Introduction to Decision Trees, Background and Application….Ensemble Overview

Brian Wright

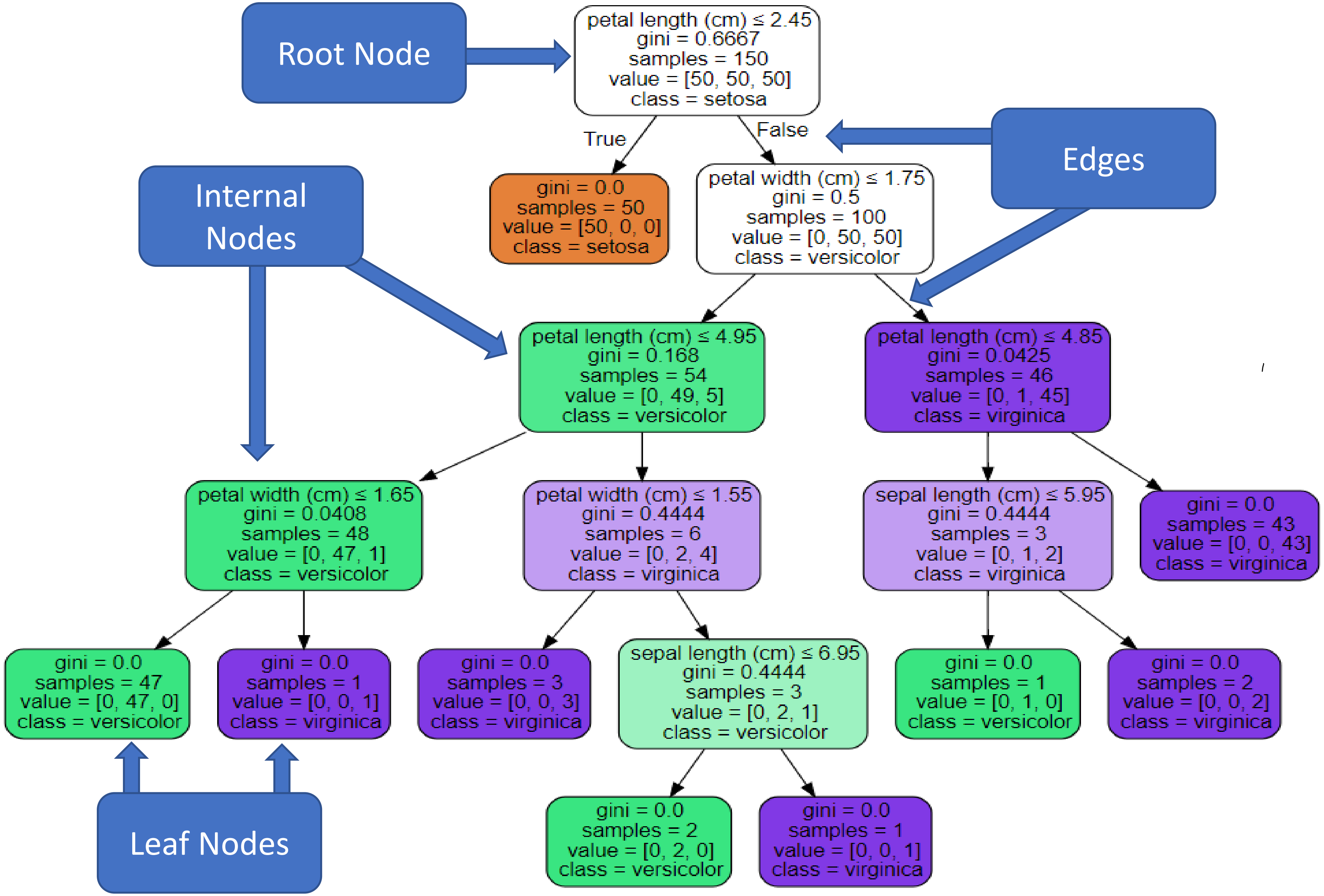
Outline

- Decision Trees
  - Basics
  - Background
  - Advantages and Limitations
  - Mathematical Approaches and Example
  - Example in R
  - Evaluation

Basics: Graph Elements

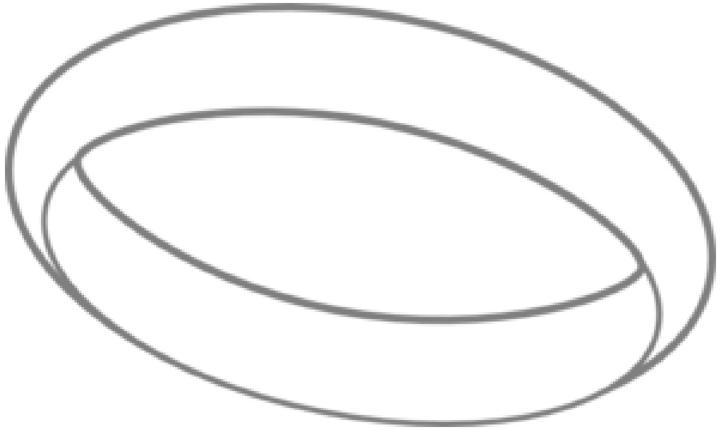
- Tree begins with a **Root Node** that has no incoming edges and two or more out going edges
- **Internal Node** – Has one incoming edge and two or more outgoing and represent test conditions at every given level
- **Leaf Node** – One incoming edge and no outgoing edges
- **Edges –** Connections between nodes

Basics: Graph Example



Basics: Intuition

1. What is the most important question to move on to a second date?
   The question with the most amount of relevant information.



