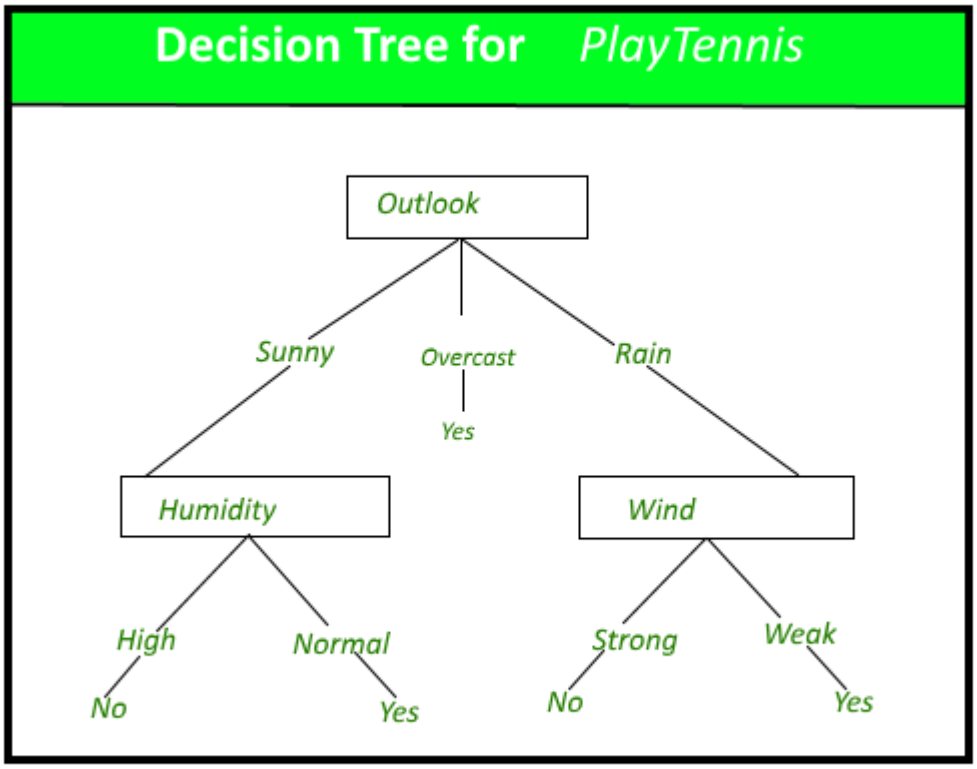
CSE/ECE-471: Statistical Methods in AI **Spring 2020**

Decision Tree

Prepared by: Nabhiraj Jain(2019201062),Damodhar Reddy Munagala (2019900072)

1 what is a Decision Tree

A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label



with the help of decision tree we can predict the output by just walking down the nodes of the decision tree according to the input features.

2 Building a Decision Tree

we can build the decision tree manually from our experience or we can learn the the decision tree from the data.

2.1 Learning the Decision Tree

we want to find as short decision tree as possible which will correctly predict the output label given a input feature. Unfortunately the space of decision tree is too big for systematic search and it is computationally infeasible to do so. Therefore we will search for the best node in a greedy manner one at a time,this can be done as follow

1. Select the best attribute using Attribute Selection Measures(ASM) to split the records.
2. Make that attribute a decision node and break the data set into smaller subsets.
3. Starts tree building by repeating this process recursively for each child until one of the condition will match:
   (a) All the tuples belong to the same attribute value.
   (b) There are no more remaining attributes.
   (c) There are no more instances.

3 Which decision attribute is Best?

A a decision attribute is better than other decision attributes if it can separate the different output labels into different child nodes more than other decision attributes.
Therefore we need a quantitative measure which can tell how well a given attribute separate the training example according to their target classification.This measure will be used to select the decision attribute among the candidate attributes at each step while growing the tree.

3.1 Entropy and Information Gain

Entropy is the measure of disturbance in the system
let p be the proportion of positive exaple in set S
let n be the proportion of negative example in set S
then entropy of the set S will be

$$entropy(s) = -plog(p) - nlog(n) \tag{1}$$

on the other hand information Gain is the measure of how much we can reduce uncertainty

$$Gain(S, A) = entropy(S) - \sum_{v}(num(Sv)/num(S))entropy(S) \tag{2}$$

Therefore we want to choose the attribute which gives the heigst information gain

3.2 Other similar measures

like Information gain there are many other measure some of them are

1. Gini index,
2. Gain Ratio,
3. Reduction in Variance
4. Chi-Square

3.2.1 Gini Index:

You can understand the Gini index as a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions and is easy to implement whereas information gain favors smaller partitions with distinct values. Gini Index works with the categorical target variable "Success" or "Failure". It performs only Binary splits.

$$Gini = 1 - \sum_{i=1}^{c}(p_i^2) \tag{3}$$

Higher the value of Gini index higher the homogeneity. Steps to Calculate Gini index for a split

1. Calculate Gini for sub-nodes, using the above formula for success(p) and failure(q)
2. Calculate the Gini index for split using the weighted Gini score of each node of that split.

