Predictive Modelling Tutorials 5 & 6: Decision Trees

Dr Liew How Hui

Jan 2021

Tut 5: Decision Trees

Gini Impurity

$$G(S) = 1 - \sum_{j=1}^{K} p_j^2, \quad p_j = \frac{|S_j|}{|S|} \tag{1}$$

$$G^T(\underbrace{\{S_1, \cdots, S_L\}}_{S^A}) = \sum_{i=1}^{L_A} \frac{|S_i|}{|S|} G(S_i). \tag{2}$$

Tut 5: Decision Trees

(Information) Entropy:

$$H(S) = -\sum_{i=1}^{K} p_i \log_2 p_i \tag{3}$$

$$H^T(\underbrace{\{S_1, \cdots, S_L\}}_{S^A}) = \sum_{i=1}^{L_A} \frac{|S_i|}{|S|} H(S_i). \tag{4}$$

Information Gain

$$IG(S^A) = H(S) - \sum_{i \in Values(A)} \frac{|S_i|}{|S|} H(S_i) \tag{5}$$

Tut 5: Decision Trees

Intrinsic Information

$$H(A) = -\sum_{i=1}^{K} \frac{|A_i|}{|A|} \log_2 \frac{|A_i|}{|A|}. \tag{6}$$

Gain ratio

$$R(S^A) = \frac{IG(S^A)}{H(A)} \tag{7}$$

Tutorial 5, Q2

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.

Tutorial 5, Q3

You are a team member from the data team of ClickMe, an online platform selling electronic products. Your company would like to do target marketing. You were asked to build a model to predict whether the customer will buy the product based on their demographic information such as age, race, gender, and income. The data collected is shown in the table below.

Tutorial 5, Q3 (cont)

Obs	Age	Race	Gender	Income	Buy
1	52	Malay	Male	11500	Not Buy
2	22	Chinese	Male	6500	Buy
3	30	Indian	Male	8000	Buy
4	26	Indian	Female	8500	Buy
5	27	Malay	Female	6500	Buy
6	32	Malay	Female	6000	Not Buy
7	33	Malay	Male	9500	Not Buy
8	50	Indian	Female	4000	Not Buy
9	31	Chinese	Male	10500	Buy
10	27	Malay	Male	10000	Buy
11	25	Indian	Female	5500	Not Buy
12	40	Indian	Female	3000	Not Buy
13	48	Malay	Female	8000	Not Buy
14	46	Chinese	Female	7000	Buy
15	51	Indian	Female	3000	Not Buy
16	42	Indian	Male	5000	Buy

Tutorial 5, Q3 (cont)

Additional Information:

- Use L-way split for qualitative predictors
- Split for *Age* is 31.5
- Split for *Income* is 6250

(a) Construct a decision tree with pure leaves by using information gain as criterion.

