Tut 7: Decision Tree Models

Jan 2023

Classification Tree

1. Use **gain ratio** to determine which split is better:

   Split 1: Leaf $A = [20+, 15-]$; Leaf $B = [5+, 20-]$

   Split 2: Leaf $A = [10+, 2-]$; Leaf $B = [15+, 33-]$

   Remark: The larger "information gain" and "gain ratio", the better.

   Solution. Total, $Tbl(S) = [25+, 35-]$ implies $H(S) = -(\frac{25}{60}\log_2\frac{25}{60} + \frac{35}{60}\log_2\frac{35}{60}) = 0.9799$

   For Split 1:

   $Tbl(S_1|A) = [20+, 15-]$ implies $H(S_1|A) = 0.9852$
   $Tbl(S_1|B) = [5+, 20-]$ implies $H(S_1|B) = 0.7219$
   $IG(S_1) = 0.9799 - \left[\frac{35}{60}(0.9852) + \frac{25}{60}(0.7219)\right] = 0.1044$
   $I(S_1) = -\left[\frac{35}{60}\log_2\left(\frac{35}{60}\right) + \frac{25}{60}\log_2\left(\frac{25}{60}\right)\right] = 0.9799$
   $R(S_1) = \frac{0.1044}{0.9799} = 0.1065$

   For Split 2:

   $Tbl(S_2|A) = [10+, 2-]$ implies $H(S_2|A) = 0.6500$
   $Tbl(S_2|B) = [15+, 33-]$ implies $H(S_2|B) = 0.8960$
   $IG(S_2) = 0.9799 - \left[\frac{12}{60}(0.6500) + \frac{48}{60}(0.8960)\right] = 0.1331$
   $I(S_2) = -\left[\frac{12}{60}\log_2\left(\frac{12}{60}\right) + \frac{48}{60}\log_2\left(\frac{48}{60}\right)\right] = 0.7219$
   $R(S_2) = \frac{0.1331}{0.7219} = 0.1844$

   Split 2 has a higher gain ratio, hence Split 2 is preferred. □

2. (Jan 2022 Final Q4(b)) A classification tree is being constructed to predict whether the credit card application approval is positive. Consider the two splits below:

   - **Split 1**: The left node has 178 observations with 68 positives and the right node has 295 observations with 144 positives.
   - **Split 2**: The left node has 136 observations with 83 positives and the right node has 337 observations with 129 positives.

   By calculating the information gains, determine which split is better. (7 marks)

Solution. First, we calculate the entropy of Y:

$$H(Y) = -\left(\frac{212}{473}\log_2\frac{212}{473} + \frac{261}{473}\log_2\frac{261}{473}\right) = 0.9922448 \qquad \text{[2 marks]}$$

The entropy of **Split 1** is

$$\begin{aligned} H_1 &= \frac{178}{473}\left[-\frac{68}{178}\log_2\frac{68}{178} - \frac{110}{178}\log_2\frac{110}{178}\right] + \frac{295}{473}\left[-\frac{144}{295}\log_2\frac{144}{295} - \frac{151}{295}\log_2\frac{151}{295}\right] \\ &= 0.9844898 \end{aligned}$$

[1.5 marks]

The information gain for **Split 1** is

$$IG_1 = H(Y) - H_1 = 0.007755 \qquad \text{[0.5 mark]}$$

The entropy of **Split 2** is

$$\begin{aligned} H_2 &= \frac{136}{473}\left[-\frac{83}{136}\log_2\frac{83}{136} - \frac{53}{136}\log_2\frac{53}{136}\right] + \frac{337}{473}\left[-\frac{129}{337}\log_2\frac{129}{337} - \frac{208}{337}\log_2\frac{208}{337}\right] \\ &= 0.961317 \end{aligned}$$

[1.5 marks]

The information gain for **Split 2** is

$$IG_2 = H(Y) - H_2 = 0.0309278 \qquad \text{[0.5 mark]}$$

Split 2: One node has 10 observations with 8 lapses and one node has 90 observations with 22 lapses.

$$\begin{aligned} H(S) &= -\left[\frac{30}{100}\log_2\frac{30}{100} + \frac{70}{100}\log_2\frac{70}{100}\right] = 0.8813 \\ H(Split2) &= -\frac{10}{100}\left[\frac{8}{10}\log_2\frac{2}{10} + \frac{2}{10}\log_2\frac{2}{10}\right] - \frac{90}{100}\left[\frac{22}{90}\log_2\frac{22}{90} + \frac{68}{90}\log_2\frac{68}{90}\right] \\ &= -0.1 \times (-0.7219) - 0.9 \times (-0.8024) = 0.7943 \\ IG(Split2) &= 0.8813 - 0.7943 = 0.0870 \end{aligned}$$

Since **Split 2** has a higher information gain, it is a better split. ..... [1 mark]

□

3. (May 2020 Final Q4(b)(ii)) In trying to build a model that is able to predict whether or not an email message is spam based on the following predictors:

