UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

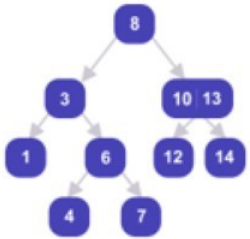
Decision Trees

Building tree-based models

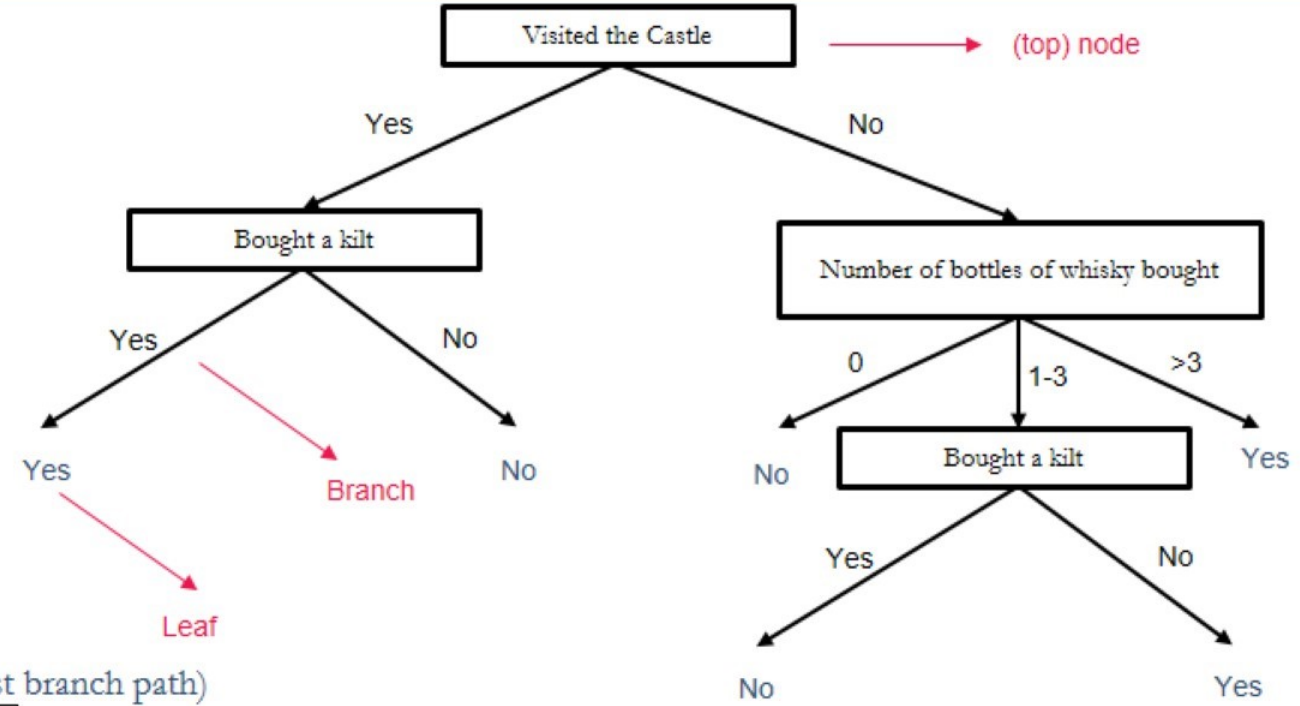
Instructions

- Create two trees that use the variables as splitting points and for the branches use the values of the variables. Try to make a tree that is good at classifying the data (e.g., has a high accuracy) for the label Tourist. You can draw this by hand, or use text in an innovative way like this:
- Bought a kilt Yes -> Tourist = Yes
- No -> Tourist = No
- Next, try to make another different one and compare you models. Which tree is larger, or smaller, classifying more observations correctly, etc.

