Tutorial 2

Decision Tree, Cross-validation, Precision and Recall

Luke Chang

The University of Auckland

Mar. 2021

Objectives

1. Fully understand Decision Tree and able to explain it in your own words
2. Compute splitting criterion using Entropy and Information Gain
3. Fully understand Cross-Validation and know how to use it
4. Use built-in methods in `sklearn` to compute ROC and AUC

Decision Tree Overview

- A decision tree models different possible decision paths, where each decision node is a conditional.
- Decision nodes are created based on a **splitting criterion**.
- Most common splitting criterions are: **Entropy** and **Gini**.
- A tree is created recursively from the root node to the leaf nodes.
- Decision tree can be used for both classification and regression.

Entropy

Informally, entropy measures the amount of uncertainty in a system.

Shannon Entropy

$$H(X) := -\sum_x P(X = x) \log_2 P(X = x)$$

Where $0 \log_2(0) \equiv 0$, since $\lim_{x \to 0} x \log_2(x) = 0$.

- What is the possible maximum entropy?
- What does it mean?
- What is the possible minimum entropy?
- What does it mean?

Entropy

Informally, entropy measures the amount of uncertainty in a system.

Shannon Entropy

$$H(X) := -\sum_x P(X=x)\log_2 P(X=x)$$

Where $0\log_2(0) \equiv 0$, since $\lim_{x\to 0} x\log_2(x) = 0$.

- What is the possible maximum entropy? No upper bound
- What does it mean? No better than random guess.
- What is the possible minimum entropy? 0
- What does it mean? You are certain about the outcome. No uncertainty.

Example: A six-sided Die

Suppose we have a standard six-sided die with each of the faces has a probability of $1/6$.

$$P(X=i)=\frac{1}{6}, i \in \{1,2,3,4,5,6\}$$

Entropy:

$$H(X)=-\sum \frac{1}{6}\log_2(\frac{1}{6})=-6[\frac{1}{6}\log_2(\frac{1}{6})]=\log_2(6)\approx 2.6$$

Suppose we have a loaded six-sided die. All 6 faces are printed “1”.

$$P(X=i)=\begin{cases}1 & \text{if } i=1,\\ 0 & otherwise.\end{cases}$$

Entropy:

$$H(X)=-log_2(1)=0$$

Information Gain

Information Gain

Information Gain = Parent Entropy - Current Conditional Entropy

$$\mathsf{IG}(Y,X) := H(Y) - H(Y|X)$$

Where $H(Y|X)$ is the conditional entropy of the target variable Y given attribute X (Weighted sum):

$$H(Y|X) := \sum_x P(X=x)H(Y|X=x)$$

And $H(Y|X=x)$ is the conditional entropy of the target variable Y given $X=x$:

$$H(Y|X=x) := -\sum_y P(Y=y|X=x)\log_2 P(Y=y|X=x)$$

Estimating Probabilities

Let D_n be the subset of the data at node n in the tree, we estimate the probability by counting the number of success:

$$P(Y=y) := \frac{|\{i \in D_n : i_Y = y\}|}{|D_n|}$$

Similarly:

$$P(X=x) := \frac{|\{i \in D_n : i_X = x\}|}{|D_n|}$$

Conditional probability:

$$P(Y=y|X=x) := \frac{|\{i \in D_n : i_Y = y \wedge i_X = x\}|}{|\{i \in D_n : i_X = x\}|}$$

Example 1: Boolean Functions

Give decision trees to represent the following boolean functions:

1. $A \wedge \neg B$
2. $A \vee (B \wedge C)$
3. $A \oplus B$ (XOR)

Question 1: $A \wedge \neg B$

A	B	¬B	Y
0	0	1	0
1	0	1	1
0	1	0	0
1	1	0	0

$A \wedge \neg B$ Probabilities

A	B	¬B	Y
0	0	1	0
1	0	1	1
0	1	0	0
1	1	0	0

$$P(Y=1) = \frac{1}{4}$$

$$P(Y=1|A=1) = \frac{1}{2}$$

$$P(Y=1|A=0) = 0$$

$$P(Y=1|B=1) = 0$$

$$P(Y=1|B=0) = \frac{1}{2}$$

Root Entropy:

$$\begin{aligned} H(Y) &= -P(Y=1)\log_2 P(Y=1) - P(Y=0)\log_2 P(Y=0) \\ &= -\frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4}) \\ &\approx 0.5 + 0.31 = 0.81 \end{aligned}$$