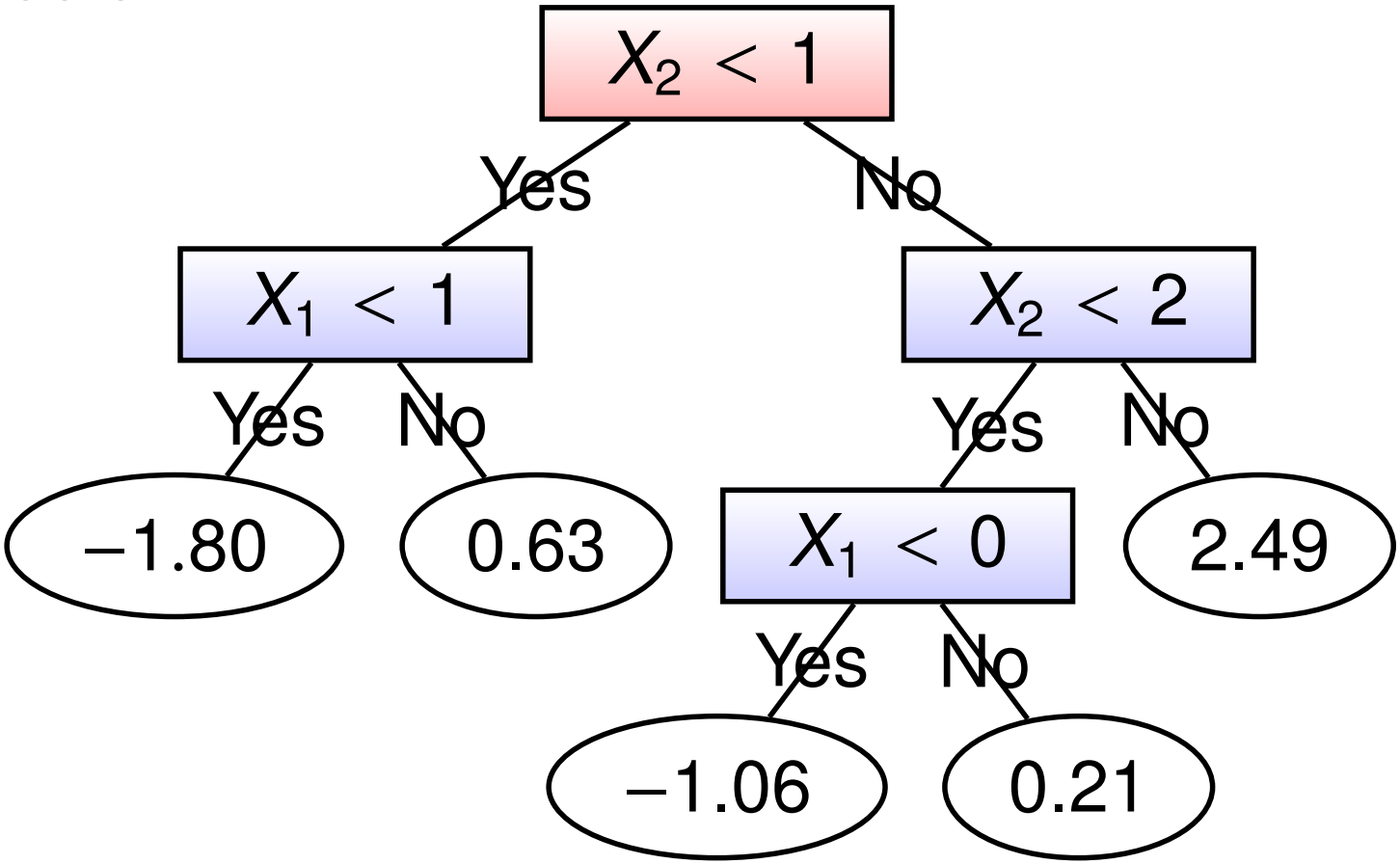
(b) Construct a decision tree with pure leaves by using Gini impurity as criterion.

Tutorial 5, Q1

(a) Sketch the (regression) tree corresponding to the partition of the predictor space illustrated in the figure below.

Tutorial 5, Q1 (cont)

(b) Create a partition space using the tree illustrated below.



FA May 2020 Q4 (b)

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.

FA May 2020 Q4 (b)