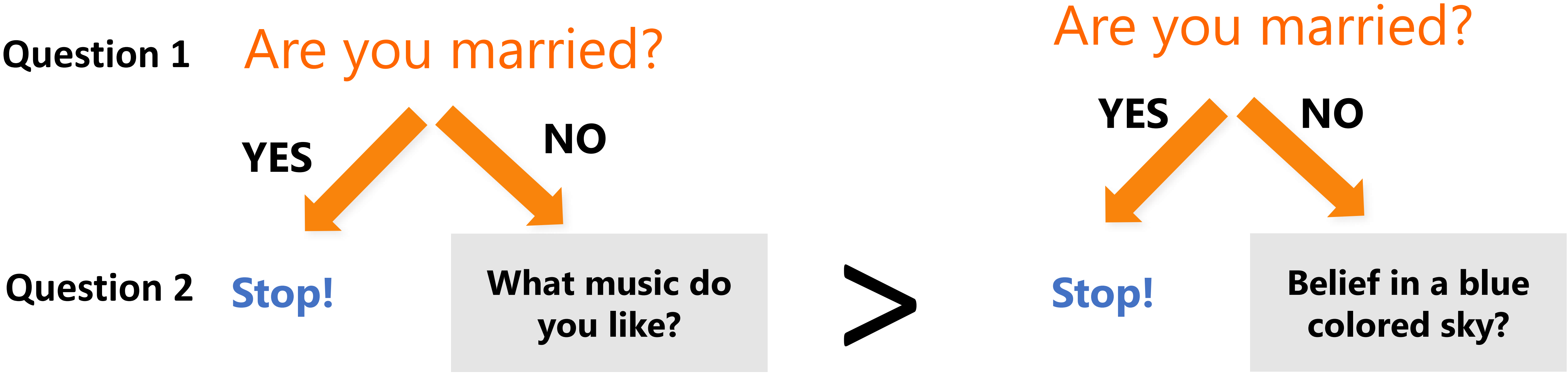
Are you married?

>

What music do you like?

Basics: Intuition

2. How do you combine questions?
   Conditional on the first answer - select the next most important question for information gain.



Question 1

Are you married?

YES

NO

Question 2

Stop!

What music do you like?

>

Are you married?

YES

NO

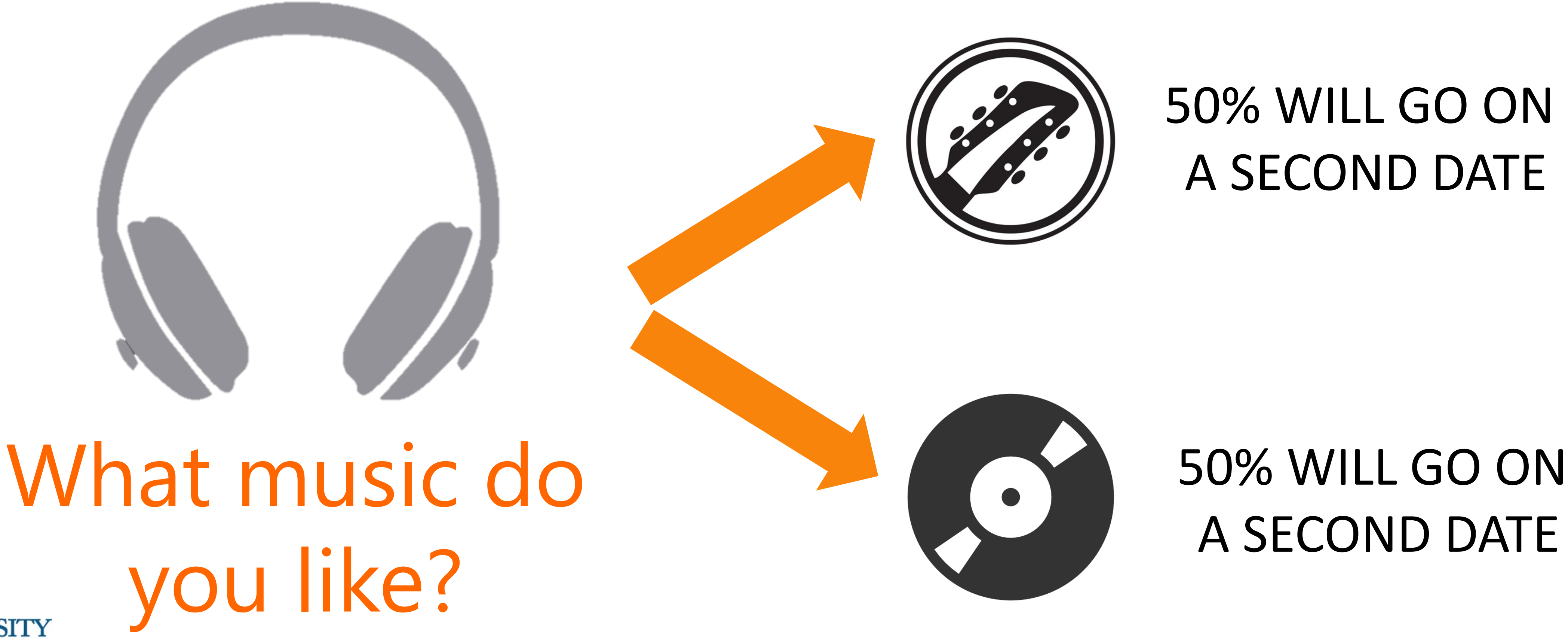
Stop!

Belief in a blue colored sky?

UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE

Basics: Intuition

3. When should you stop asking questions?
   When the answer no longer provides additional relevant information.



What music do you like?