Entropies Condition on A

$$
\begin{aligned}
H(Y|A=1) &= -P(Y=1|A=1)\log_2 P(Y=1|A=1) - P(Y=0|A=1)\log_2 P(Y=0|A=1) \\
&= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) \\
&= 1
\end{aligned}
$$

$$
\begin{aligned}
H(Y|A=0) &= -P(Y=1|A=0)\log_2 P(Y=1|A=0) - P(Y=0|A=0)\log_2 P(Y=0|A=0) \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
IG(Y|A) &= H(Y) - P(A=1)H(Y|A=1) - P(A=0)H(Y|A=0) \\
&\approx 0.81 - \frac{2}{4} - 0 = 0.31
\end{aligned}
$$

Entropies Condition on B

$$\begin{aligned} H(Y|B=1) &= -P(Y=1|B=1)\log_2 P(Y=1|B=1) - P(Y=0|B=1)\log_2 P(Y=0|B=1) \\ &= 0 \end{aligned}$$

$$\begin{aligned} H(Y|B=0) &= -P(Y=1|B=0)\log_2 P(Y=1|B=0) - P(Y=0|B=0)\log_2 P(Y=0|B=0) \\ &= 1 \end{aligned}$$

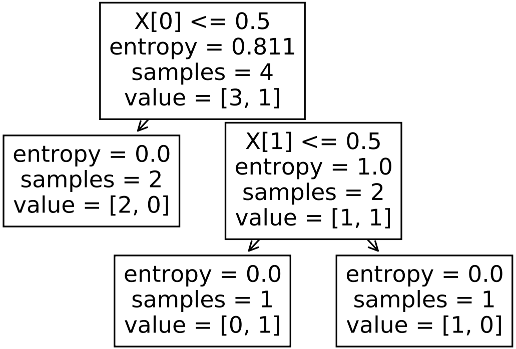
$$\begin{aligned} IG(Y|B) &= H(Y) - P(B=1)H(Y|B=1) - P(B=0)H(Y|B=0) \\ &\approx 0.81 - 0 - \frac{2}{4} = 0.31 \end{aligned}$$

Information Gain

Draw Decision Tree using `sklearn`

$$IG(Y|A) \approx 0.31$$
$$IG(Y|B) \approx 0.31$$

There is a tie on both conditions. Let's use A as root node.



Example 2: Questions from last week

Colour	Length	Size	Brightness	Shape	Class
red	long	larger	bright	triangle	TRUE
red	long	small	bright	circle	FALSE
red	long	small	bright	triangle	TRUE
red	short	larger	dull	circle	FALSE
red	short	larger	bright	triangle	TRUE
blue	short	larger	bright	triangle	FALSE

Without prior condition:

$$
\begin{aligned}
H(Y) &= -P(Y=1)\log_2 P(Y=1) - P(Y=0)\log_2 P(Y=0) \\
&= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) \\
&= 0.5 + 0.5 = 1
\end{aligned}
$$

Entropies

Colour	Length	Size	Brightness	Shape	Class
red	long	larger	bright	triangle	TRUE
red	long	small	bright	circle	FALSE
red	long	small	bright	triangle	TRUE
red	short	larger	dull	circle	FALSE
red	short	larger	bright	triangle	TRUE
blue	short	larger	bright	triangle	FALSE

Condition on *Colour*:

$$H(Y|\text{Colour} = \text{red}) = -\frac{3}{5}log_2(\frac{3}{5}) - \frac{2}{5}log_2(\frac{2}{5}) \approx 0.971$$

$$H(Y|\text{Colour} = \text{blue}) = 0$$

$$IG(Y|\text{Colour}) \approx 1 - \frac{5}{6}(0.971) - \frac{1}{6}(0) = 0.191$$

Entropies

Condition on *Length*:

$$H(Y|\text{Length} = \text{long}) = -\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) \approx 0.918$$

$$H(Y|\text{Length} = \text{short}) = -\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3}) \approx 0.918$$

$$IG(Y|\text{Length}) \approx 1 - \frac{3}{6}(0.918) - \frac{3}{6}(0.918) = 0.082$$

Condition on *Size*:

$$H(Y|\text{Size} = \text{large}) = -\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4}) = 1$$

$$H(Y|\text{Size} = \text{small}) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$

$$IG(Y|\text{Size}) = 1 - \frac{4}{6} - \frac{2}{6} = 0$$

Entropies

Condition on *Brightness*:

$$H(Y|\text{Brightness} = \text{dull}) = -\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$$

$$H(Y|\text{Brightness} = \text{bright}) = -\frac{3}{5}log_2(\frac{3}{5}) - \frac{2}{5}log_2(\frac{2}{5}) \approx 0.971$$

$$IG(Y|\text{Brightness}) \approx 1 - \frac{1}{6}(0) - \frac{5}{6}(0.971) = 0.191$$

Condition on *Shape*:

$$H(Y|\text{Shape} = \text{triangle}) = -\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4}) \approx 0.821$$

$$H(Y|\text{Shape} = \text{circle}) = -\frac{0}{2}log_2(\frac{0}{2}) - \frac{2}{2}log_2(\frac{2}{2}) = 0$$

$$IG(Y|\text{Shape}) = 1 - \frac{4}{6}(0.821) - \frac{2}{6}(0) = 0.453$$

Entropies

Attribute	Information Gain
Colour	0.191
Length	0.082
Size	0
Brightness	0.191
Shape	0.453

Shape has the largest IG. It is the top of the tree.
Circle branch is pure, so it is a leaf.
triangle must recurse.