	Visited the Castle	Bought a kilt	#bottles of whisky bought	Tourist
Visitor 1	Yes	Yes	4	Yes
Visitor 2	Yes	No	1	No
Visitor 3	No	No	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 5	No	Yes	2	No



A solution (not necessarily a good one)



Depth = 3 (longest branch path)

Size = 4 (#nodes)

#leaves = 6

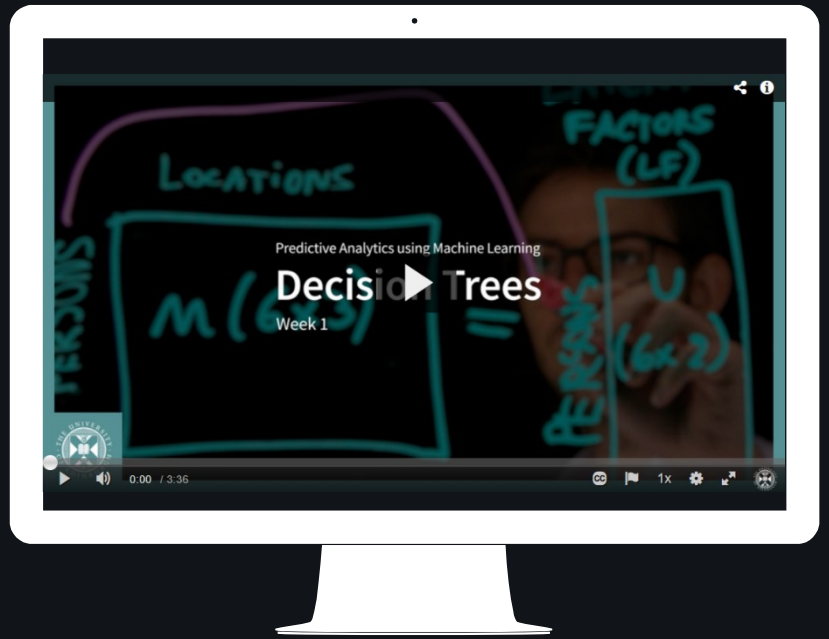
Reflecting on decision trees

Now, see how they perform for these new examples, see how your own trees compare:

- Are they overfitting or too general?
- What is their depth?
- Do they have high accuracy?

	Visited the Castle	Bought a kilt	#bottles of whisky bought	Tourist
Visitor 6	Yes	Yes	1	Yes
Visitor 7	Yes	No	10	No
Visitor 8	No	No	5	Yes
Visitor 9	No	Yes	1	Yes
Visitor 10	No	Yes	0	No

Decision trees:
rules and topology



Please watch the following video:

https://media.ed.ac.uk/media/Decision+Trees/1_d0r6ad7h/141757871

Splitting criteria

- Are used to find good splitting nodes from the dataset
- Three common ones:
  - Information gain – categorical (and numeric)
  - Gini index – categorical (and numeric)
  - Sum of squared errors (SSE) - numeric
- They are all essentially trying to find the best reduction in 'unstructuredness'

Introduction to splitting criteria

Information gain (mutual information)

Entropy is the expected amount of useful of information

We use the reduction in entropy (called information gain) to find a good variable to use for splitting – less entropy means more homogeneous information

- Entropy $H(Z)$ of a categorical variable Z is expressed as follows:

$$H(Z) = -\sum_{c \in C} p(Z = c) \log_2 p(Z = c)$$

where C represents the set of different levels c of the variable (e.g. Bought a kilt: Yes/No).

Introduction to splitting criteria

- **Entropy** $H(Z)$ of a categorical variable Z is expressed as follows:

$$H(Z) = -\sum_{c \in C} p(Z = c) \log_2 p(Z = c)$$

where C represents the set of different levels of the variable.

- We measure the reduction in entropy if we split the node based on variable K, called the **conditional entropy**:

$$H(Z \mid K) = -\Sigma_{z \in Z, k \in K}\, p(Z = z, K = k) \log_2 \frac{p(Z = z, K = k)}{p(Z = z)}$$

- The **information gain** $IG(Z, K)$ of a certain split for a variable K is then:

$$IG(Z, K) = H(Z) - H(Z \mid K)$$

Using information gain allows us to find the variable that **offers us the greatest reduction of entropy**, or the best reduction in terms of making the data more homogeneous (towards the dependent variable).

