Supervised Machine Learning…

Decision Trees

Supervised Machine Learning…

Decision Trees -

1. Classifiers utilize a tree structure to model relationships among the features and the potential outcomes

2. Decision trees consist of nodes and branches. Nodes represent a decision function while branch represents the result of the function. Thus it is a flow chart for deciding how to classify a new observation:

3. The nodes are of three types, Root Node (representing the original data), Branch Node (representing a function), Leaf Node (which holds the result of all the previous functions that connect to it)

Supervised Machine Learning…

Decision Trees -

4. For classification problem, the posterior probability of all the classes is reflected in the leaf node and the Leaf Node belongs to the majority class.

5. After executing all the functions from Root Node to Leaf Node, the class of a data point is decided by the leaf node to which it reaches

6. For regression, the average/ median value of the target attribute is assigned to the query variable

7. Tree creation splits data into subsets and subsets into further smaller subsets. The algorithm stops splitting data when data within the subsets are sufficiently homogenous or some other stopping criterion is met

Supervised Machine Learning…

<u>Decision Trees </u>-

1. The decision tree algorithm learns (i.e. creates the decision tree from the data set) through optimization of a loss function

2. The loss function represents the loss of impurity in the target column. The requirement here is to minimize the impurity as much as possible at the leaf nodes

3. Purity of a node is a measure of homogeneity in the target column at that node

Supervised Machine Learning…

Decision Trees -

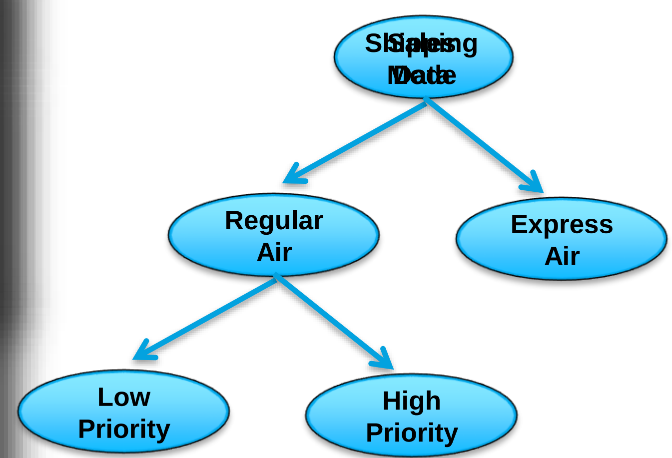
1. There is a bag of 50 balls of red, green, blue, white and yellow colour respectively
2. You have to pull out one ball from the bag with closed eyes. If the ball is -
   a. Red, you loose the prize money accumulated
   b. Green, you can quit
   c. Blue you loose half prize money but continue
   d. White you loose quarter prize money & continue
   e. Yellow you can skip the question
3. This state where you have to decide and your decision can result in various outcomes with equal probability is said to be state of maximum uncertainty

Supervised Machine Learning…

Decision Trees -

Suppose we wish to find if there was any influence of shipping mode, order priority on customer location. Customer location is target column and like the bag of coloured balls