50% WILL GO ON A SECOND DATE

50% WILL GO ON A SECOND DATE

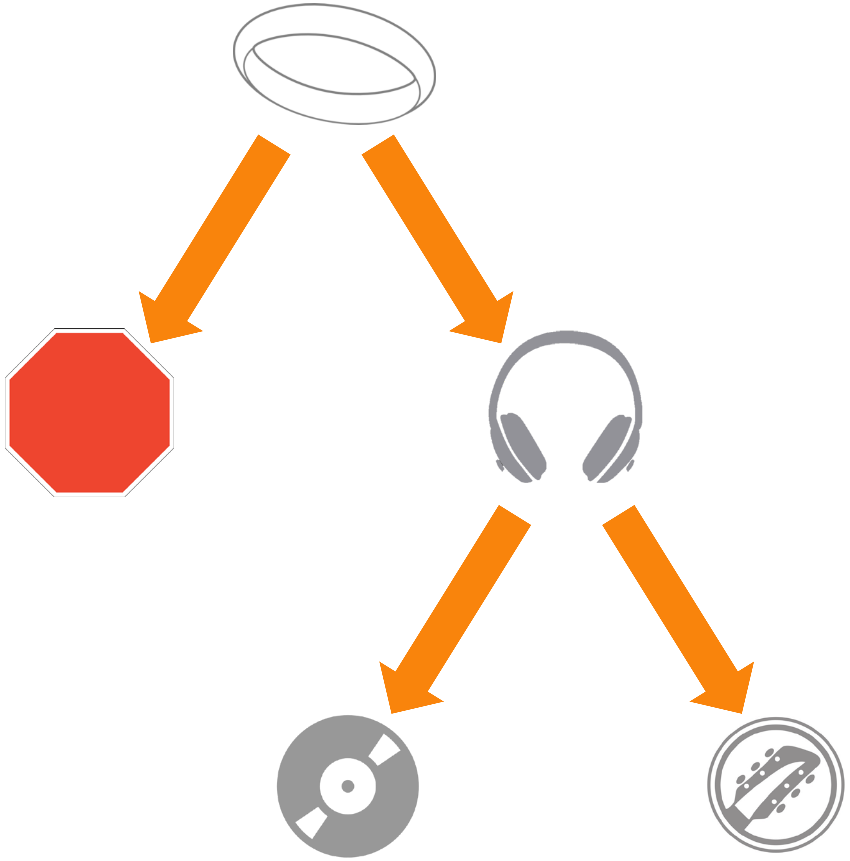
Basics: Building a tree in four steps

- **Step 1:** Ask the question with the most amount of information, where "most amount of information" is based on some objective criteria.

- **Step 2:** Conditional on the first answer, select the next most important question.

- **Step 3:** When the answer no longer provides additional information (no information gain), stop growing the branch.

- **Step 4:** Repeat steps 2 and 3 for each question branch.



UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE

Basics

- Decision trees are a hierarchical technique
  - Meaning that a series of decisions are made until a predetermined metric is met
    - Model is built such that a sequence of ordered decisions concerning values of data features results in assigning **class labels**

- Nonparametric
  - Number of parameters is not pre-determined as is the case with linear models that have pre-determine parameters thus limiting their **degrees of freedom**
  - No assumptions need to be met concerning parameters or distributions

- Best recognized through graphs produced
  - Type of Acyclic graph - are used to model probabilities, connectivity, and causality. A “graph” in this sense means a structure made from nodes and edges
  - Trees consist of nodes and edges defined by decisions rules applied to the data features

Background

Source: https://www.thehindu.com/features/friday-review/where-sanskrit-meets-computer-science/article7061379.ece

Background

- Uses **recursive binary splitting** - Considering every possible partition of space is computationally infeasible, a **greedy** approach is used to divide the space.

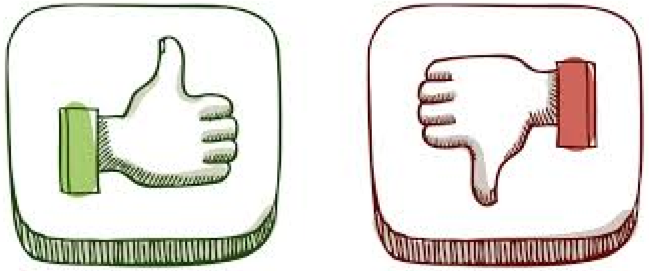
- **Greedy algorithm** because at each step of the tree building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in the future.

- Trees can be regression or classification based, but in both instances will use **recursive binary splitting**
  - The difference is that in regression based trees we are predicting the actual class whereas in classification we are generating the probability of class inclusion as the determinate of the splitting
  - This probability measure that drives the splitting for classification comes in two forms: Gini Index or Entropy

Background: CART Algorithm and C4.5

- Classification and Regression Tree (CART) – 1984 Breiman, Friedman, Olshen and Stone – Binary Trees
  - Can be used on numerical or categorical data
  - First splits the training data in two subsets using a single feature k and a threshold t_k
  - Searches through all possible pairs (k, t_k) to identify the split that produces the purest subsets, based on weighted average of information gain.
  - Stops once it cannot find a split that reduces impurity or by a pre-determine node size (hyperparameter).
- C4.5 – Grew out of ID3 (early version) in the late 1980s early 90s both from J. Ross Quinlan, uses gain ratio, accepts cont. and discrete, introduced pruning and the application of different weights to variables
- C5.0 – Next version of C4.5 – performance improvements, computationally more efficient and allows for boosting

Advantages and Limitations



Source: https://socialmediamanager.tweetinggoddess.com/2018/04/20/advantages-and-disadvantages-of-twitter/

Advantages

- Simple to understand and to interpret through visualization.

- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.

- Able to handle both numerical and categorical data.

- Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. By contrast, in a black box model (some neural network approaches), results may be more difficult to interpret.

- Fairly straight forward to evaluate and understand reliability of the model. ROC/Hit Rate/Error Rate/

Limitations

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called **overfitting**.
  - Compare terminal nodes to data points, use the depth of the tree to calculate terminal nodes, for example 6 levels = 2^6 or 64 terminal nodes, if you have 100 data points that’s a lot of single data terminal nodes. Leaf nodes roughly double with every additional level of the tree.
  - Mechanisms such setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree can be used to avoid this problem.
- Decision trees can be unstable as small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an **ensemble** like Random Forest.

UNIVERSITY *of* VIRGINIA
DATA SCIENCE INSTITUTE

Limitations

- Practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node.
    - Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.

- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to **balance** the dataset prior to fitting if necessary

Mathematical Approaches and Examples

Mathematical Approaches: Node split criterion

- Decision Trees can use several different types of node split criteria depending on the data or data scientist’s preference
  - Regression/MSE – Continuous data
  - Entropy – Binary data splits
  - Gini Coefficient – Most common approach
- Let’s take a look at each approach

Both Entropy and Gini Coefficient use Information Gain to determine variable split criteria

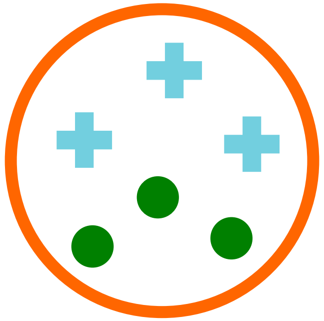
Tree Based Methods

- Several approaches to tree building
  - CART – Gini Index – Binary Trees
  - ID3 – Information Gain
  - C4.5 – Gains Ratio – Introduced pruning
  - C5.0 – Improvement on C4.5 – Boosting, computationally efficient

