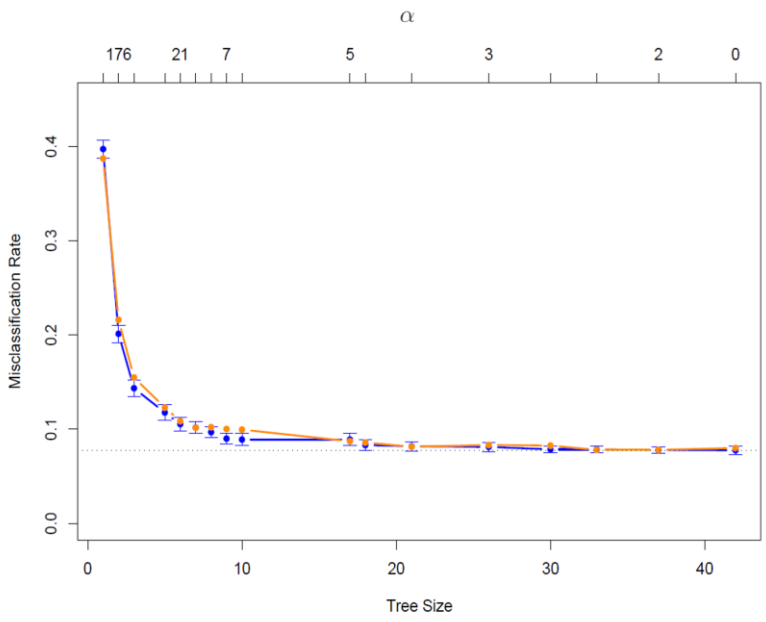
Problem 1 for Project 1:

Figure of Project 1:



Citation of Problem 1:

Hastie et al. Figure 9.4, page: 314.

Problem Setting of problem 1:

In our project we are going to use dataset having information of email. This email is collected from office and personal email. Here as an input different feature of email will be used as input like percent of common words, length of capital latter etc. This is mainly a spam-based data set where output will be the class of spam or not. But the ultimate output of this project is a figure where cross-validation estimate of misclassification rate of tree size and cross validation.

Here a tree base algorithm compares the values of the attributes and move forward in a recursive way to find out the class is a spam or not.

DataSource of Problem 1:

Here problem is to find out the misclassification rate based on tree size for Spam data. We will analyze email span data set by tree-based algorithm and in the output we will get the email is spam (1) or not (0). From the analysis we will find the cross-validation estimate of misclassification rate for different tree size.

Data Set:

Spam Data by George Forman from Hewlett-Packard laboratories, Palo Alto, California.

URL:
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spambase/

Dataset dimension:
Number of Instances: 4601
Number of Attributes: 58

Detail information of Attributes:

Input X features:
Here in the data set there are 57 observation which is used as input X

48 continuous real [0,100] attributes of type word_freq_WORD = percentage of words in the e-mail that match WORD,

6 continuous real [0,100] attributes of type char_freq_CHAR = percentage of characters in the e-mail that match CHAR,

1 continuous real [1,...] attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

Output y label:

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0),

Here y level is indicating 1or 0 to identify type is spam or not.

Algorithm:

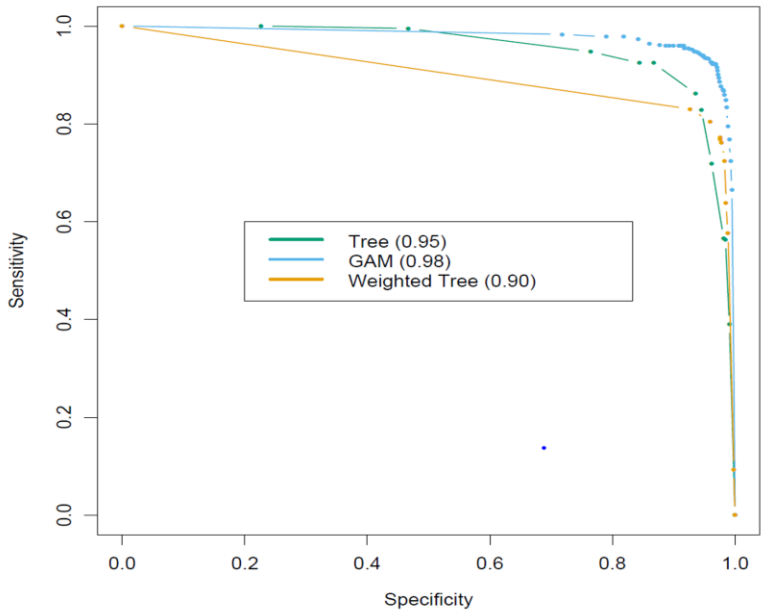
Pseudo Code:
1.Initialise the data set and Start at root node
2.For each ordered variable X,convert it to an unordered variable X' by grouping it values in the node into small number of intervals if X is unordered, then set X'=X
3. Perform a chi-squared test of independence of each X' variable versus Y on the data in the node and compute it significant probability
4. Choose the variable X* associated with X' that has the smallest probability
5. Find the split set {X* € S*} that minimizes the sum of Gini indexes and use it to split the node into two child nodes
6. 2-5 is repeated until it reached to stopping criteria
7.Prune the tree with the CART algorithm
8.Calculate misclassification rate from the classification result

