Data Science Concepts

Lesson04–Decision Tree Concepts

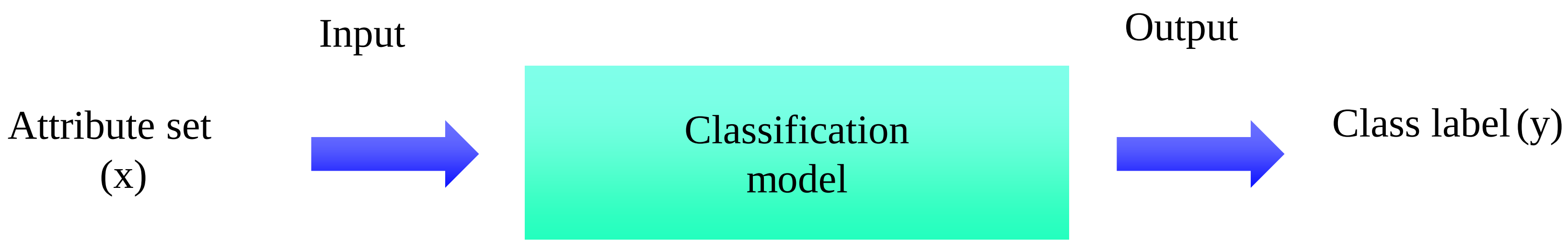
Objective

After completing this lesson you will be able to:

- Explain Decision Trees and its applications
- Explain the various parameters which are used to evaluate the outcome of the decision trees.

Decision Trees

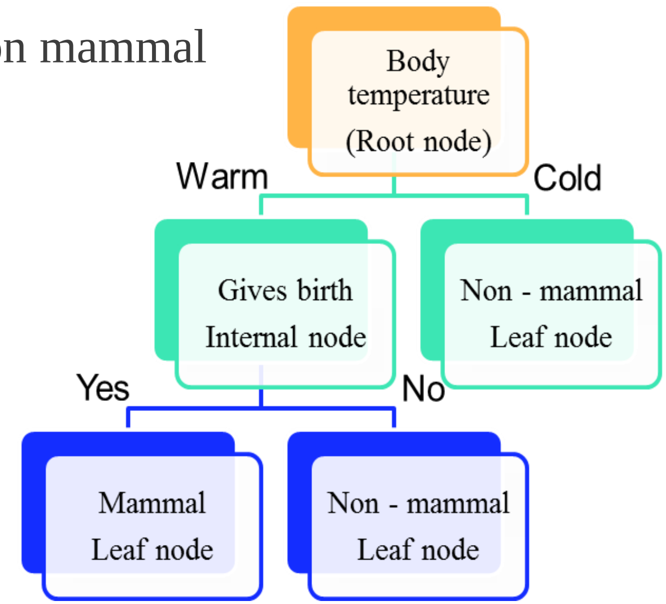
- Classification is a task of assigning objects to one of the several pre-defined categories.
  - Descriptive modelling: Can be used as an explanatory tool to distinguish between objects of different classes.
  - Predictive modelling: Can be used to predict the class label of unknown records.



- Objective is to build a learning algorithm with good generalization capability.

Decision Tree–Concept Development

Classifying species as mammal or non mammal



CART	C5.0	CHAID
Hunt’s algorithm	Hunt’s algorithm	CHAID algorithm
Split: Gini Index	Split: Entropy	Split: x^2 test

Criteria for comparing different methods: Predictive accuracy, speed, robustness, scalability, Interpretability

Decision Tree - CART

- CART (Classification and Regression Tree) always performs binary splits.
  - Gini Index is a measure of impurity at the node. If sample is completely homogenous then less impurity. If sample is equally divided then more impurity.