Introduction to splitting criteria

$$H(Z) = -\sum_{c \in C} p(Z = c) \log_2 p(Z = c)$$

$$H(Z \mid K) = -\Sigma_{z \in Z, k \in K} p(Z = z, K = k) \log_2 \frac{p(Z = z, K = k)}{p(Z = z)}$$

$$IG(Z, K) = H(Z) - H(Z \mid K)$$

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought	Tourist (T)
Visitor 1	Yes	Yes	4	Yes
Visitor 2	Yes	No	1	No
Visitor 3	No	No	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 5	No	No	2	No

For our root node, we have the following entropy for our dependent variable 'Tourist' (T) with its two levels 'Yes' (3/5) and 'No' (2/5):

$$H(T) = -\frac{3}{5}\log_2 \frac{3}{5} - \frac{2}{5}\log_2 \frac{2}{5} = 0.971$$

For variable 'Visited the Castle' (V), we have:

$$H(T \mid V = \text{Yes}) = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} = 1$$

We get these fractions as we have 1 visitor of the castle who is a tourist, and 1 visitor who is not a tourist.

$$H(T \mid V = \text{No}) = -\frac{2}{3}\log \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3} = 0.918$$

$$IG(T, V) = H(T) - \frac{2}{5}H(T \mid V = Yes) - \frac{3}{5}H(T \mid V = No) = 0.971 - 0.4 - 0.551 = 0.02$$

Introduction to splitting criteria

Gini index (or Gini impurity)

It is a measure for the dispersion of the data and is based on the frequency distribution of the data. The Gini index can be written as follows:

$$GI = 1 - \Sigma_{i=1}^{n} p_i^2$$

with p_i the probability of belonging to a particular class.

For two classes, this shortens to $GI = p_1(1 - p_1) + p_2(1 - p_2) = 2p_1p_2$.

Our example:

Gini index prior to the split, we have $p_t = 3/5$ and $p_{nt} = 2/5$ (t:tourist; nt: not tourist)

$$G(prior\ to\ split) = 2 \cdot p_t \cdot p_{nt} = 2\frac{3}{5}\frac{2}{5} = 0.48$$

split based on buying a kilt or not (b/nb):

$$G\,(afterthesplit) = p_t\,(2 \cdot p_{b,t} \cdot p_{nb,t}) + p_{nt}\,(2 \cdot p_{b,nt} \cdot p_{nb,nt}) = \frac{3}{5}(2 \cdot \frac{2}{3} \cdot \frac{1}{3}) + \frac{2}{5}(2 \cdot \frac{0}{2} \cdot \frac{2}{2}) = 0.266$$

Reduction from 0.48 to 0.266

Gini vs information gain

- Gini impurity:
  - Goes for variables with fewer values with higher proportions
  - minimise the Gini after split

- Information gain:
  - Goes for variables with more distinct values with smaller proportions
  - Look for large reduction of entropy

Introduction to splitting criteria

Sum of squared errors (SSE)

Used for numeric values (but still based on a particular value to split)

$$SSE = \Sigma_{i \in S_1} (y_i - \bar{y}_1)^2 + \Sigma_{i \in S_2} (y_i - \bar{y}_2)^2$$

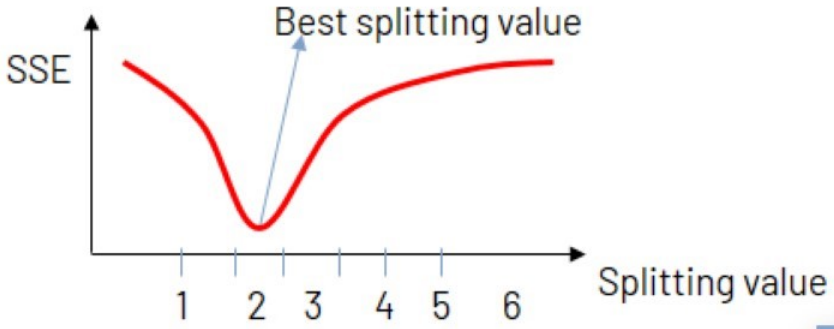
Where S_1 and S_2 are the two non-overlapping sets that contain the observations based on the split.

