Ryley Higa

Netid: higa2

Homework 1 Report: Pima Indians Classifier

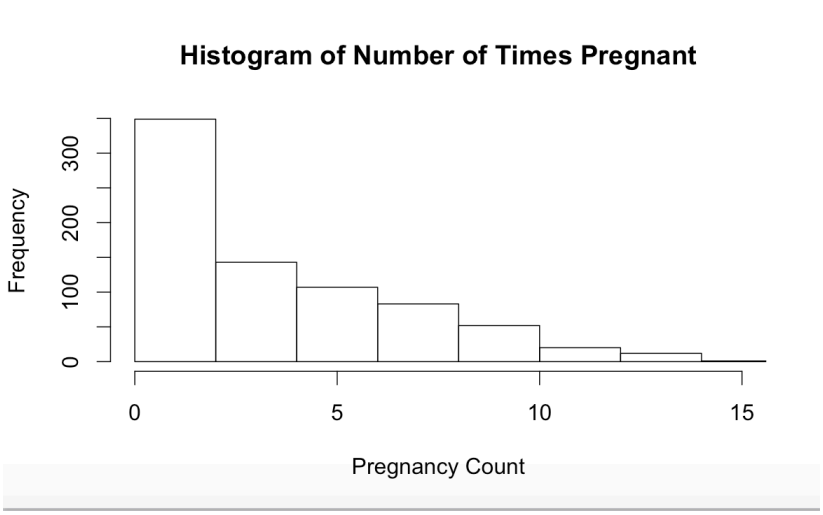
Introduction

The Pima Indians Dataset uses features described in the link, http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.names, to classify whether or not patients show signs of diabetes. Missing data in features were labeled as 0 and may skew the mean and standard deviation. Parts A, C, and D ignore such skewness. To calculate the accuracies for all parts, I averaged the accuracies over 10 trials of 80%-20% train-test splits on 10 different models.

	Part A	Part B	Part C	Part D
Train	0.752195121951	0.7404878048780	0.7764227642276	0.7609756097560
Test	0.742483660130	0.7254901960784	0.7686274509803	0.7542483660130

Analysis

In part A, training and testing examples were classified with a Naïve Bayes Classifier under the assumption that every feature was conditionally normally distributed. Features such as pregnant counts were not normally distributed accounting for mediocre accuracy.



In part B, training and testing were tested with the same model, except missing values labeled as 0 were replaced with NA, and were ignored during training and testing. The training and testing accuracy decreased as a result of eliminating missing values, which may suggest that missing values were unevenly distributed between the two classes.

In part C, I used the klaR and caret packages to train and test a new Naïve Bayes model. The new model uses 10 fold cross-validation and does not assume features are conditionally normal. The increase in accuracy by about 2-4% from Part A and B suggests that throwing out the normality assumption improved the model.

In part D, I used klaR, caret, and SVMLight to train a SVM to classify the data using the svmlight function. The training and testing accuracy decreased from the Naïve Bayes. In terms of accuracy, the best classifier seemed to be the Naïve Bayes Classifier built using caret and klaR.