Mathematical Approaches: Classification, Entropy (C4.5)

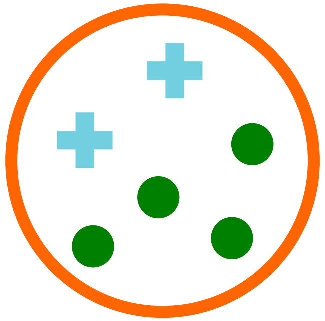
- The formula for entropy is below, where Pi is the probability that a random selection would have a state i

$$\text{Entropy} = \text{sum}(-P_i * \log_2 P_i)$$



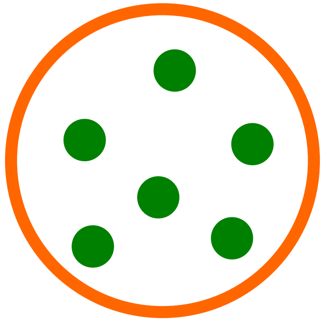
✚ = 3

● = 3



✚ = 2

● = 4



✚ = 0

● = 6

$(3/6 \log_2 3/6) – (3/6 \log_2 3/6) = 1$

$(2/6 \log_2 2/6) - (4/6 \log_2 4/6) = 0.92$

$(6/6 \log_2 6/6) = 0$

Mathematical Approaches: Classification, Entropy

- Information gain helps us understand how important an attribute is in the data
- We can use it to decide how to order the nodes of the decision tree

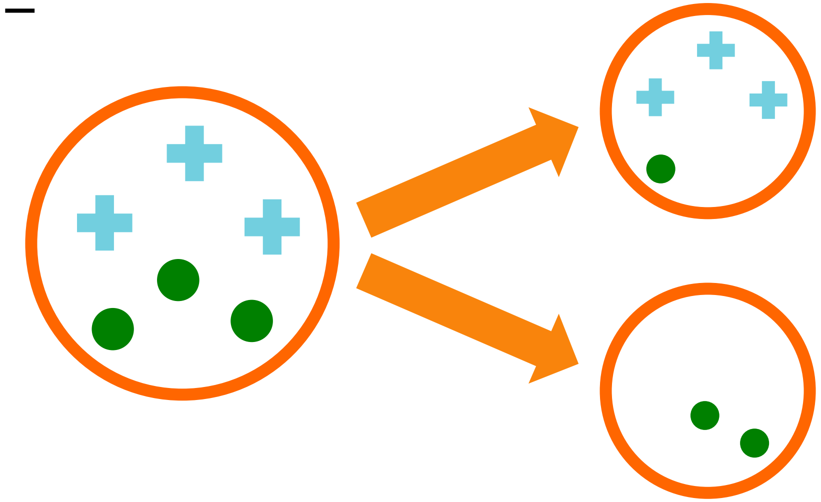
Information gain = entropy (parent) – average entropy (children)

entropy (parent)

$-(3/6 \log_2 3/6) - (3/6 \log_2 3/6) =$

1

average entropy (children)



$-(3/4 \log_2 3/4) - (1/4 \log_2 1/4)$

$= 0.81$

$-(2/2 \log_2 2/2)$

$= 0$

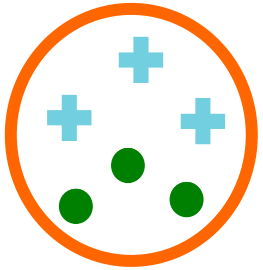
Mathematical Approaches: Classification, Entropy

- In order to calculate the average entropy for the split, we need to weigh the split by the number of data points in each node. So we create a weighted average of the entropy of the children nodes.

Information gain (ratio) = entropy (parent) – average entropy (children)

entropy (parent)

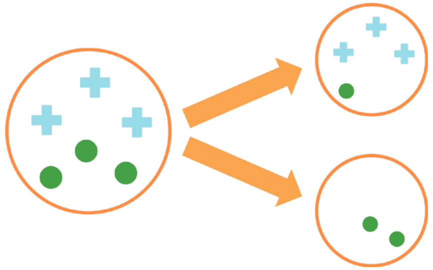
= 1



Weighted average entropy (children)

= (2/6 * 0) + (4/6 * 0.81)

= 0.54



UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE

Mathematical Approaches: Information gain + entropy: example

- In order to construct the tree, we need to follow three steps:

1. Choose the attribute with the highest information gain
2. Construct the child nodes
3. Repeat steps 1 and 2 recursively until no more information can be gained

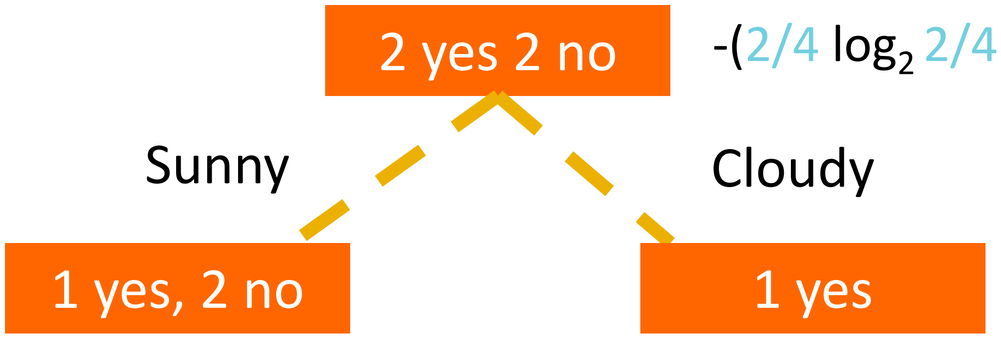
Should you play outside?