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
```

FA May 2020 Q4 (b) cont

(i) Calculate the odds of an email being a spam if "winner" appeared in the email and interpret the results on the likelihood of receiving a spam having the word "winner" over a spam not having the word "winner".

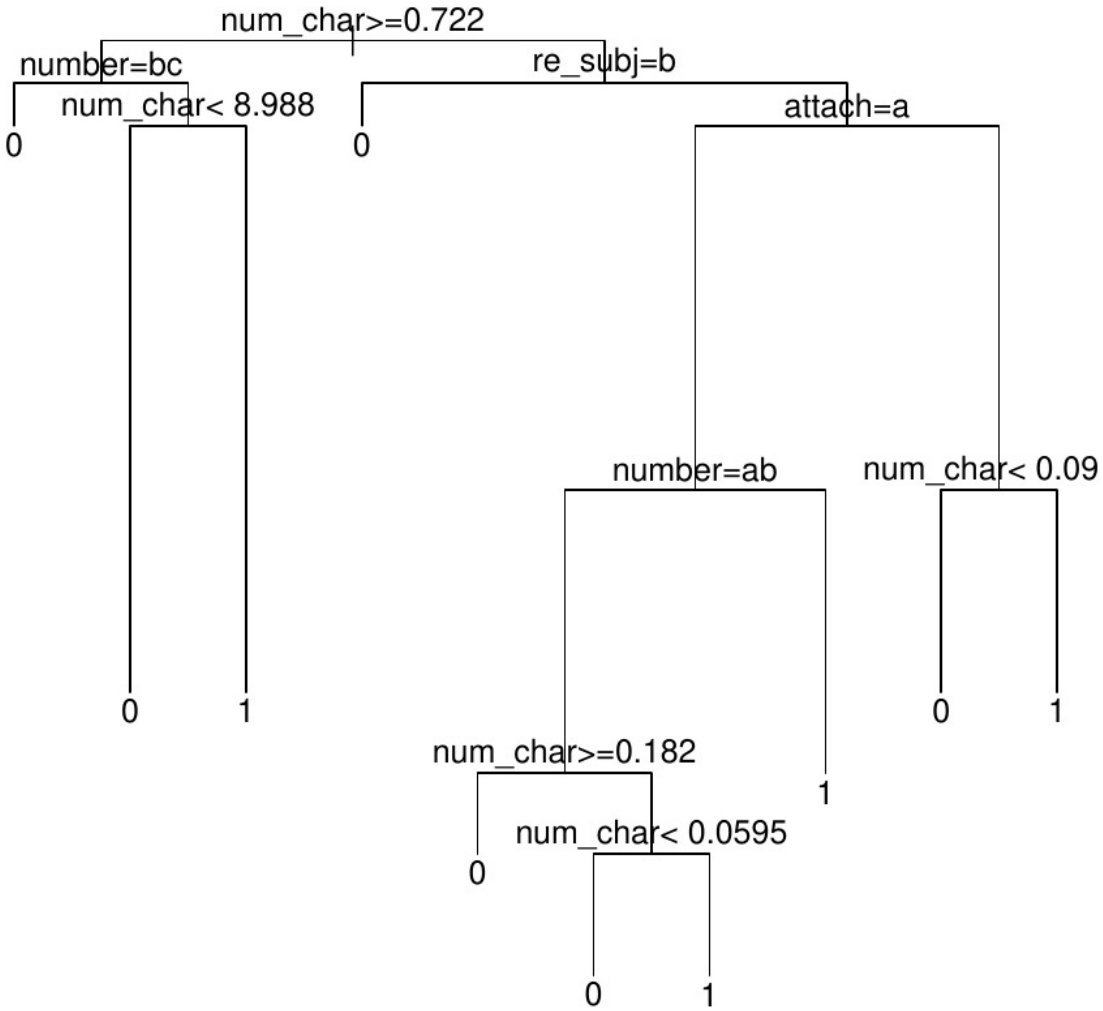
(1.5 marks)

FA May 2020 Q4 (b) cont

(ii) 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.

FA May 2020 Q4 (b) cont

Figure 4.3: The trained decision tree model.



(4.5 marks)

Tutorial 5, Q3 (cont)

If the following is not discussed in Practical Lab, they can be discussed in tutorial:

```
d.f = read.csv("tut5q3.csv")  # Original Data
d.f = read.csv("tut5q3b.csv")  # Transformed Data
C50model = C5.0(d.f[ , 1:4], d.f[ , 5])
plot(C50model)
C50model = C5.0(d.f[ , 1:4], d.f[ , 5], control=C5.0Control(
  subset=FALSE,noGlobalPruning=TRUE,earlyStopping=FALSE,
  minCases=1,CF=1))
dev.new()
plot(C50model)
```