Row ID	Order ID	Order Date	Order Priorit	Order Quantit	Sales	Discoun	Ship Mod	Profit	Unit Pric	Shipping Cos	Customer Nam	Province
6555	46599	13-09-2010	Not Specified	4	284	0.01	Express Air	208.31	62.18	10.84	Victoria Brennan	Alberta
7110	50726	31-10-2011	Low	21	98.88	0.07	Express Air	36.01	4.13	0.99	Ruben Dartt	Alberta
7269	51872	23-04-2012	Medium	11	812.498	0	Express Air	128.95	85.99	1.25	Shui Tom	Alberta
7658	54913	03-07-2010	High	11	54.61	0.08	Express Air	14.99	4.76	0.88	Tonja Turnell	Alberta
7738	55424	02-03-2010	Medium	6	25.84	0.04	Express Air	-4.13	1.76	4.86	Victoria Brennan	Alberta
7880	56327	11-01-2010	Medium	47	1276.73	0.08	Express Air	357.23	29.18	8.55	Victoria Brennan	Alberta
805	5767	28-04-2012	High	36	163.54	0.03	Express Air	-95.06	4.13	5.04	Jessica Myrick	Alberta
6492	46212	12-09-2012	Not Specified	43	322.47	0.09	Express Air	72.28	7.78	2.5	Grant Donatelli	Alberta
7396	52706	09-07-2012	Low	34	1041.66	0.02	Express Air	480.53	28.53	1.49	Harry Greene	Alberta
7906	56550	08-04-2011	Not Specified	37	823.78	0.03	Express Air	343.05	22.23	5.08	Frank Hawley	Alberta
7914	56581	08-02-2009	High	20	2026.01	0.1	Express Air	580.43	105.98	13.99	Grant Donatelli	Alberta
1	3	13-10-2010	Low	6	261.54	0.04	Regular Air	-213.25	38.94	35	Muhammed MacIntyr	Nunavut
50	293	01-10-2012	High	27	244.57	0.01	Regular Air	46.71	8.69	2.99	Barry French	Nunavut
80	483	10-07-2011	High	30	4965.7595	0.08	Regular Air	1198.97	195.99	3.99	Clay Rozendal	Nunavut
85	515	28-08-2010	Not Specified	19	394.27	0.08	Regular Air	30.94	21.78	5.94	Carlos Soltero	Nunavut
86	515	28-08-2010	Not Specified	21	146.69	0.05	Regular Air	4.43	6.64	4.95	Carlos Soltero	Nunavut
97	613	17-06-2011	High	12	93.54	0.03	Regular Air	-54.04	7.3	7.72	Carl Jackson	Nunavut
98	613	17-06-2011	High	22	905.08	0.09	Regular Air	127.70	42.76	6.22	Carl Jackson	Nunavut
107	678	26-02-2010	Low	44	228.41	0.07	Regular Air	-226.36	4.98	8.33	Dorothy Badders	Nunavut
127	807	23-11-2010	Medium	45	196.85	0.01	Regular Air	-166.85	4.28	6.18	Neola Schneider	Nunavut
128	807	23-11-2010	Medium	32	124.56	0.04	Regular Air	-14.33	3.95	2	Neola Schneider	Nunavut
134	868	08-06-2012	Not Specified	32	716.84	0	Regular Air	134.72	21.78	5.94	Carlos Daly	Nunavut
135	868	08-06-2012	Not Specified	31	1474.33	0.04	Regular Air	114.46	47.98	3.61	Carlos Daly	Nunavut
149	933	04-08-2012	Not Specified	15	80.61	0.02	Regular Air	-4.72	5.28	2.99	Claudia Miner	Nunavut
160	995	30-05-2011	Medium	46	1815.49	0.03	Regular Air	782.91	39.89	3.04	Neola Schneider	Nunavut
161	998	25-11-2009	Not Specified	16	248.26	0.07	Regular Air	93.80	15.74	1.39	Allen Rosenblatt	Nunavut
176	1154	14-02-2012	Critical	11	663.784	0.25	Regular Air	-481.04	71.37	69	Sylvia Foulston	Nunavut



When sub branches are created, the total entropy of the sub branches should be less than the entropy of the parent node. More the drop in entropy, more the information gained

Supervised Machine Learning…

<u>Decision Trees</u> – Shannon's Entropy

a. Imagine a bag contains 6 red and 4 black balls.

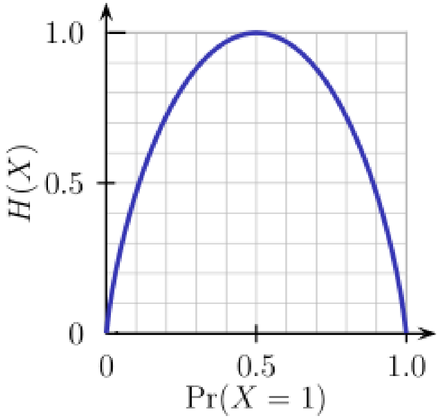
b. Let the two classes Red -> class 0 and Black -> class 1

c. Entropy of the bag (X) will be calculated as per the formula $H(X) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$