Outlook	Temp	Humidity	Play
Sunny	Hot	High	No
Sunny	Hot	Low	No
Cloudy	Cool	High	Yes
Sunny	Cool	Low	Yes

UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE

Mathematical Approaches: Information gain + entropy: example

1. Choose the attribute with the highest information gain



$-(2/4 \log_2 2/4) - (2/4 \log_2 2/4) = 1$

$-(1/3 \log_2 1/3) - (2/3 \log_2 2/3)$
$= 0.92$

$-(1/1 \log_2 1/1) = 0$

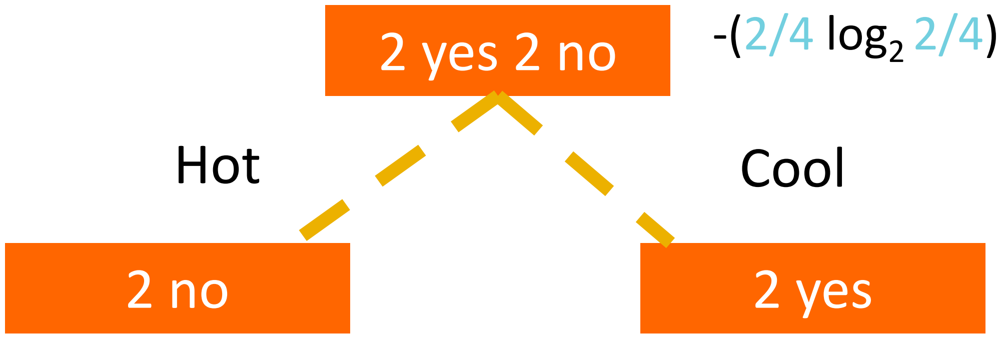
$$\text{I.G.} = 1 - \overbrace{((3/4 * 0.92) + (1/4 * 0))}^{\textit{weighted average}} = 0.31$$

Should you play outside?

Outlook	Temp	Humidity	Play
Sunny	Hot	High	No
Sunny	Hot	Low	No
Cloudy	Cool	High	Yes
Sunny	Cool	Low	Yes

Mathematical Approaches: Information gain + entropy: example

1. Choose the attribute with the highest information gain



2 yes 2 no

Hot

Cool

2 no

2 yes

$$-(2/4 \log_2 2/4) - (2/4 \log_2 2/4) = 1$$

$$-(2/2 \log_2 2/2) = 0$$

$$-(2/2 \log_2 2/2) = 0$$

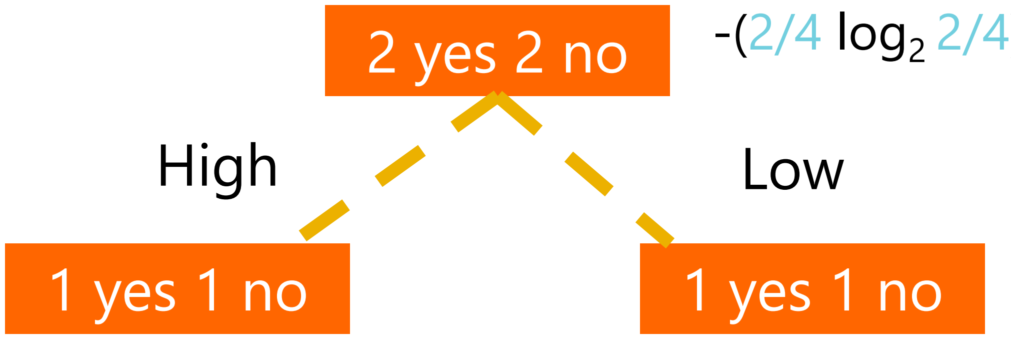
$$I.G. = 1 - \overbrace{((2/4 * 0) + (2/4 * 0))}^{\textit{weighted average}} = 1$$

Should you play outside?

	Outlook	**Temp**	**Humidity**	**Play**
	Sunny	Hot	High	No
	Sunny	Hot	Low	No
	Cloudy	Cool	High	Yes
	Sunny	Cool	Low	Yes
I.G.	0.31			

Mathematical Approaches: Information gain + entropy: example

1. Choose the attribute with the highest information gain



2 yes 2 no

High

Low

1 yes 1 no

1 yes 1 no

$-(2/4 \log_2 2/4) - (2/4 \log_2 2/4) = 1$

$-(1/2 \log_2 1/2) - (1/2 \log_2 1/2) = 1$ $\quad -(1/2 \log_2 1/2) - (1/2 \log_2 1/2) = 1$

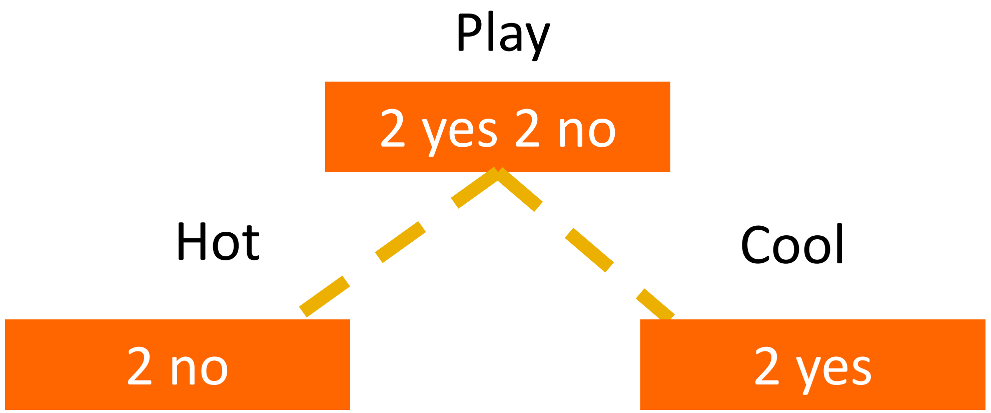
$$\text{I.G.} = 1 - \overbrace{((2/4 * 1) + (2/4 * 1))}^{\textit{weighted average}} = 0$$

Should you play outside?

	Outlook	Temp	Humidity	Play
	Sunny	Hot	High	No
	Sunny	Hot	Low	No
	Cloudy	Cool	High	Yes
	Sunny	Cool	Low	Yes
I.G.	0.31	1		

Mathematical Approaches: Information gain + entropy: example

- Temp has the highest information gain resulting in an entropy of 1, meaning that this attribute perfectly matches class prediction



Should you play outside?