Shape (Root)

circle (Leaf) triangle

Splitting on Shape

Colour	Length	Size	Brightness	Shape	Class
red	long	larger	bright	triangle	TRUE
red	long	small	bright	circle	FALSE
red	long	small	bright	triangle	TRUE
red	short	larger	dull	circle	FALSE
red	short	larger	bright	triangle	TRUE
blue	short	larger	bright	triangle	FALSE

Given Shape = triangle, condition on *Colour*:

$$H(Y|\text{Shape} = \text{triangle}) \approx 0.821$$

$$H(Y|\text{Colour} = \text{red}, \text{Shape} = \text{triangle}) = -\frac{3}{3}log_2(\frac{3}{3}) - \frac{0}{3}log_2(\frac{0}{3}) = 0$$

$$H(Y|\text{Colour} = \text{blue}, \text{Shape} = \text{triangle}) = -\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1}) = 0$$

$$IG(Y|\text{Colour}, \text{Shape} = \text{triangle}) \approx 0.821 - 0 - 0 = 0.821$$

Splitting on Shape

$$H(Y|\text{Shape} = \text{triangle}) \approx 0.821$$

Given Shape = triangle, condition on *Length*:

$$H(Y|\text{Length} = \text{long}, \text{Shape} = \text{triangle}) = -\frac{2}{2}log_2(\frac{2}{2}) - \frac{0}{2}log_2(\frac{0}{2}) = 0$$

$$H(Y|\text{Length} = \text{short}, \text{Shape} = \text{triangle}) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$

$$IG(Y|\text{Length}, \text{Shape} = \text{triangle}) \approx 0.821 - 0 - \frac{2}{4} = 0.321$$

Given Shape = triangle, condition on *Size*:

$$H(Y|\text{Size} = \text{larger}, \text{Shape} = \text{triangle}) = -\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) = 0.918$$

$$H(Y|\text{Size} = \text{small}, \text{Shape} = \text{triangle}) = -\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$$

$$IG(Y|\text{Size}, \text{Shape} = \text{triangle}) \approx 0.821 - \frac{3}{4}(0.918) - 0 \approx 0.132$$

Splitting on Shape

$$H(Y|\text{Shape} = \text{triangle}) \approx 0.821$$

Given Shape $=$ triangle, condition on *Brightness*:

$$H(Y|\text{Brightness} = \text{dull}, \text{Shape} = \text{triangle}) = -\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4}) = 0.821$$

$$H(Y|\text{Brightness} = \text{bright}, \text{Shape} = \text{triangle}) = 0$$

$$IG(Y|\text{Brightness}, \text{Shape} = \text{triangle}) \approx 0.821 - \frac{4}{4}(0.821) - 0 = 0$$

Splitting on Colour

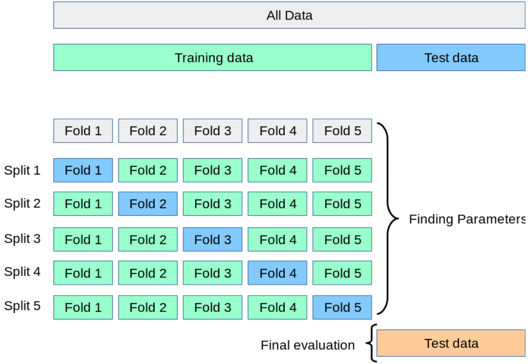
Attribute	Information Gain
Colour	0.821
Length	0.321
Size	0.132
Brightness	0

The next node is *Colour*.
Blue branch is pure and is a leaf.
Red branch is pure and is a leaf.
Recursion stops. The entropies of leaf nodes are 0.

- Shape (Root)
  - circle (Leaf)
  - triangle
    - Colour
      - blue(Leaf)
      - red(Leaf)

Cross-Validation (CV)

- K-folds Cross-validation uses $k-1$ of the folds as training data.
- The performance measure reported by CV is then the average of the values computed in the loop.
- CV should be used only for finding hyper-parameters.
- CV does not shuffle the test set.
- Once we obtained the optimal hyper-parameters, we retrain the model with the entire training set.



Code Demos

Check notebook `tutorial_02.ipynb` for code examples.