   a. H(X) = - (0.6 * log2(0.6)) - (0.4 * log2(0.4)) = 0.9709506

d. Suppose we remove all red balls from the bag and then entropy will be

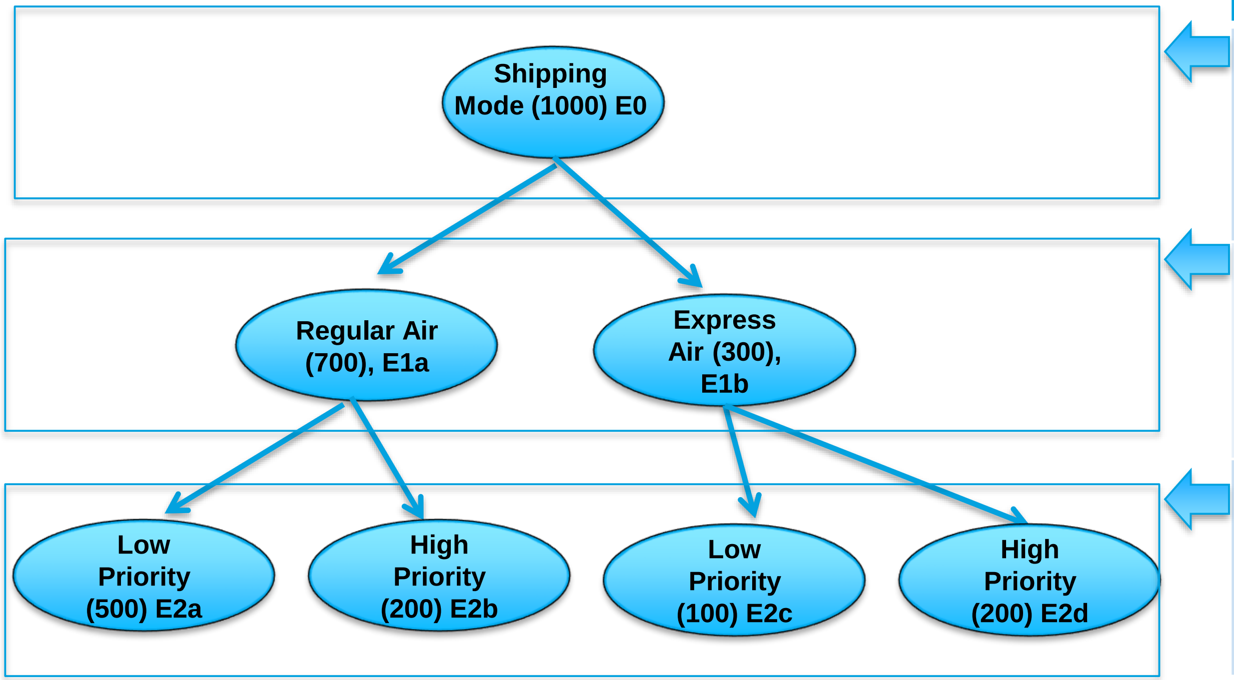
   a. H(X) = - 1.0 *log2(1.0) – 0.0 * log2(0) = 0 ## Entropy is 0! i.e. Information is 100%



Supervised Machine Learning…

Machine Learning (Decision Tree Classification)

Decision Trees -



Entropy	Info Gain
E0 = max entropy say 1	0
E1 = (E1a*700/1000) + (E1b * 300/1000)	E0 – E1
E2 = (E2a * 500/700) + (E2b * 200/700) + (E2c * 100/300) + (E2d * 200/300)	E1 – E2

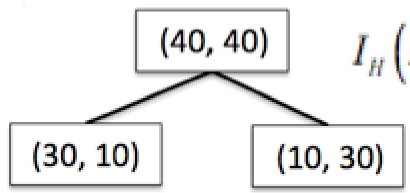
Tree will stop growing when stop criterion for the splitting is reached which could be -

a. Tree has reached certain pre-fixed depth (longestt path from root node to leaf node)
b. Tree has achieve maximum number of nodes (tree size)
c. Exhausted all attributes to split
d. Leaf node on split will have less than predefined number of data points