Outlook	Temp	Humidity	Play
Sunny	Hot	High	No
Sunny	Hot	Low	No
Cloudy	Cool	High	Yes
Sunny	Cool	Low	Yes

I.G.	0.31	1	0

UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE

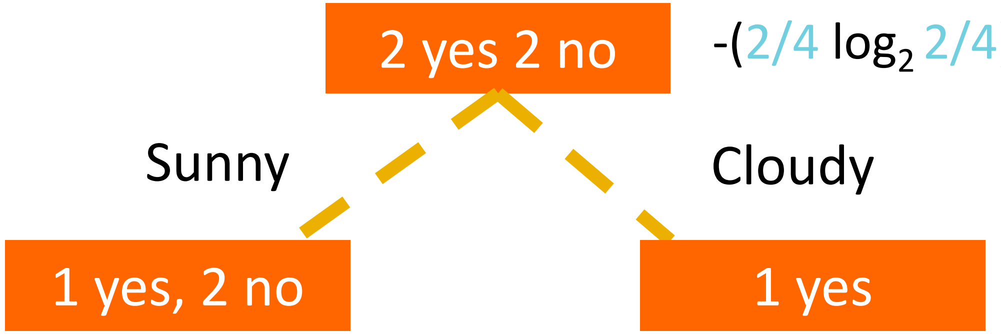
C4.5/C5.0 and Gains Ratio

C4.5/C5.0 – Uses Gain Ratio that extends the previous example but dividing the information gain by the overall split ratio for the feature

$$G_r(S,A) = \underset{\substack{\Big\uparrow \\ \text{Information} \\ \text{Gain}}}{G(S,A)} / \underset{\substack{\Big\uparrow \\ \text{Split} \\ \text{Information}}}{\mathit{Split}(S,A)}$$

Mathematical Approaches: Information gain + entropy + gains ratio (C4.5)

1. Choose the attribute with the highest info gain ratio



$-(2/4 \log_2 2/4) - (2/4 \log_2 2/4) = 1$

$-(1/3 \log_2 1/3) - (2/3 \log_2 2/3)$

$= 0.92$

$-(1/1 \log_2 1/1) = 0$

Should you play outside?

Outlook	Temp	Humidity	Play
Sunny	Hot	High	No
Sunny	Hot	Low	No
Cloudy	Cool	High	Yes
Sunny	Cool	Low	Yes

$$\text{I.G.} = 1 - \overbrace{((3/4 * 0.92) + (1/4 * 0))}^{\textit{weighted average}} = 0.31$$

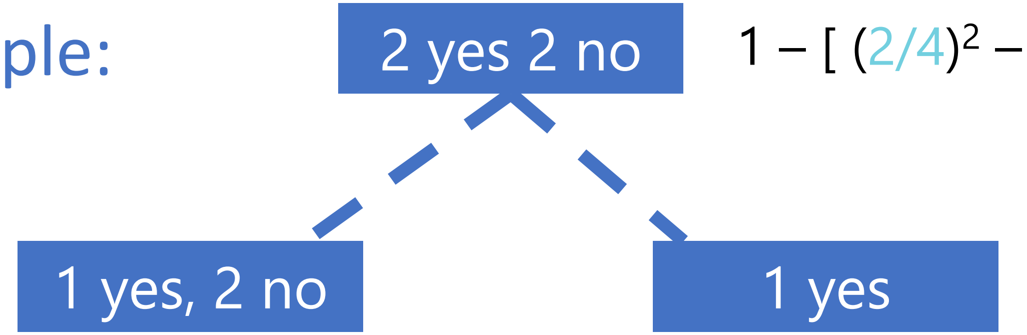
$$\text{Split} = -\,3/4\log_2 3/4 - 1/4\log_2 1/4 = .811$$

$$\text{G Ratio} = .31/.811 = .38$$

UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE

Mathematical Approaches: Classification, Gini Coefficient (CART)

- Gini
  - Gini Impurity = 1 – sum[$(Pi)^2$]
  - Pi – Represents the probability that a random selection would have state i (kinda like a target)
  - Same mathematical process as entropy
- Example:



$1 – [(2/4)^2 – (2/4)^2] = 0.5$

$1 – [(1/3)^2 + (2/3)^2] = 0.44$

$1 – [(1/1)^2] = 0$

$$\text{Gini Impurity} = 0.5 – \overbrace{((3/4 * 0.44) + (1/4 * 0))}^{\text{weighted average}} = 0.19$$

Mathematical Approaches: Regression/MSE

- Works to identify the split point in the data set that minimizes **mean squared error (MSE)** point
- The average of each of the groups is the term that minimizes the mean squared error
  - MSE – is the average of the difference between the prediction and actual values

$$\sqrt{\frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n}}$$

- The decision tree algorithm searches through all variables and all possible split points to identify the point that minimizes error