To decide a good splitting value:

Calculating all the possible splits will return the best split and is called **recursive partitioning**.

Various split values should be considered, often by doing left/right search until no more changes

However, this entails possibly calculating lots of splits.



Introduction to splitting criteria

$$SSE = \Sigma_{i \in S_1} (y_i - \bar{y}_1)^2 + \Sigma_{i \in S_2} (y_i - \bar{y}_2)^2$$

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought	Tourist (T)
Visitor 1	Yes	Yes	4	Yes
Visitor 2	Yes	No	1	No
Visitor 3	No	No	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 5	No	No	2	No

Split >2:

$$S_1 = [1, 2], \bar{y}_1 = 1.5$$

$$S_2 = [3, 4, 4], \bar{y}_2 = 3.666$$

$$SSE = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 3.666)^2 + (3 - 3.666)^2 + (4 - 3.3666) + = 1.167$$

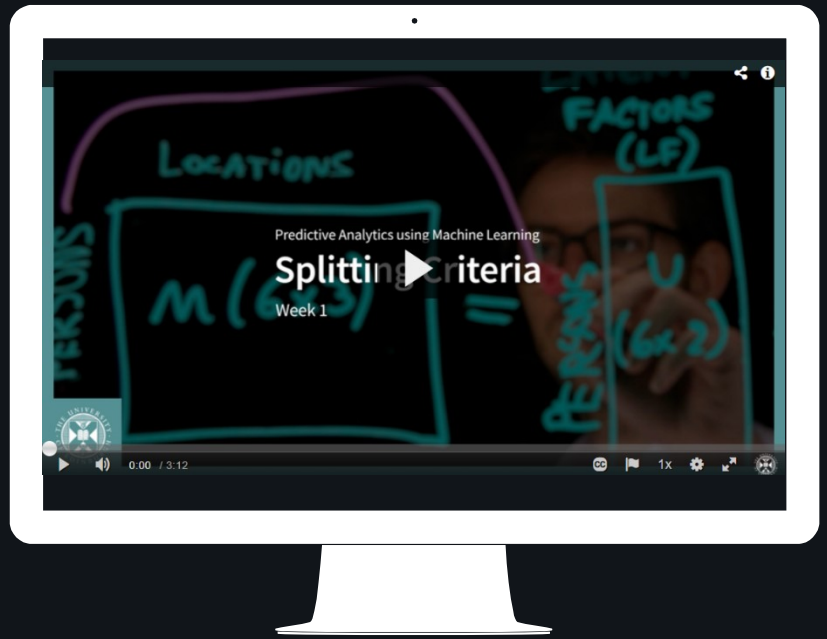
Split >3:

$$S_1 = [1, 2, 3], \bar{y}_1 = 2$$

$$S_2 = [4, 4], \bar{y}_2 = 4$$

$$SSE = (1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2 + (4 - 4)^2 + (4 - 4)^2 = 2$$

Using splitting criteria



Please watch the following video:

https://media.ed.ac.uk/media/Splitting+Criteria/1_nhttzl76/141757871

Quiz

Calculate the information gain if you split based on the number of bottles of whisky bought (>2).

- A. IG(T,N) = 1
- B. IG(T,N) = 0.0722
- C. IG(T,N) = 0.971
- D. IG(T,N) = 0.0996

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought (N)	Tourist (T)
Visitor 1	Yes	Yes	4	Yes
Visitor 2	Yes	No	1	No
Visitor 3	No	Yes	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 5	No	Yes	2	No
Visitor 6	Yes	Yes	1	Yes
Visitor 7	Yes	No	10	No
Visitor 8	No	No	5	Yes
Visitor 9	No	Yes	1	Yes
Visitor 10	No	Yes	0	No

