CS5805 : Machine Learning I
Lecture # 10

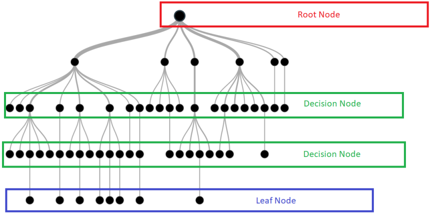
Reza Jafari, Ph.D

Collegiate Associate Professor
rjafari@vt.edu

COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH.

Definition

- A Decision Tree is a series of nodes, a directional graph that starts at the base with a single node (root) and extends to the many leaf nodes that represent the categories that the tree can classify.
- Another way to think of a decision tree is as a flow chart, where the flow starts at the root node and ends with a decision made at the leaves.
- It is a decision-support tool. It uses a tree-like graph to show the predictions that result from a series of feature-based splits.

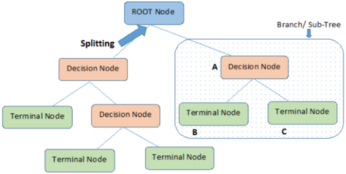


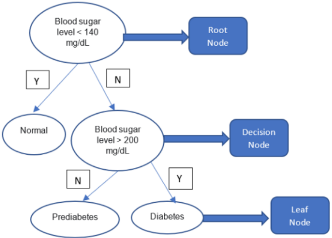
Terminologies

- A Supervised Machine Learning Algorithm, used to build classification and regression models in the form of a tree structure.

terminology

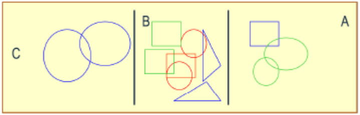
- Node- a feature (attribute)
- Branch - a decision (rule)
- Leaf- an outcome(categorical or continuous)



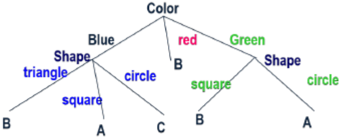


Definition

- Decision Tree is a hierarchical data structure that represents data through a divide and conquer strategy.
- Decision tree can be applied to categorical labels but non-parametric classification and regression can be performed with decision trees as well.
- Decision tress are classifiers for instances represented as feature vectors e.g. color=?, shape = ?, label = ?
- Nodes



(a) Example Data



(b) Decision Tree

Advantages of decision tree

Advantages

- Easy to Understand: no need for statistical knowledge.
- Its graphical representation is very intuitive.
- Useful in Data exploration: Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables.
- Helps to create new features that has better power to predict target variable.
- i.e. In hundreds of variables, it will help to identify most significant variable.
- Less data cleaning. It is not influenced by outliers and missing values to a fair degree.
- Can handle both numerical and categorical variables.
- Non Parametric method.

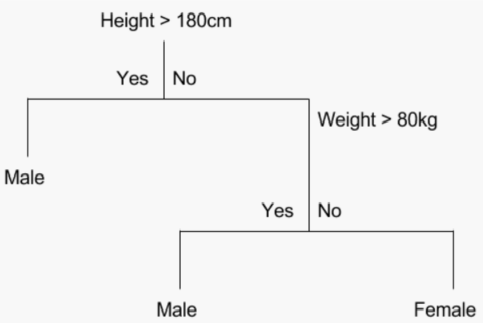
Disadvantages of decision tree

Disadvantages

- Overfitting: Over fitting is one of the most practical difficulty for decision tree models.
- Overfiting problem gets solved by setting constraints on model parameters and pruning.
- Its Not fit for continuous variables: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories.

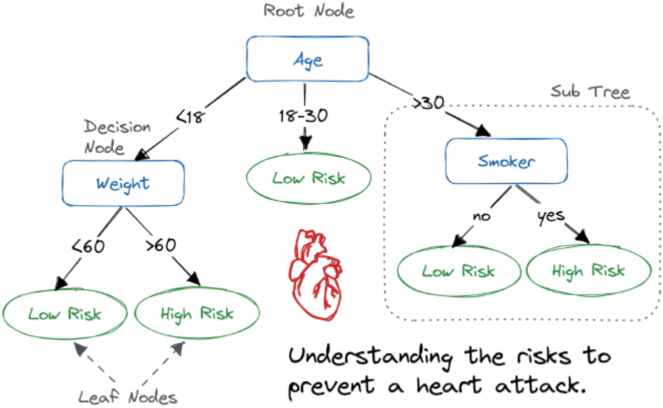
Classification and Regression Trees (CART)

- The terminal nodes (or leaves) lies at the bottom of the decision tree
- This means that decision trees are typically drawn upside down such that leaves are the bottom & roots are the tops (shown below).