Practice…Poll

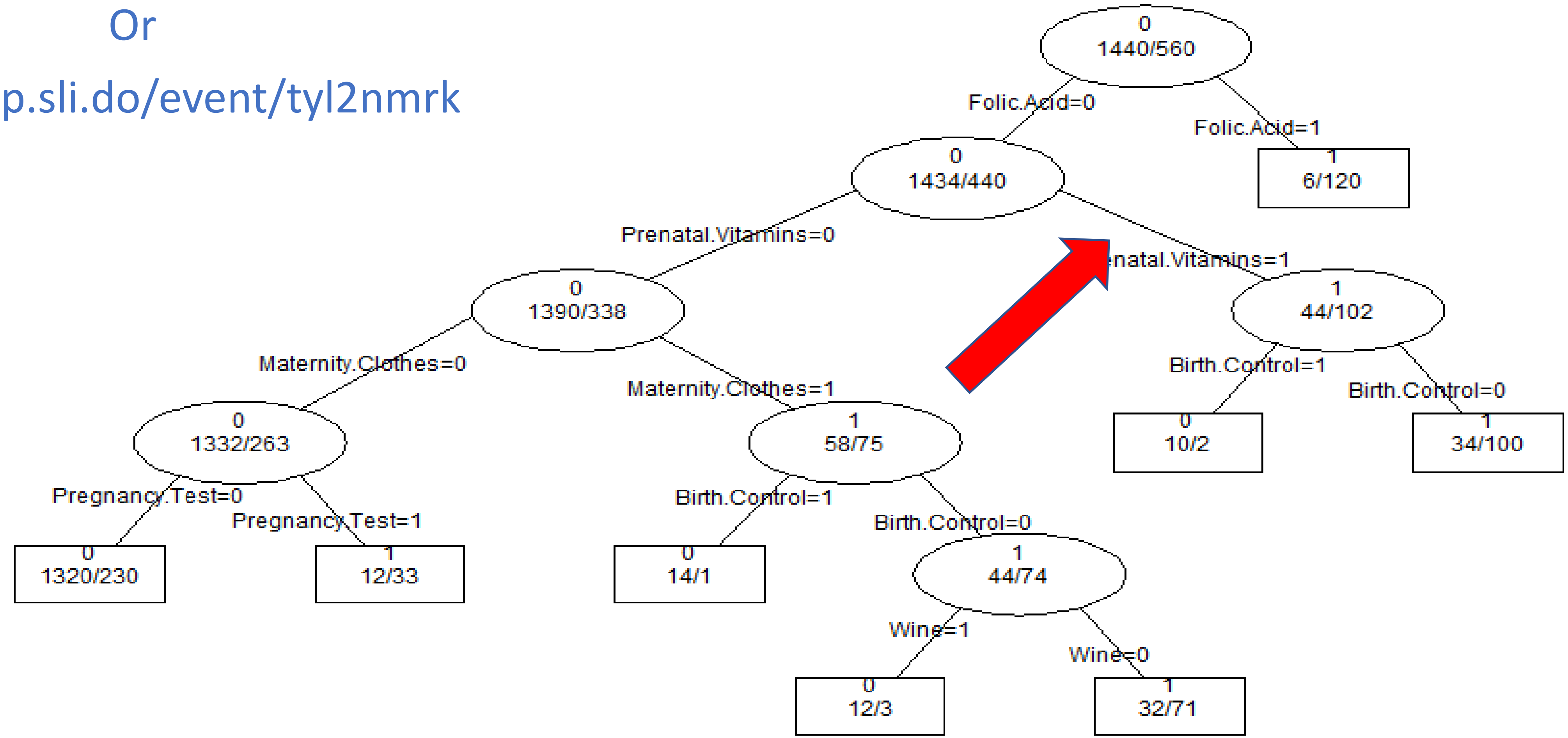
https://www.sli.do/

Use # 68881

Or

https://app.sli.do/event/tyl2nmrk

Tree for Pregnancy - gini



Decision Trees: Overfitting and Hyper-parameters

- Decision trees are often prone to overfitting, one solution is to utilize the hyper-parameters to control how the tree grows

- Another option is to use an ensemble method via bagging or what's known as Random Forest

Decision Trees: Hyper-parameter tuning (Pruning)

- **Minimum samples for a node split** Minimum number of samples (or observations) which are required in a node to be considered for splitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample. It should be tuned using cross validation.

- **Minimum samples for a terminal node (leaf)** The minimum number of samples (or observations) required in a terminal node or leaf. For imbalanced class problems, a lower value should be used since regions dominant with samples belonging to minority class will be much smaller in number.

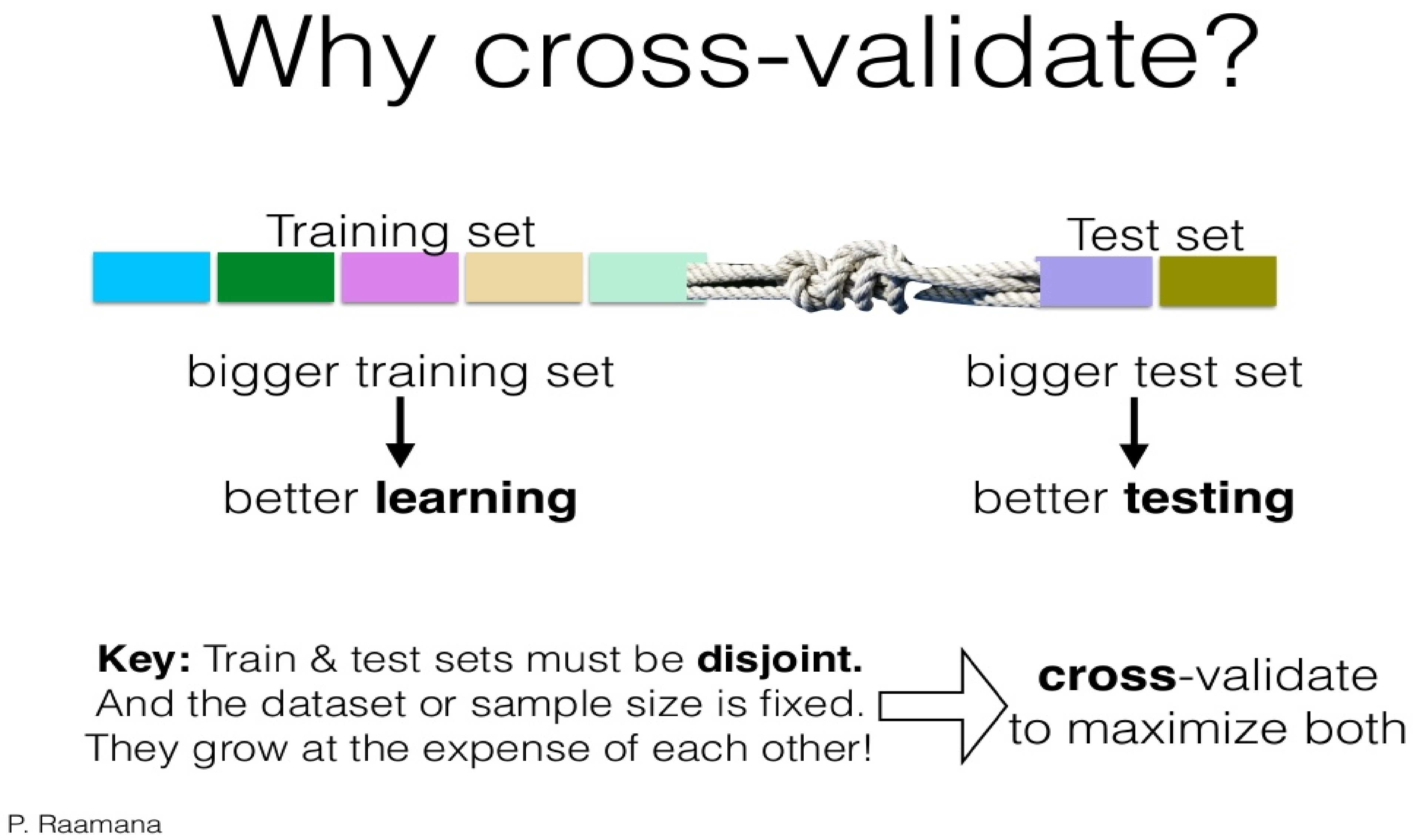
- **Maximum depth of tree (vertical depth)** The maximum depth of trees, lower values prevent a model from learning relations which might be highly specific to the particular sample. It should be tuned using cross validation.