$$\boldsymbol{i(t) = Gini(t) = \sum_{j=1}^{J} P(j \mid t) * \left(1 - P(j \mid t)\right)}$$

$$where P(j|t)\ is\ the\ proportion\ of\ cateogy\ j\ at\ node\ t.$$

$$Change\ in\ impurity = \ [\boldsymbol{i(t) - P_L * i\big(t(L)\big) - P_R * i\big(t(R)\big)}]$$

$$P_L = Proportion\ of\ obs\ in\ left\ branch$$

$$P_R = Proportion\ of\ obs\ in\ left\ branch$$

  - The variable which maximizes the change in impurity is picked up for building decision tree
  - In case of a two category, minimum value of Gini is 0 and Maximum value of Gini can be 0.5 (50% zeros and 50% ones as the two categories).
  - If a variable has more than two classes, the classes are combined and then Gini index is computed:

$$No\ of\ combinations = \ 2^{k-1} - 1$$

Decision Tree - CART

- Entropy is another measure to select the best split

$$\boldsymbol{Entropy(t) = -\sum_{j=1}^{J} P(j \mid t) * log_2\big(P(j \mid t)\big)}$$

$$where P(j|t)\ is\ the\ proportion\ of\ cateogy\ j\ at\ node\ t.$$

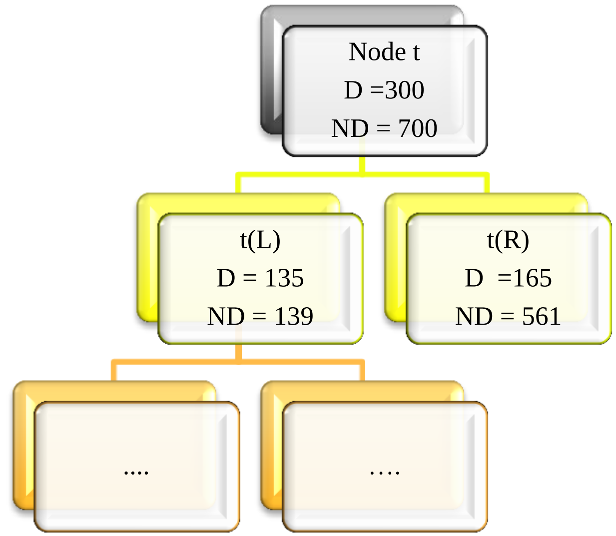
  - The variable which maximizes the change in impurity is picked up for building decision tree
  - In case of a two category, minimum value of entropy is 0 and Maximum value is 1 (50% zeros and 50% ones as the two categories).
  - If a variable has more than two classes, the classes are combined and then Gini index is computed:

$$No\ of\ combinations = 2^{k-1} - 1$$

Decision Tree - CART

Contingency Table

Gender (X)	**NPA Status**		**Total**
	1	**0**	
Node (t(L): Male	135	139	274
Node (t(R): Female	165	561	726
Total	300	700	1000



Calculation Table

Node	Proportion of the Class	i(t) = P(j | K)*(1-P(j | K))		Proportion	$\Delta i(t)$
t():	P(D | t) = 300/1000 P(ND | t) = 700/1000	0.30*(1-0.30) = 0.21 0.70*(1.0.70) = 0.21	**0.42**		
t(L): Male	P(D | t(L)): 135/274 = 0.49 P(ND | t(L)): 139/274 = 0.51	(0.49)*(1-0.49) = 0.25 (0.51)*(1-0.51) = 0.25	**0.50**	274/1000= 0.274	[0.42 − (0.27∗0.50) − (0.726∗0.34)]= 0.038
t(R): Female	P(D | t(R)): 165/726 = 0.23 P(ND | t(R)): 561/726 = 0.77	(0.23)*(1-0.23) = 0.17 (0.77)*(1-0.77) = 0.17	**0.34**	726/1000= 0.726	

Decision Tree–Classification Matrix

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) = \frac{4}{7} = 57.1\%$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) = \frac{17}{17} = 100\%$$