Quiz

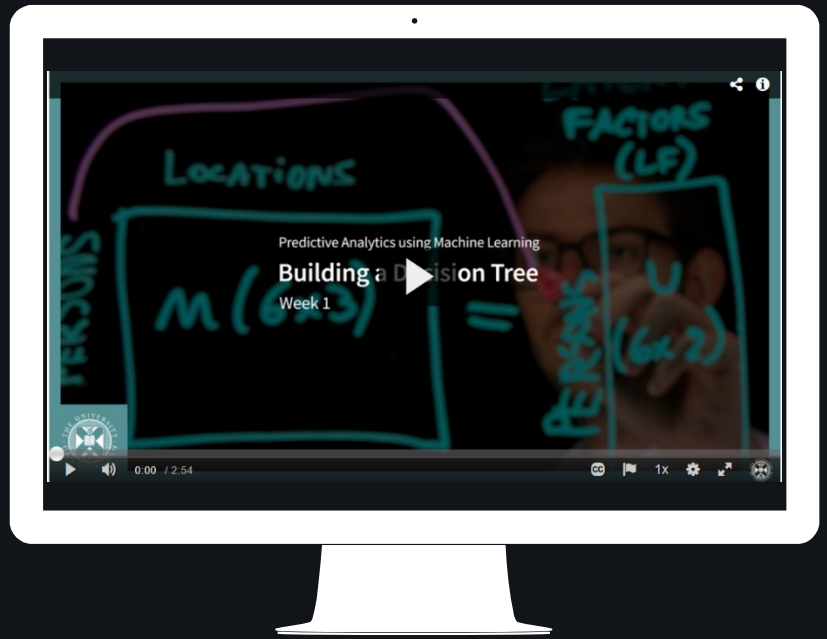
Calculate the Gini impurity after splitting based on N<2, N<4, N>=4.

- A. G(after N split) = 0.033
- B. G(after N split) = 0.025
- C. G(after N split) = 0.0583
- D. G(after N split) = 0.1

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought (N)	Tourist (T)
Visitor 1	Yes	Yes	4	Yes
Visitor 2	Yes	No	1	No
Visitor 3	No	Yes	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 5	No	Yes	2	No
Visitor 6	Yes	Yes	1	Yes
Visitor 7	Yes	No	10	No
Visitor 8	No	No	5	Yes
Visitor 9	No	Yes	1	Yes
Visitor 10	No	Yes	0	No

Building a Tree

Building a tree



Please watch the following video:

https://media.ed.ac.uk/media/Building+a+Decision+Tree/1_q4hb7lw3/141757871

Building a tree

- Straightforward solution:
  - Iteratively keep dividing up the dataset based on the best splits
  - Calculate splitting criteria for all variables and split until no further splits possible/nodes are pure

Building a tree

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought (N)	Tourist (T)
Visitor 1	Yes	Yes	4	Yes
Visitor 2	Yes	No	1	No
Visitor 3	No	Yes	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 5	No	Yes	2	No
Visitor 6	Yes	Yes	1	Yes
Visitor 7	Yes	No	10	No
Visitor 8	No	No	5	Yes
Visitor 9	No	Yes	1	Yes
Visitor 10	No	Yes	0	No

$$IG(T,V) = H(T) - \frac{4}{10}H(T|V = Yes) - \frac{6}{10}H(T|V = No)$$
$$= 0.971 - \frac{4}{10}1 - \frac{6}{10}0.92 = 0.02$$

$$IG(T,B) = H(T) - \frac{4}{10}H(T|B = Yes) - \frac{6}{10}H(T|B = No)$$
$$= 0.971 - \frac{7}{10}0.86 - \frac{3}{10}0.92 = 0.093$$

$$IG(T,N) = H(T) - \frac{5}{10}H(T|N > 2) - \frac{5}{10}H(T|N \leq 2)$$
$$= 0.971 - \frac{5}{10}0.722 - \frac{5}{10}0.971 = 0.125$$