Decision Tree Example

- Decision tree is a hierarchical data structure that represents data through a divide and conquer strategy.



Regression Trees vs Classification Trees

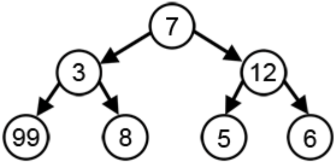
- Regression trees used when dependent variable is continuous.
- Classification trees are used when dependent variable is categorical.
- Regression tree: the value obtained by terminal nodes in the training data is the mean response of observation falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mean value.
- Classification tree: the value (class) obtained by terminal node in the training data is the mode of observations falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mode value.
- Both the trees divide the predictor space (independent variables) into distinct and non-overlapping regions. For the sake of simplicity, you can think of these regions as high dimensional boxes or boxes.

Regression Trees vs Classification Trees

- Both the trees follow a top-down greedy approach. Greedy algorithm cares about only the current split, and not about future splits which will lead to a better tree.
- This splitting process is continued until a user defined stopping criteria is reached.
- i.e,. Stop once the number of observations per node becomes less than 50.
- In both, the splitting process results in fully grown trees until the stopping criteria is reached.
- But, the fully grown tree is likely to overfit data, leading to poor accuracy on unseen data.
- Pruning is one of the technique used tackle overfitting.

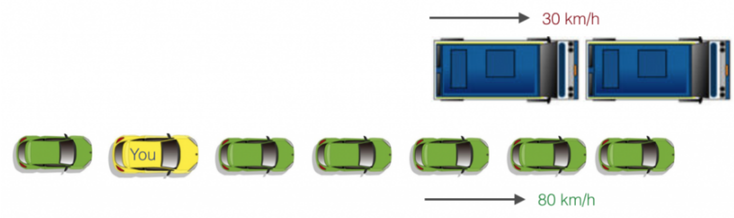
What is a Greedy algorithm?

- A greedy algorithm is an algorithm that at every step we check for optimal solution(local) in order to reach to optimal (global) solution.
- It is not guaranteed that we will always reach the optimal solution.
- In this programming paradigm we will never traverse back.
- Example: Does the local optimum solution to find max sum the same as global optimum solution in the following example?



Greedy Algorithm Example

- Decision is a greedy algorithm. it will check for the best split instantaneously and move forward until one of the specified stopping condition is reached.
- Let consider a case that there are 2 lanes in the highway:
  1. A lane with cars moving at 60mph
  2. A lane with trucks moving at 30mph
- Two choices:
  1. Take left and overtake 2 cars quickly.
  2. Keep moving in the present lane



Attribute selection

- While implementing a decision tree we must identify the attributes that will be used as root node and decision nodes.
- This process is called attribute selection. Two most popular attribute selection methods are:

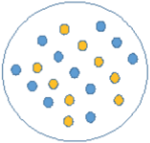
Attribute selection

- Information gain (IG)
- Gini Index

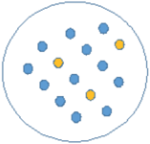
Information gain provides how much information each attribute contains. To calculate information gain first we must calculate Entropy.

Information Gain

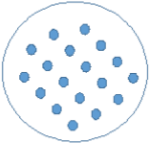
- Look at the image below. Which node can be described easily?
- ...requires less information to be described because all values are similar.
- ...requires more information to be described because it is more impure.



A



B



C

Information Gain

- Information theory is a measure to define this degree of disorganization in a system known as Entropy.
- If the sample is completely homogeneous, then the entropy is zero and if the sample is an equally divided (50% – 50%), it has entropy of one.

Entropy formula

$$E = \sum_{i=1}^{c} -pi * \log(pi)$$

c : number of classes, pi:probability associated with ith class.

- It chooses the split which has lowest entropy compared to parent node and other splits. The lesser the entropy, the better it is.

What is Entropy?

- Entropy measures the impurity in each dataset.
- In Physics and Mathematics, entropy is referred to as the randomness or uncertainty of a random variable X
- In information theory, it refers to the impurity in a group of examples.

What is information gain?

