Classification and Regression Trees (with missingness)

PH240C Lab 02

Classification and Regression Trees (CART)

univariate covariate

1. Suppose we have i.i.d. sample with pairs (Y_i, X_i), $i = 1, \ldots, n$, and X_i lives in a discrete sample space $X_i \in \mathcal{X} = \{x_1, \ldots, x_d\}$;

2. For $j = 1 : d$

   2.1 Split the data set into two groups:

   $$G_{\text{left}}(j) = \{i : X_i \leq x_j\}, \quad G_{\text{right}}(j) = \{i : X_i > x_j\};$$

   2.2 Calculate the within group "measure of similarity":
   - Sum of squares for regression trees (RSS), $s^2_{\text{left}}(j)$ and $s^2_{\text{right}}(j)$;
   - Impurity measure for classification trees.

   2.3 Calculate the split "quality":
   - Regression tree – the split total RSS: $s^2(j) = \frac{|G_{\text{left}}(j)|}{n} s^2_{\text{left}}(j) + \frac{|G_{\text{right}}(j)|}{n} s^2_{\text{right}}(j)$;
   - Classification tree – the split total weighted impurity;

3. Split the data into two groups with threshold that maximize between nodes difference and the within node similarly;

4. Keep splitting with in each group following Step 2.

Classification tree with different impurity measures

- Suppose $Y_1, \ldots, Y_n$ are the binary responses in a classification tree;
- We consider a simple scenario that we split the parent node R into two child nodes R_1 and R_2;
- Define the proportion:

$$p_0(R_j) = \frac{1}{|Rj|} \sum_{i \in R_j} (1 - Y_i), \quad p_1(R_j) = \frac{1}{|Rj|} \sum_{i \in R_j} Y_i, \quad j = 1, 2.$$

- Possible impurity functions calculated in each node, for $j = 1, 2$:
  - Entropy function: $E(R_j) = -p_0(R_j) \log p_0(R_j) - p_1(R_j) \log p_1(R_j)$;
  - Gini index: $G(R_j) = p_0(R_j)(1 - p_0(R_j)) + p_1(R_j)(1 - p_1(R_j))$
- Then the split impurity is calculated via, take Entropy for example:

$$\frac{|R_1|}{n} E(R_1) + \frac{|R_2|}{n} E(R_2).$$

What if we have some missing values in the response?

Subject	Y	Weight	Height
1	1	10	10
2	1	9	9
3	NA	8	8
4	1	7	7
5	1	6	5
6	0	5	6
7	0	4	4
8	0	3	3
9	0	2	2
10	0	1	1

1. may not be helpful for prediction (can be deleted)
2. can be treated as a separate category

What if we have some missing values in the covariates?

Subject	Y	Weight	Height
1	1	10	10
2	1	9	9
3	1	NA	8
4	1	7	7
5	1	6	5
6	0	5	6
7	0	4	4
8	0	3	3
9	0	2	2
10	0	1	1

If $R_1 = \{1, 2, 3, \ldots, 7\}$ and $R_2 = \{8, 9, 10\}$,

- Without missing value, we calculate the split impurity measure as:

$$\frac{7}{10}\left(-\frac{2}{7}\log\frac{2}{7} - \frac{2}{7}\log\frac{2}{7}\right) + \frac{3}{10}E(R_2) = 0.418.$$

- With missing value, we calculate the split impurity measure as:

$$\frac{6}{9}\left(-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}\right) + \frac{3}{9}E(R_2) = 0.424$$

The region without missing values receives **higher** weight.

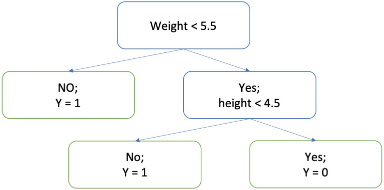
Missing covariates in the training data

- When we have missing covariates in the training data, we need to adjust the impurity measure;
- The impurity measures (either Gini index or Entrooy) are calculated only over the observations which are not missing a particular predictor.
- To weight the calculated impurity measures, the weighting probabilities are also calculated only over the non-missing observations.
- Problems? Issues with this construction? Can you identify a case that this construction is flawed? Hint: What happens if one variable has only two observations which are not missing? (Homework question)

What if we have some missing values in the covariates?