Decision Trees: Hyper-parameter tuning (Pruning)

- **Maximum number of terminal nodes** Also referred as *number of leaves*. Since binary trees are created, a depth of *n* would produce a maximum of 2^n leaves.

- **Maximum features to consider for split** The number of features to consider (selected randomly) while searching for a best split. A typical value is the square root of total number of available features. A higher number typically leads to over-fitting but is dependent on the problem as well.

Cross- Validation



Why cross-validate?

bigger training set ↓ better **learning**

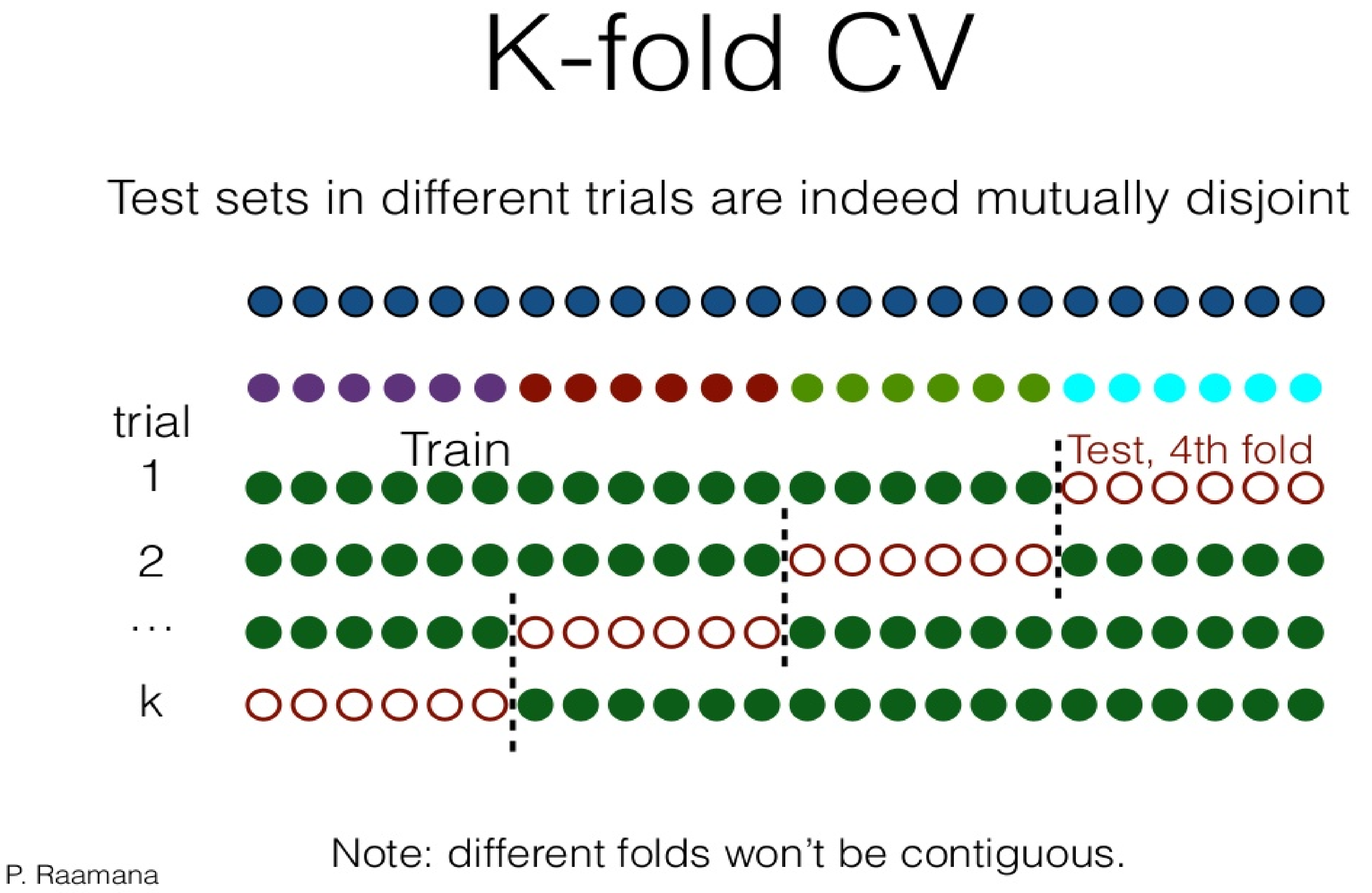
bigger test set ↓ better **testing**

Key: Train & test sets must be **disjoint.**
And the dataset or sample size is fixed.
They grow at the expense of each other!

⇨ **cross**-validate to maximize both

P. Raamana

Cross-Validation



K-fold CV

Test sets in different trials are indeed mutually disjoint

Note: different folds won't be contiguous.

P. Raamana

Decision Trees: Definitions

- Overfitting – model becomes overly complex and as a result is predicting noise or the space between features (random error) instead of the true relationship. It is in theory possible to create a leaf node for every data point.
- Ensemble methods – Process of running numerous models and codifying them using a decision rule to choose the optimal model result – example is majority vote on feature inclusion
- Heuristic algorithms – approaches designed for operational efficiency generating an approximation to the ideal result but does not guarantee the best model

Ensemble Methods

A standard error is by definition the standard deviation of the sampling distribution of a parameter estimate, generated by repeated sampling.

xerror reflects the mean of the sample means (of the errors) from the ten folds;

xstd reflects the standard deviation of the sample means (of the errors) from the ten folds. Thus, xstd is a standard deviation of sample means, which is also known as the standard error of the mean.

Example in R

UNIVERSITY of VIRGINIA
DATA SCIENCE INSTITUTE