Supervised Machine Learning…

Decision Trees - Information Gain using Entropy

$$H(X) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$



$$I_H(D_p) = -\left(0.5\ \log_2(0.5) + 0.5\ \log_2(0.5)\right) = 1$$

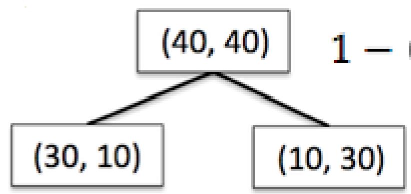
$$-\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.81 \qquad -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) = 0.81$$

Information Gain = reduction in entropy = $1 - \frac{4}{8}0.81 - \frac{4}{8}0.81 = 0.19$

Supervised Machine Learning…

Decision Trees - Information Gain using Gini index

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$



$$1 - (0.5^2 + 0.5^2) = 0.5$$

$$1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = \frac{3}{8} = 0.375$$

$$1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = \frac{3}{8} = 0.375$$

Information Gain = reduction in Gini index = $0.5 - \frac{4}{8}0.375 - \frac{4}{8}0.375 = 0.125$

Supervised Machine Learning…

<u>Decision Trees</u> -

Common measures of purity

1. Gini index – is calculated by subtracting the sum of the squared probabilities of each class from one
   a. Uses squared proportion of classes
   b. Perfectly classified, Gini Index would be zero
   c. Evenly distributed would be 1 – (1/# Classes)
   d. You want a variable split that has a low Gini Index
   e. Used in CART algorithm

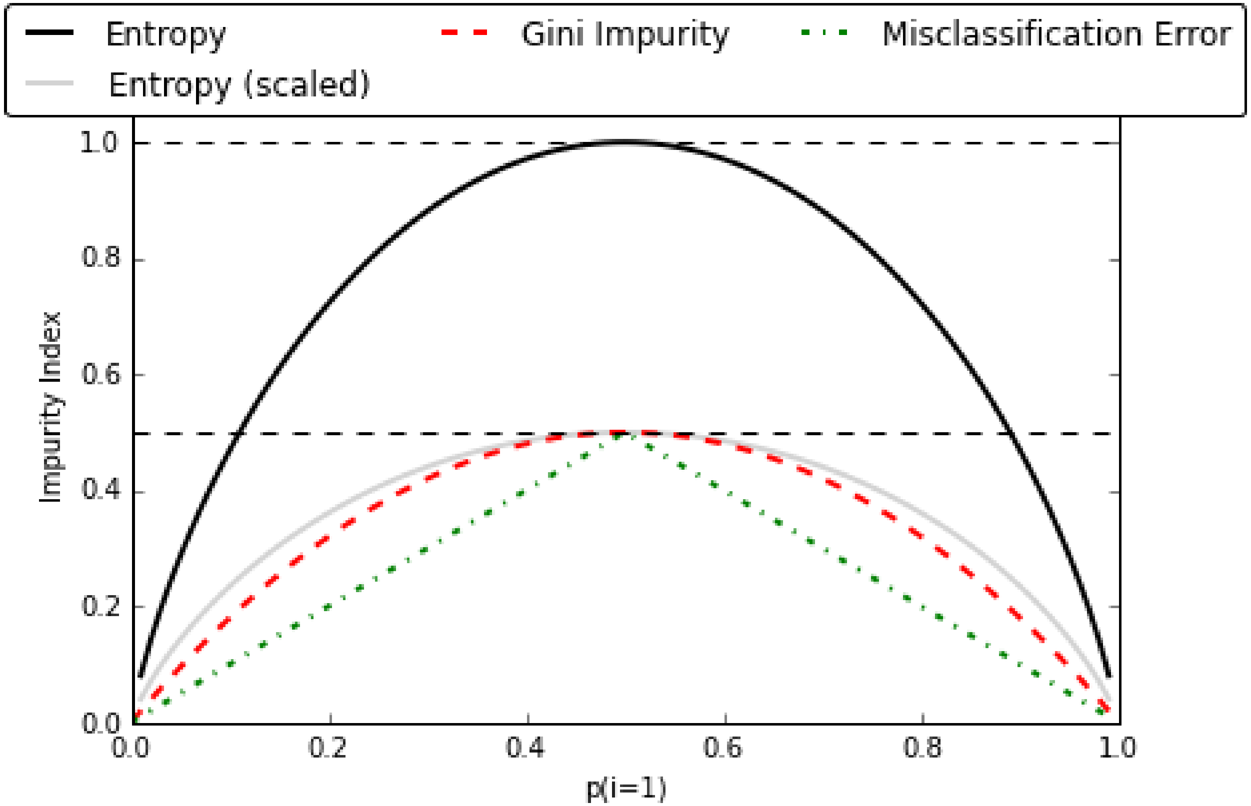
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