9. Repeat 1-8 for different cross validation
10. create cross validation misclassification rate vs tree size graph

Here the primary output of the tree will be the classification of the spam or not. From the result error rate will be measured and plotted. Here we can learn classification tree, CART algorithm. Cross validation techniques.
To implement the project, I will use python where I have to use some frameworks like pandas, scikit ,numpy. So this project will help me to learn this frameworks deeply.

Problem 2 for Project 1:

Figure of Project 2:



Citation of Problem 2:

Hastie et al. Figure 9.5, page: 316.

Problem Setting of problem 2:

In our project we are going to use dataset having information of email. This email is collected from office and personal email. Here as an input different feature of email will be used as input like percent of common words, length of capital latter etc. This is mainly a spam-based data set where output will be the class of spam or not. But the ultimate output of this project is a figure where sensitivity vs specificity for tree, GAM, and weighted tree algorithm .

Here a tree base algorithm compares the values of the attributes and move forward in a recursive way to find out the class is a spam or not. From the classification rate we will measure the

sensitivity and specificity to understand the performance of the three Tree, GAM and weighted tree.

DataSource of Problem 2:

Here problem is to find out the misclassification rate based on tree size for Spam data. We will analyze email span data set by tree-based algorithm and in the output, we will get the email is spam (1) or not (0). From the analysis we will find the cross-validation estimate of misclassification rate for different tree size. Then from the error rate, sensitivity and specificity will be measured.

Data Set:

Spam Data by George Forman from Hewlett-Packard laboratories, Palo Alto, California.

URL:
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spambase/

Dataset dimension:
Number of Instances: 4601
Number of Attributes: 58

Detail information of Attributes:

Input X features:
Here in the data set there are 57 observation which is used as input X

48 continuous real [0,100] attributes of type word_freq_WORD = percentage of words in the e-mail that match WORD,

6 continuous real [0,100] attributes of type char_freq_CHAR = percentage of characters in the e-mail that match CHAR,

1 continuous real [1,...] attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

Output y label:

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0),

Here y level is indicating 1or 0 to identify type is spam or not.

Algorithm of project 2:

Pseudo Code:

1.Initialise the data set and Start at root node
2.For each ordered variable X, convert it to an unordered variable X' by grouping it values in the node into small number of intervals if X is unordered, then set X'=X
3. Perform a chi-squared test of independence of each X' variable versus Y on the data in the node and compute it significant probability
4. Choose the variable X* associated with X' that has the smallest probability
5. Find the split set {X* € S*} that minimizes the sum of Gini indexes and use it to split the node into two child nodes
6. 2-5 is repeated until it reached to stopping criteria
7.Prune the tree with the CART algorithm
8. Calculate error rate
9.Calculate sensitivity
10. Calculate specificity
11. Plot Sensitivity vs specificity
12. This approach is for tree model; we need to bring some small changes 4-5 for weighted tree and GAM.

This project will help me to learn different tree-based algorithms like tree, weighted tree, GAM and their performances and differences. Moreover, I have mentioned earlier for tree algorithm I am going to use pandas, numpy and scikit, it will help me a lot to deeply learn python framework to implement the machine learning algorithm.