Classification matrix		
	Predicted	
Observed	**Class=1 (Positive)**	**Class=0 (Negative)**
Class =1 (Positive)	f_{11}= 4 [TP]	f_{10}= 3 [FN]
Class =0 (Negative)	f_{01}= 0 [FP]	f_{00}= 17 [TN]

$$Model\ accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) = \frac{21}{24} = 87.5\%$$

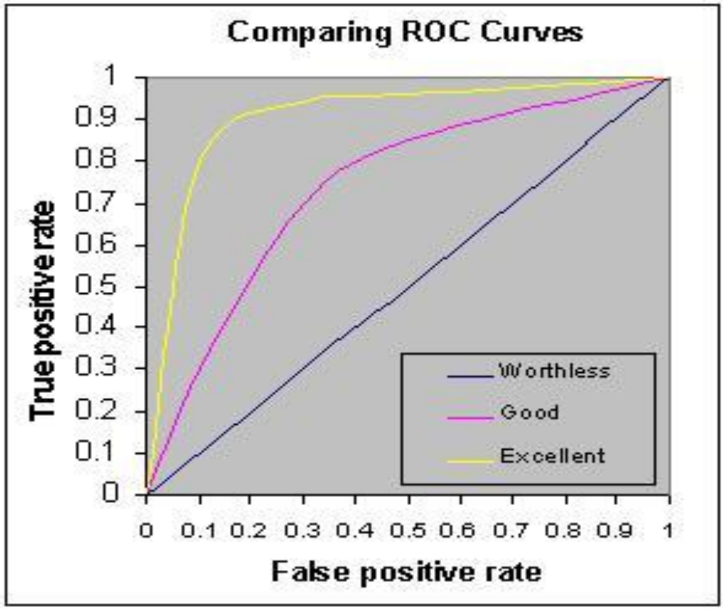
Sensitivity is the probability that predicted class is 1 when observed class is 1.
Specificity is the probability that the predicted class is 0 when the observed class is 0.

Decision Tree–ROC Curve

- Receiver operating characteristics (ROC) Curve is a useful way to determine cut-off point which maximizes sensitivity and specificity.
- Sensitivity and specificity measures are computed based on a sequence of cut-off points to be applied to the model for predicting observations into Positive or Negative.

An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). AUC values between:

- 0.9-1 indicate perfect sensitivity and specificity,
- 0.8-0.9 indicate good sensitivity and specificity,
- 0.7-0.8 indicate fair sensitivity and specificity,
- 0.6-0.7 is poor
- 0.6 and below indicate by chance outcome



Decision Tree–Pruning

Pruning is applied to overcome the under fitting or over fitting issues in the decision tree model

Pre-pruning

Stop the algorithm before it becomes a fully grown tree:

- Stop if number of instances is less than some user specified threshold.
- Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain) by at least some threshold

This is more efficient but less accurate.

Post Pruning

Grow decision tree to its entirety. Trim the nodes of the decision tree in a bottom-up fashion

- If generalization error improves after trimming, replace sub-tree by a leaf node.
- Class label of leaf node is determined from majority class of instances in the sub-tree

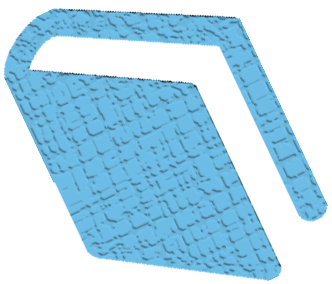
This is more accurate but less efficient.

Misclassification error pruning: Decision tree pruning stops when number of cases in a terminal node becomes less than a threshold

Decision Tree in R Using an Example

Summary

Summary of the topics covered in this lesson:



- Decision Tree is one of the most widely used data mining technique.
- The outcome of decision tree can be used for exploration of data as well as to build in predictive model.
- Unlike regression and logistic regression model, there are no statistical attributes which can suggest that the decision tree model is good and generalizable.

End of Lesson04–Decision Tree Concepts