- Information gain is the decrease in entropy.
- Information gain computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

Information Gain

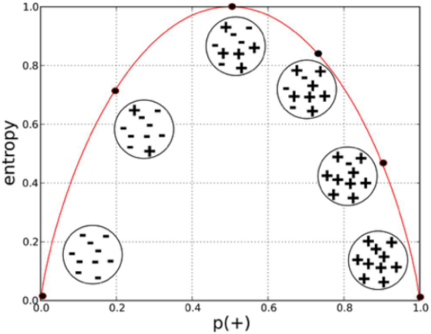
- Information Gain is entropy of full data set minus entropy of the dataset given some feature.

$$IG = E(Y) - E(Y|X)$$

- The ID3 (Iterative Dichotomiser) Decision Tree uses information gain method for attribute selection.
- By calculating decrease in entropy measure of each attribute we can calculate their information gain. The greater the reduction, the more information is gained about Y from X.

Entropy

- Entropy is the measure of disorder or uncertainty.
- The goal of machine learning and data scientists in general is to reduce uncertainty.



What are the characteristics of ID3 algorithm?

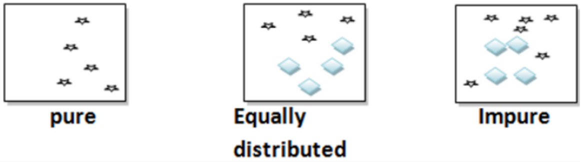
- ID3(Iterative Dichotomiser 3) uses a greedy approach that's why it does not guarantee an optimal solution; it can get stuck in local optimums.
- ID3 can overfit to the training data (to avoid overfitting, smaller decision trees should be preferred over larger ones).
- This algorithm usually produces small trees, but it does not always produce the smallest possible tree.
- ID3 is harder to use on continuous data (if the values of any given attribute is continuous, then there are many more places to split the data on this attribute, and searching for the best value to split by can be time consuming).

Gini Impurity

- Gini Impurity calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.
- If all the elements are linked with a single class then it is called pure.

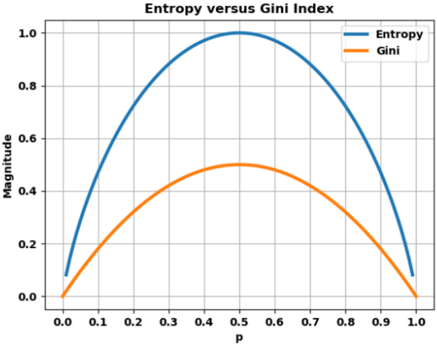
$$Gini = 1 - \sum_{i=1}^{c} (pi)^2$$

where pi is the probability associated with i^{th} class.



Gini Index v.s Entropy

- Lower the gini impurity, higher the homogeneity.
- Works only with categorical targets.
- Only performs binary splits.



What is pruning ?

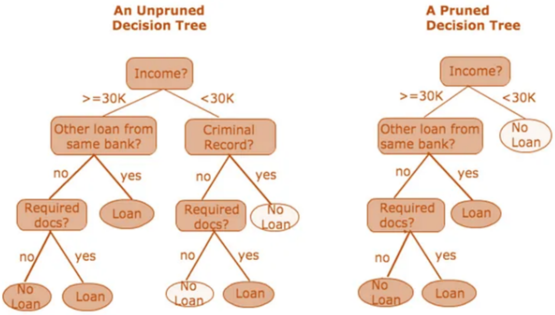
- In general pruning is a process of removal of selected part of plant such as bud, branches and roots.
- In Decision Tree pruning does the same task it removes the branches of decision tree to overcome the overfitting condition of decision tree.
- This can be done in two ways, we will discuss both the techniques in detail.

Pruning

- The this example clearly depicts the difference of an unpruned and a pruned tree. The unpruned tree looks denser and complex with high variance and hence overfitting the model.

Pre-pruning

Post-pruning



Setting constraints on tree based algorithm

Maximum depth of tree (vertical depth)

- The maximum depth of a tree.
- Should be tuned using CV.
- Used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample.

Maximum number of terminal nodes

- The maximum number of terminal nodes or leaves in a tree.
- Can be defined in place of max_depth. Since binary trees are created, a depth of n would produce a maximum of 2^n leaves

Maximum features to consider for split

Setting constraints on tree based algorithm

Minimum samples leaf

- min samples leaf hyperparameter is set to 10. This means that a leaf node, or a terminal node, in the decision tree must have at least 10 samples in order for it to be created. If a certain split results in a leaf node with only 5 samples, the split will not be allowed

Maximum samples split

- hyperparameter is set to 20. This means that an internal node must have at least 20 samples to be split. If certain internal node has only 15 samples, the node will not be allowed to split further, meaning the decision for classifying a new sample will be based on the majority of the class in that leaf node.