   - to_multiple: Indicator for whether the email was addressed to more than one recipient;
   - image: Indicates whether any images were attached;
   - attach: Indicates whether any files were attached;
   - dollar: Indicates whether a dollar sign or the word 'dollar' or 'ringgit' appeared in the email;
   - winner: Indicates whether "winner" appeared in the email;
   - num_char: The number of characters in the email, in thousands;

- format: Indicates whether the email was written using HTML (e.g. may have included bolding or active links) or plaintext;
- re_subj: Indicates whether the subject started with "Re:", "RE:", "re:", or "rE:";
- number: Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

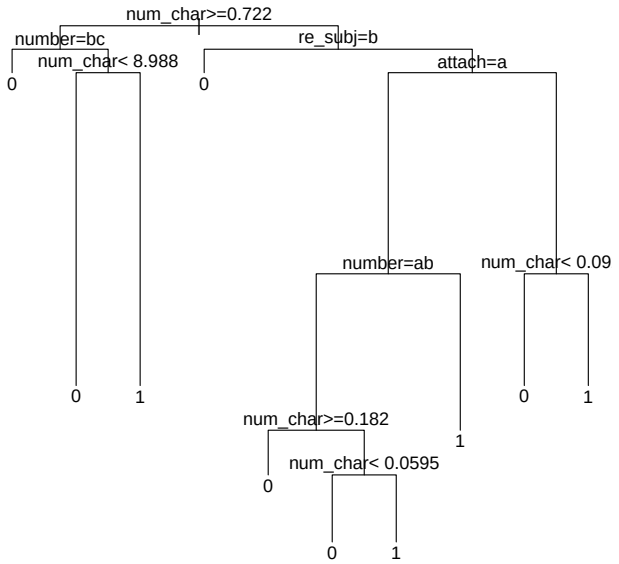
Note that "spam" is denoted with the value 1 while "non-spam" is denoted with the value 0. The trained logistic regression model has the parameters given in Figure 4.2.

Table 4.2: Coefficients of Logistic Regression

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.468478   0.181285  -8.100 5.48e-16 ***
to_multipleyes  -2.152057   0.349538  -6.157 7.42e-10 ***
imageyes        -1.467843   0.797895  -1.840 0.065820 .
attachyes        0.957716   0.281455   3.403 0.000667 ***
num_char        -0.014651   0.007199  -2.035 0.041849 *
dollaryes        0.453477   0.197009   2.302 0.021346 *
winneryes        1.994563   0.392252   5.085 3.68e-07 ***
numbersmall     -1.227981   0.186300  -6.591 4.36e-11 ***
numberbig       -0.561313   0.263563  -2.130 0.033195 *
formatPlain      1.032511   0.171915   6.006 1.90e-09 ***
re_subjyes      -2.447223   0.398309  -6.144 8.05e-10 ***
---
Signif. :  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If an email does not address to multiple, has no image, no attached file(s), no "dollar" sign, does not have the word "winner", has 20.133×10^3 number of characters and is in HTML format, has no subject starting with "Re:" and has a small number in the email. **Determine** whether the email is a spam using the trained logistic regression model and using the decision tree model (you will need to interpret the decision tree model based on your knowledge of "rpart" algorithm) given in Figure 4.3.

Figure 4.3: The trained decision tree model.



(4.5 marks)

Solution. Given the predictors, the probability of spam is

$$\begin{aligned}\mathbb{P}(Y=1|X=x) &= \frac{1}{1+\exp(-(-1.468478+\beta^T x))} \\ &= \frac{1}{1+\exp(1.468478+1.52295)} = 0.047815\end{aligned}$$ [1 mark]

where

$$\begin{aligned}\beta^T x = &-2.152057*0-1.467843*0+0.957716*0-0.014651*20.133 \\ &+0.453477*0+1.994563*0-1.227981+1.032511*0 \\ &-2.447223*0=-1.52295\end{aligned}$$ [1.5 marks]

Since the probability of the email being spam is smaller than 0.5, we predict it to be a **non-spam**. ..... [0.5 mark]

The prediction with tree is as follows:

(a) number of characters = $20.133 > 0.722$, go to left subtree;

(b) number = small. It should match "b", therefore, go to left subtree, given us 0, i.e. **non-spam**.

Marks are deducted if no justification is given. ..... [1.5 marks] □

4. (Jan 2021 Final Q2(a)) The dataset in Table 2.1 is used to build a classification tree which predicts if a student pass predictive modelling (Pass or Fail, P, F for short), based on their previous GPA (High, Medium, or Low, H, M, L for short) and whether they have or have not (Y or N in short) put in significant efforts in their study.

Table 2.1: Training dataset for classification problem.