2. Entropy –
   a. Favors splits with small counts but many unique value
   b. Weights probability of class by log(base=2) of the class probability
   c. A smaller value of Entropy is better. That makes the difference between the parent node's entropy larger
   d. Information Gain is the Entropy of the parent node minus the entropy of the child nodes

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Supervised Machine Learning…

Decision Trees – Gini , Entropy , Misclassification Error



Note: Misclassification Error is not used in Decision Trees

Supervised Machine Learning…

Decision Trees - Algorithms

1. ID3 (Iterative Dicotomizer 3) – developed by Ross Quinlan. Creates a multi branch tree at each node using greedy algorithm. Trees grow to maximum size before pruning

2. C4.5 succeeded ID3 by overcoming limitation of features required to be categorical. It dynamically defines discrete attribute for numerical attributes. It converts the trained trees into a set of if-then rules. Accuracy of each rule is evaluated to determine the order in which they should be applied

3. C5.0 is Quinlan's latest version and it uses less memory and builds smaller rulesets than C4.5 while being more accurate

4. CART (Classification & Regression Trees) is similar to C4.5 but it supports numerical target variables and does not compute rule sets. Creates binary tree. Scikit uses CART

Supervised Machine Learning…

Decision Trees -

Advantages -

1. Simple , Fast in processing and effective
2. Does well with noisy data and missing data
3. Handles numeric and categorical variables
4. Interpretation of results does not required mathematical or statistical knowledge

Dis-advantages -

1. Often biased towards splits or features have large number of levels
2. May not be optimum as modelling some relations on axis parallel basis is not optimal
3. Small changes in training data can result in large changes to the logic
4. Large trees can be difficult to interpret

Supervised Machine Learning…

Decision Trees - Preventing overfitting through regularization

1. Decision trees do not assume a particular form of relationship between the independent and dependent variables unlike linear models for e.g.

2. DT is a non-parametrized algorithm unlike linear models where we supply the input parameters

3. If left unconstrained, they can build tree structures to adapt to the training data leading to overfitting

4. To avoid overfitting, we need to restrict the DT's freedom during the tree creation. This is called regularization

5. The regularization hyperparameters depend on the algorithms used

Supervised Machine Learning…

Decision Trees - **Regularization parameters**

1. max_depth – Is the maximum length of a path from root to leaf (in terms of number of decision points. The leaf node is not split further. It could lead to a tree with leaf node containing many observations on one side of the tree, whereas on the other side, nodes containing much less observations get further split

2. min_sample_split - A limit to stop further splitting of nodes when the number of observations in the node is lower than this value

3. min_sample_leaf – Minimum number of samples a leaf node must have. When a leaf contains too few observations, further splitting will result in overfitting (modeling of noise in the data).

Supervised Machine Learning…

<u>Decision Trees - </u>Regularization parameters (Contd…)

4. min_weight_fraction_leaf – Same as min_sample_leaf but expressed in fraction of total number of weighted instances

5. max_leaf_nodes – maximum number of leaf nodes in a tree

6. max_feature_size - max number of features that are evaluated for splitting each node

Supervised Machine Learning…

Decision Tree -

Lab- 5 Model to predict potential credit defaulters

Description – Sample data is available at local file system as credit.csv

The dataset has 16 attributes described at
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
 or in the notes page of this slide

Sol: Regularization+Credit+Decision+Tree.ipynb