Post pruning

- As the name implies pruning involves cutting back the tree.
- After a tree has been built (and in the absence of early stopping discussed below) it may be overfitted.
- Overfitting means: The tree learned the data exactly, but a new data point that differs very slightly might not be predicted well.
- Two methods of pruning:

1- Post pruning

- This technique is used after construction of decision tree.
- This technique is used when decision tree will have very large depth and will show overfitting of model.
- This technique is used when we have infinitely grown decision tree.
- Here we will control the branches of decision tree that is maximum depth and minimum samples split using cost complexity pruning.

Cost complexity Pruning

Cost complexity Pruning

- In Post pruning, we grow a large tree and then prune it back in order to obtain a subtree such that we get the lowest test error rate.
- The problem with this algorithm is that we don't want to go to every subtree and choose each one of them to calculate the change in the test error rate
- Cost complexity Pruning/Weakest link Pruning helps us with that. It introduces a new term, α.
- For each value of α we have a subtree which can minimize the value of:

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

Cost complexity Pruning

- where R_m is the rectangle corresponding to m^{th} terminal node.
- $\hat{y}_{R_m}$ is the mean of the training observations in R_m

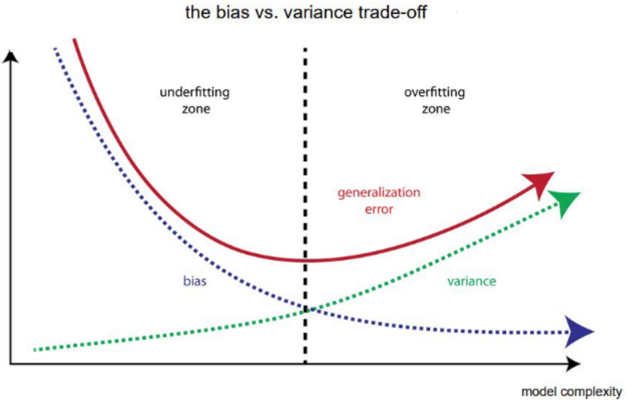
Steps involved in building Regression Tree using Tree Pruning

- Grow the large stopping tree, stopping only when the terminal node contains fewer than some minimum number of observations. For example, we will keep dividing until each region has less than 20 data points.
- Apply cost complexity pruning to the large tree and get the sequence of best subtrees as a function of α. The idea is to minimize the cost-complexity function.

$$C_\alpha = R(T) + \alpha|T|$$

T: is the number of leaves of the tree and $R(T)$ is the loss function calculated across the leaves.

Bias and variance Trade-off



Pre-pruning

2- Pre pruning

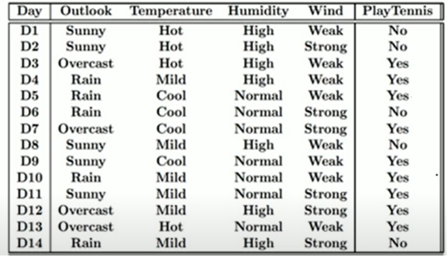
- This technique is used before construction of decision tree.
- Pre-Pruning can be done using Hyperparameter tuning.
- Overcome the overfitting issue.
- In this lecture we will use GridSearchCV for Hyperparameter tuning.
- What is Hyperparameter Tuning ?
  grid_param=”criterion”:[”gini”,”entropy”],
  ”splitter”:[”best”,”random”], ”max_depth”:range(2,50,1),
  ”min_samples_leaf”:range(1,15,1),
  ”min_samples_split”:range(2,20,1)

Example

- Consider an example where we are building a decision tree to predict whether a loan given to a person would result in a write off or not.
- Population : 30
- 16 write off, 14 non-write off
- features: Balance and Residence
- Balance $= \begin{cases} < 50K \\ > 50K \end{cases}$
- Residence $= \begin{cases} own \\ rent \\ other \end{cases}$
- we want to find out which feature provides more information gain or reduce uncertainty about our target using Entropy and Information Gain.

Example

- Consider a dataset shown below. Develop a decision tree model then classify the following test observation.
- We want to know if Outlook = sunny and Temp = cool and humidity = high and wind is strong, then play tennis YES or NO?



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No