabetic or not diabetic).
   (b) I think all variables will be important for analysis. This hypothesis was confirmed by constructing a control random forest model and examining the mean decrease in Gini coeficient.
   (c) The dataset is high quality, but it could be better. Several features have missing values, and it could benefit from having more observations. Overall, I think the flaws will have only a small negative impact on the analysis.
2. Description of the research question:
   (a) "Can machine learning be used to diagnose type-2 diabetes in women belonging to the Pima indain tribe?"
   (b) If the models are successful in diagnosing diabetes, they may help uncover some socio-economic factors that lead to diabetes. For example, if plasmaGlucose and bmi are significant features in the model construction, we might be able to conclude and diet is a contributing factor to diabetes.
3. Description of your analysis:
   (a) I will compare the performance of three machine learning algorithms: random forest, support vector machine, and neural network. The models will be construced using the caret package for R.
   (b) Models will be selected by maximizing the area under the an ROC curve. In order to limit overfitting, models will be trained with 10-fold cross validation.
   (c) The three seleted models will be compared by evaluating their accuracy in predicting classes of the test set.
4. Description of Progress:
   (a) I have imported the data and dealt with missing values. I have also constructed general models using the three algorithms. I still need to tune each one to maximize performance. I also still need to compare the performance of the tuned final models on the test set.