GPA	Studied	Pass
L	N	F
L	Y	P
M	N	F
M	Y	P
H	N	P
H	Y	P

Construct and plot the ID3 classification tree (using information gain) with appropriate labels. You must show all the calculation steps. (5 marks)

Solution. First, we calculate the entropy

$$H = -(\frac{2}{6}\log_2\frac{2}{6}+\frac{4}{6}\log_2\frac{4}{6}) = 0.9182958$$ [1 mark]

The information gain for the variable GPA is

$$\begin{aligned}IG_1 = H &- (-\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}) \\ &- \frac{1}{3}(\frac{2}{2}\log_2\frac{2}{2}+\frac{0}{2}\log_2\frac{0}{2})) \\ &= 0.9182958 - \frac{1}{3}(1+1+0) = 0.2516291\end{aligned}$$ [1 mark]

The information gain for the variable Studied is

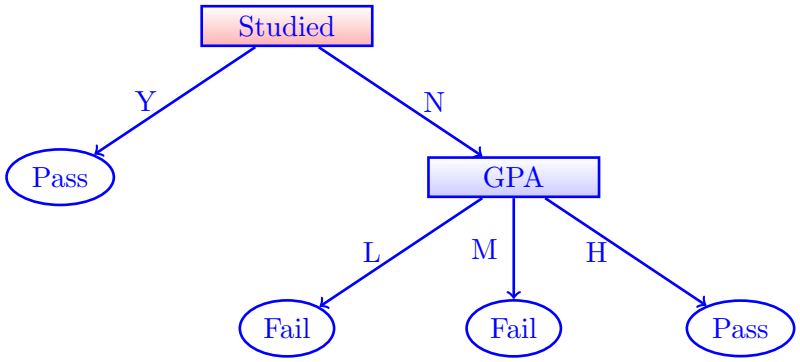
$$IG_2 = H - \left(-\frac{1}{2}(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) - \frac{1}{2}(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3})\right)$$
$$= 0.9182958 - \frac{1}{2}(0.9182958 + 0) = 0.4591479$$

[1 mark]

The variable "Studied" is choice for the ID3 split because it has higher information gain.

For "Studied=N", we have three more branches because the output variable Pass is not pure.

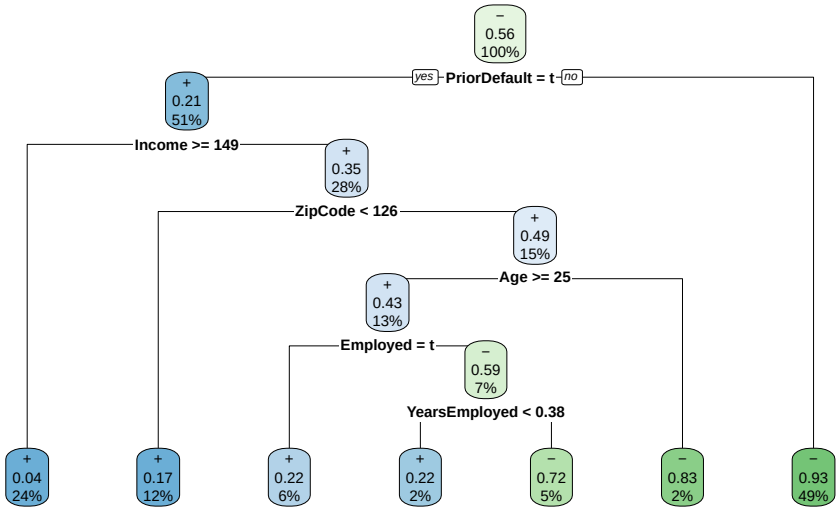
For "Studied=Y", the output variable Pass is already pure. ..... [0.5 mark]



..... [1.5 marks] □

5. (Jan 2022 Final Q2(b)) For the same training data (as Tutorial 4 Q1, i.e. Jan 2022 Final Q2(b)), use the CART tree in Figure 2.1 to predict the the credit card application being approved (positive or negative) for a male of age 22.08 with a debt of 0.83 unit who has been employed for 2.165 years with no prior default and is currently unemployed, has a credit score 0 and a zip code 128 with income 0.

Figure 2.1: CART tree for credit card application approval data.



You need to show your workings by explaining the steps to move left or right in the tree travesal to reach the prediction. (4 marks)

Solution. From the decision tree, we move right (no prior default) directly to negative (Probability=0.93. It consists of 49% of the training data). ..... [3 marks]

The credit card application being approved is negative. ..... [1 mark] □

6. (Jan 2022 Final Q2(c)) Compare the ability of the logistic regression model and the C4.5 tree model in the handling missing values and the prediction of highly nonlinear data. (4 marks)

Solution. Logistic regression model will omit data with missing values since the mathematical model does not allow the arithmetic calculation for missing value. ..... [1 mark]

C4.5 tree model will ignore the missing value in the feature and compute the gain ratio. [1 mark]

Logistic regression model performs poorly when it is used to predict highly nonlinear data since the model is linear. ..... [1 mark]

C4.5 tree model performs better compare to logistic regression model when it is used to predict highly nonlinear data since it is nonlinear. However, it may overfit and does not generalise well. [1 mark] □