Homework 5: Decision Trees

Sreesha Venkat

April 3, 2017

Collaborators: Sydney McMuldroch, Lynn Kong, Roshni Patel

Q1. Decision Trees

a. See appendix for decision tree code.

Q2. Random Forests

a. See appendix for random forest code.

Q3. Implementation Details

a. I turned each categorical feature into a binary value, so that my decision tree could use the same 0-1 decision process throughout the entire data structure. I replaced qualitative missing values with the mean of that feature, and replaced qualitiative missing values with the mode.

b. I tested several different tree depths, and decided to stop once I noticed that an additional layer did not lower my error rate.

c. I wrote my imputed data to a csv that I could access during training, so that I could save time that went to preprocessing.

d. I used the decision tree class I already created, and trained several trees on my random forest data. Then, I took the average of the predicted values from each of those classifiers, and assigned that rounded value to be the final label for each data point.

e. From a coding architecture standpoint, I think the layout of my preprocessing, random forest, and decision tree classes/functions are cool because I've made it simple to calculate error rates across the different classes.

Q4. Performance Evaluation

a. Census

   i. Decision Tree, Training Accuracy: 0.8655

   ii. Decision Tree, Validation Accuracy: 0.8533

   iii. Random Forest, Training Accuracy: 0.8588

   iv. Random Forest, Validation Accuracy: 0.8515

b. Titanic

   i. Decision Tree, Training Accuracy: 0.8735

   ii. Decision Tree, Validation Accuracy: 0.8614

iii. Random Forest, Training Accuracy: 0.8811

iv. Random Forest, Validation Accuracy: 0.8702

c. Spam

i. Decision Tree, Training Accuracy: 0.7821

ii. Decision Tree, Validation Accuracy: 0.7693

iii. Random Forest, Training Accuracy: 0.8023

iv. Random Forest, Validation Accuracy: 0.7932

c. Kaggle Scores, Username: SreeshaVenkat

i. Census: 0.84055

ii. Titanic: 0.83226

iii. Spam: 0.76780

Q5. Spam Writeup

a. I did not use any other features or feature transformations.

b. i. ("!") $\geq$ 1

ii. ("money") $\geq$ 1

iii. ("energy") $\geq$ 1

iv. Therefore this email was spam.

c. i. ("!") $<$ 1 (12 trees)

ii. ("money") $<$ 1 (10 trees)

iii. ("prescription") $<$ 1 (8 trees)

Q6. Census Writeup

a. I did not use any other features or feature transformations.

b. i. ("marital-status-Married-civ-spouse") $<$ 1

ii. ("capital-gain") $<$ 5178

iii. ("age") $<$ 61

iv. ("education-num") $\geq$ 12

v. Therefore this person was rich

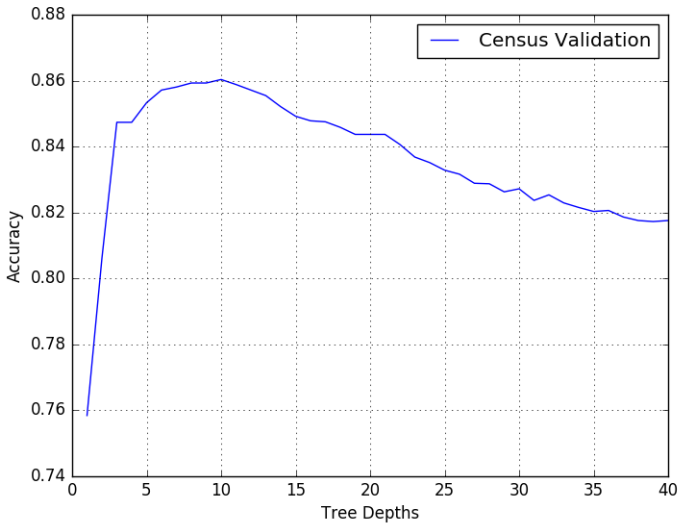
c. i. ("age") $<$ 29 (12 trees)

ii. ("capital-gain") $<$ 7262 (7 trees)

iii. ("relationship-Husband") $<$ 1 (11 trees)

d. I had the highest validation accuracy at a depth of 11. This is likely because depths lower than 11 were underfitting, whereas depths that were greater than 11 ended up overfitting the data, ultimately resulting in poorer overall predictions.

Figure 1: Census Validation Accuracies



Q7. Titanic Writeup