CART (Classification and Regression Tree) uses the Gini index method to create split points

3.2.2 Gain Ratio:

Gain ratio overcomes the problem with information gain by taking into account the number of branches that would result before making the split. It corrects information gain by taking the intrinsic information of a split into account.

Let us consider if we have a dataset that has users and their movie genre preferences based on variables like gender, group of age, rating, etc. With the help of information gain, you split at 'Gender' (assuming it has the highest information gain) and now the variables 'Group of Age' and 'Rating' could be equally important and with the help of gain ratio, it will penalize a variable with more distinct values which will help us decide the split at the next level.

$$GainRation = InformationGain/SplitInfo = (Entropy(before) - \sum_{j=1}^{k}(Entropy(j, after))/\sum_{j=1}^{k}(w_j log_2 w_j) \tag{4}$$

where "before" is the dataset before the split, k is the number of subsets generated by the split, and (j, after) is subject j after the split.

3.2.3 Reduction in Variance:

Reduction in variance is an algorithm used for continuous target variables (regression problems). This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population:

$$Variance = (\sum(X - \overline{X})^2)/n \tag{5}$$

Steps to calculate Variance:

1. Calculate variance for each node.
2. Calculate variance for each split as the weighted average of each node variance.

3.2.4 Chi-Square:

The acronym CHAID stands for Chi-squared Automatic Interaction Detector. It is one of the oldest tree classification methods. It finds out the statistical significance between the differences between sub-nodes and parent node. We measure it by the sum of squares of standardized differences between observed and expected frequencies of the target variable.

It works with the categorical target variable "Success" or "Failure". It can perform two or more splits. Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node.

It generates a tree called CHAID (Chi-square Automatic Interaction Detector).

Mathematically, Chi-squared is represented as:

$$\chi^2 = \sum((O-E)^2/E) \tag{6}$$

where:
O = Observed score
E = expected score

Steps to Calculate Chi-square for a split:

1. Calculate Chi-square for an individual node by calculating the deviation for Success and Failure both
2. Calculated Chi-square of Split using Sum of all Chi-square of success and Failure of each node of the split

4 Input attribute with continuous values

till now we have discussed for nominal value. what if the value is continuous?in such a case we will select a value and partition the input feature into two sets one which are greater then the selected value and one which are less than that.we can do a binary split or a multi way split.
i.e

$$A_c = [\begin{cases} True\ if\ A_c < C \\ False\ otherwise \end{cases}\ ] \tag{7}$$

we can chose c in following manner

1. considers all the possible split and find the best one.
2. or partition the continuous value of attribute into desecrate set of intervals and find the best one.

5 Disadvantage of Decision Tree

1. Decision trees are prone to over fitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions.
2. Gives most optimal solution but not globally optimal solution
3. Decision trees do not have same predictive accuracy compared to other regression and classification models

6 Random forest

An ensemble learning model aggregates multiple machine learning models to give a better performance. In random forest we use multiple random decision trees for a better accuracy. Random Forest is a ensemble bagging algorithm to achieve low prediction error. It reduces the variance of the individual decision trees by randomly selecting trees and then either average them or picking the class that gets the most vote. Bagging is a method for generating multiple versions of a predictor to get an aggregated predictor

6.1 Advantages of Random forest

1. High predictive accuracy.
2. Efficient on large datasets
3. Ability to handle multiple input features without need for feature deletion
4. Prediction is based on input features considered important for classification.
5. Works well with missing data still giving a better predictive accuracy

6.2 Disadvantages

1. Not easily interpretable
2. Random forest overfit with noisy classification or regression

7 References

1. NPTEL introduction to decision tree https://www.youtube.com/watch?v=FuJVLsZYkuE
2. NPTEL learning in a decision tree https://www.youtube.com/watch?v=7SSAA1CE8Ng
3. random forest concept-1 https://www.youtube.com/watch?v=sQ870aTKqiM
4. random forest concept-2 https://www.youtube.com/watch?v=J4Wdy0Wc_xQ