Subject	Y	Weight	Height
1	1	10	10
2	1	9	9
3	1	NA	8
4	1	7	7
5	1	6	5
6	0	5	6
7	0	4	4
8	0	3	3
9	0	2	2
10	0	1	1

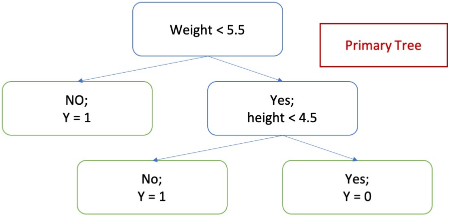
Following the new measure of split, we grow the primary tree in `rpart`:



Missing covariates in the testing data

Predict for the responses in the testing data:

Subject	Y	Weight	Height
1	?	NA	3
2	?	NA	4
3	?	NA	5
4	?	NA	6
5	?	3	3
6	?	4	4
7	?	5	5
8	?	6	6

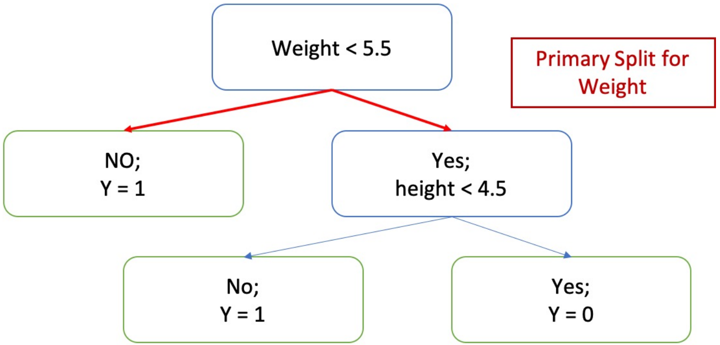


Surrogate Splits (1)

- Decision trees can handle missing values without imputation;
- When an observation is missing, *primary tree* cannot make a decision.
- What if we pretend this variable is just not there?
  1. As when the variable is missing, we cannot split based on this variable either;
  2. Instead, we want to find a *replacement split* by using other variables.
- Ideally, we want the replacement split to be similar to the primary split;
- If a case with a missing variable used in a *primary split* has to be predicted, a surrogate split is used instead.

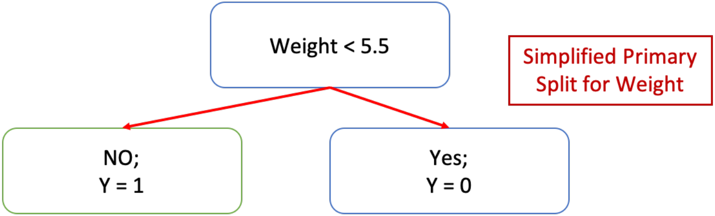
Surrogate Splits (2)

In our tree, the primary split for the missing variable "weight" is:



Surrogate Splits (3)

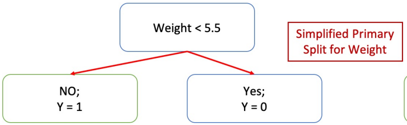
In our tree, the primary split for the missing variable "weight" can be further simplified:

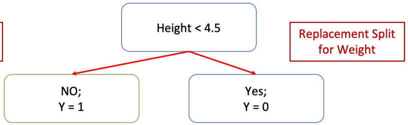


When weight is missing, question is can we find a replacement split that is similar to this primary split?

Surrogate Splits (4)

Are these two splits similar?

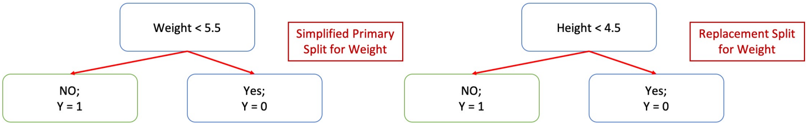




The original data:

Subject	Y	Weight	Height
1	1	10	10
2	1	9	9
3	1	NA	8
4	1	7	7
5	1	6	5
6	0	5	6
7	0	4	4
8	0	3	3
9	0	2	2
10	0	1	1

Surrogate Splits (5)



- As these two splits produces similar classification results of the responses, we call the second split as the "Surrogate Split" for the primary split for weight;
- The benefit is that we can still carry out meaningful prediction with missing covariates;
- Non-missing data are still predicted based on the primary split.

Prediction with missing covariates in surrogate splits

Subject	Y	Weight	Height
1	?	NA	3
2	?	NA	4
3	?	NA	5
4	?	NA	6
5	?	3	3
6	?	4	4
7	?	5	5
8	?	6	6

$\rightarrow$

Subject	Y	Weight	Height
1	1	NA	3
2	1	NA	4
3	0	NA	5
4	0	NA	6
5	?	3	3
6	?	4	4
7	?	5	5
8	?	6	6