Number of bottles of whisky bought

>2

<=2

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought (N)	Tourist (T)
Visitor 1	Yes	Yes	4	Yes
Visitor 3	No	Yes	3	Yes
Visitor 4	No	Yes	4	Yes
Visitor 7	Yes	No	10	No
Visitor 8	No	No	5	Yes

$$H(T) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.72$$

$$H(T|V = Yes) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$H(T|V = No) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$$H(T|B = Yes) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$$H(T|B = No) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$IG(T,V) = 0.72 - \frac{2}{5}1 = 0.32$$

$$IG(T,B) = 0.72 - \frac{2}{5}1 = 0.32$$

In case of tie, select variable with fewest distinct values

	Visited the Castle (V)	Bought a kilt (B)	#bottles of whisky bought (N)	Tourist (T)
Visitor 2	Yes	No	1	No
Visitor 5	No	Yes	2	No
Visitor 6	Yes	Yes	1	Yes
Visitor 9	No	Yes	1	Yes
Visitor 10	No	Yes	0	No

$$H(T) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$H(T|V = Yes) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$H(T|V = No) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$

$$H(T|B = Yes) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

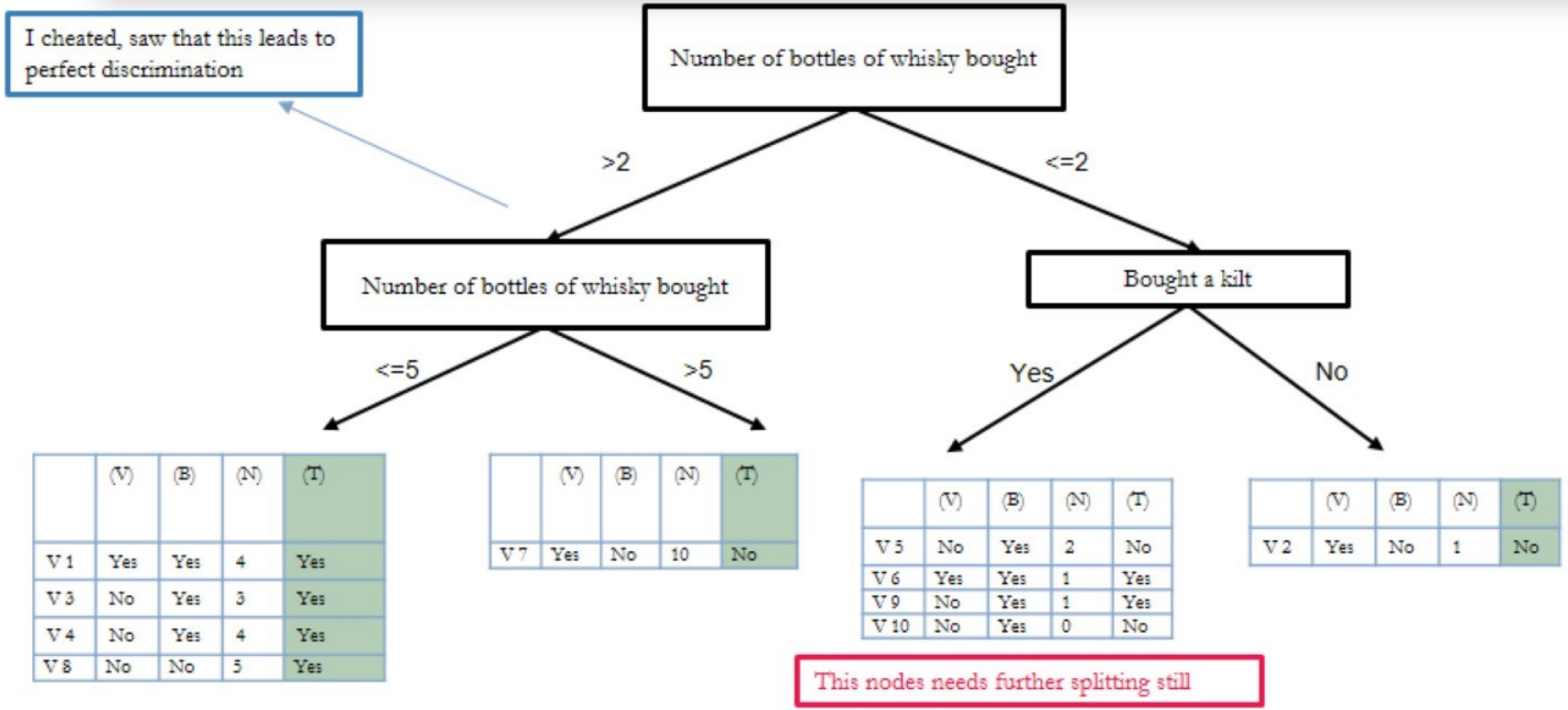
$$H(T|B = No) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} = 0$$

