UNIVERSITY OF EDINBURGH
Business School

# Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

**Dr Xuefei Lu**
The University of Edinburgh Business School

# 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 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 |

$$H(T) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.971$$

$$H(T|B > 2) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722$$

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

$$IG(T,N) = H(T) - \frac{4}{10}H(T|B > 2) - \frac{6}{10}H(T|B \leq 2)$$

$$= 0.971 - \frac{4}{10}0.722 - \frac{6}{10}0.971 = 0.0996$$

# 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 |

# 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 |

$$G(after\ V\ split) = p_t \cdot \left(2 \cdot p_{<2,t} \cdot p_{<4,t} \cdot p_{\leq 4,t}\right) + p_{nt} \cdot \left(2 \cdot p_{<2,nt} \cdot p_{<4,nt} \cdot p_{\leq 4,nt}\right)$$

$$= 0.6 \cdot (2 \cdot 2\backslash 6 \cdot 1\backslash 6 \cdot 3\backslash 6) + 0.4 \cdot \left(2 \cdot \frac{2}{4} \cdot \frac{1}{4} \cdot 1\backslash 4\right)$$

$$= 0.033 + 0.025 = 0.0583$$

# 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

**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.

# 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