$$IG(T,V) = 0.971 - \frac{2}{5}1 - \frac{3}{5}0.918 = 0.02$$

$$IG(T,B) = 0.918 - \frac{4}{5}1 = 0.118$$

Result

I cheated, saw that this leads to perfect discrimination



	(V)	(B)	(N)	(T)
V 1	Yes	Yes	4	Yes
V 3	No	Yes	3	Yes
V 4	No	Yes	4	Yes
V 8	No	No	5	Yes

	(V)	(B)	(N)	(T)
V 7	Yes	No	10	No

	(V)	(B)	(N)	(T)
V 5	No	Yes	2	No
V 6	Yes	Yes	1	Yes
V 9	No	Yes	1	Yes
V 10	No	Yes	0	No

	(V)	(B)	(N)	(T)
V 2	Yes	No	1	No

This nodes needs further splitting still

Pruning

Pruning

Please watch the following video:

https://media.ed.ac.uk/media/Pruning/1_r9un41me/141757871

Overfitting

Trees become large very quickly, then tends to lead to overfitting

- Stop trees after growing particular size (#nodes/depth)

Alternative, pruning afterwards:

- Error reduce pruning
  - Replace splitting node with class of highest frequency
  - Check whether misclassification is (much) higher
  - Keep removing nodes higher up in the tree until misclassification cost becomes too high

- Cost complexity tuning

Cost complexity tuning (Breiman et al. 1984)

$$SSE_\alpha = SSE + \alpha \times (\#\ \text{Terminal Nodes}),$$

- The goal of this process is to find a "right-sized tree" that has the smallest error rate.
- α complexity parameter, penalises for large trees
- For a specific value of the complexity parameter, we find the smallest pruned tree that has the lowest penalized error rate
- http://mlwiki.org/index.php/Cost-Complexity_Pruning

Tree Algorithms

Algorithms: ID3

- Original one: Iterative Dichotomiser 3 (ID3)
  - Uses information gain
  - 'greedy' search
  - Only binary splits
  - Does not reuse variables
  - Stops when no variables are useable anymore or split is pure
- Downsides:
  - Overfits quickly
  - Does not work with numeric variables (without transformation)

Algorithms: C4.5

- Successor to ID3:
- Uses information gain ratio $GR(Z|K) = IG(Z|K)/H(Z|K)$
  - To avoid variables with many distinct variables to be selected more often
- Prunes trees after training, using error reduced pruning
- Handles both continuous and numeric variables by using thresholds for the latter
- Handles missing data
- Allows for more than 2 branches per split

Algorithms: CART

- Classification And Regression Trees (CART):
  - Most popular currently
  - Uses binary splits
  - Can handle both numeric/categorical variables
  - Can be used for regression as well
  - Uses Gini index
  - Applies cost-complexity tuning

Wrap-up – Decision trees

Benefits:

- Easy to construct
- Intuitive
- Easy to interpret
- Handle both continuous and categorical data
- Have built-in feature selection

Downsides:

- Overfit quickly
- Feature space always divided into rectangular subsets (above/below split)
- Correlating variables can cause problems
- Large trees are not easy to understand
- Node selection is deterministic (always the same)

- Please read the following Jupyter file
- 1 - Building decision trees for classification.ipynb + churn_ibm.csv
- And try the following exercise:
- Activity 1 - Build a decision tree for regression.ipynb

Activity: Decision trees for classification

Quiz

A decision tree's performance cannot suffer from...

- A. Having too many nodes.
- B. Having a high depth.
- C. Dividing the feature space into rectangular subspaces.
- D. All of the above.

Quiz

Which statement is false?

- A. The Gini index and entropy both capture the unstructuredness in the data.
- B. The Gini index captures statistical dispersion of the dataset.
- C. Information gain needs to be as low as possible, Gini impurity as high as possible.

UNIVERSITY OF EDINBURGH
Business School