Introduction to Decision Trees

Friday, August 23, 2019 9:09 PM

Decision Trees -

1. Decision Tree is one of the most commonly used models in data science world
2. It also is a proven management tool used to take decisions in complex situations
3. It can be used for regression and classification, more often used for classification
4. Can be used for binary classification such as whether an applicant for loan is likely to turn into defaulter or not, whether a customer is likely to churn or not
5. It can also be used for multi-class classification for e.g. identifying the character in English alphabet
6. Decision Tree algorithm finds the relation between the target column and the independent variables and expresses it as a tree structure
7. It does so by binary splitting data using functions based on comparison operators on the independent columns

Decision Trees Structure – Training / Building

1. Suppose we are given the data about cars as shown
2. Our objective is to find if any patterns exist that connect the "Horse-Power" and "Weight" to car type (Large or Small)
3. The independent variables are "Horse-Power" and "Weight" while the target column is "Car Type"
4. The target column has binary values (L and S) in equal numbers.

Horse-Power	Weight	Car Type
130	3500	L
90	2000	S
90	1500	S
150	3000	L
270	2500	L
200	2900	L
70	2530	S
215	2000	L
80	2200	S
100	1700	S

Decision Trees Structure – Training / Building

Let us apply a function on the weight column.

Horse-Power	Weight	Car Type
130	3500	L
90	2000	S
90	1500	S
150	3000	L
270	2500	L
200	2900	L
70	2530	S
215	2000	L
80	2200	S
100	1700	S

Weight > 2000.

Horse-Power	Weight	Car Type
80	2200	S
270	2500	L
70	2530	S
200	2900	L
150	3000	L
130	3500	L

Horse-Power	Weight	Car Type
90	1500	S
100	1700	S
90	2000	S
215	2000	L

1. This smaller node on top has "L" in majority in the target column hence gets label "L"
2. The smaller node on the bottom has "S" in majority in the target column hence gets label "S"
3. The homogeneity of the target column in both the smaller nodes has increased compared to parent
4. But both the smaller nodes still have a mix of values in the target column
5. Let us split the data further using Horse_Power

Decision Trees Structure – Training / Building

Class = L (Majority)

Horse-Power	Weight	Car Type
80	2200	S
270	2500	L
70	2530	S
200	2900	L
150	3000	L
130	3500	L

HorsePower > 100.

Horse-Power	Weight	Car Type
70	2530	S
80	2200	S
80	2200	S

Horse-Power	Weight	Car Type
150	3000	L
200	2900	L
270	2500	L

Let us apply a function on the Horsepower column

Horse-Power	Weight	Car Type
90	1500	S
100	1700	S
90	2000	S
215	2000	L

Class = S (Majority)

HorsePower > 200.

Horse-Power	Weight	Car Type
90	1500	S
90	2000	S
100	1700	S

Horse-Power	Weight	Car Type
215	2000	L

1. The smaller node on the top now is perfectly homogenous in target column and belongs to class "S"
2. The second node similarly belongs to "L"
3. The third node belongs to "S"
4. The fourth node belongs to "L"
5. There is no further need to split the data as it is perfectly homogenous!

Decision Trees Structure – Training / Building

The tree thus has given us the following relation between "Weight", "Horse-Power" and "Car Type" as

```
If wt > 2000
   if hp > 100
      class = "L"
   else
      class  = "S"

If wt <= 2000
   if hp > 200
      class = "L"
   else
      class = "S
```

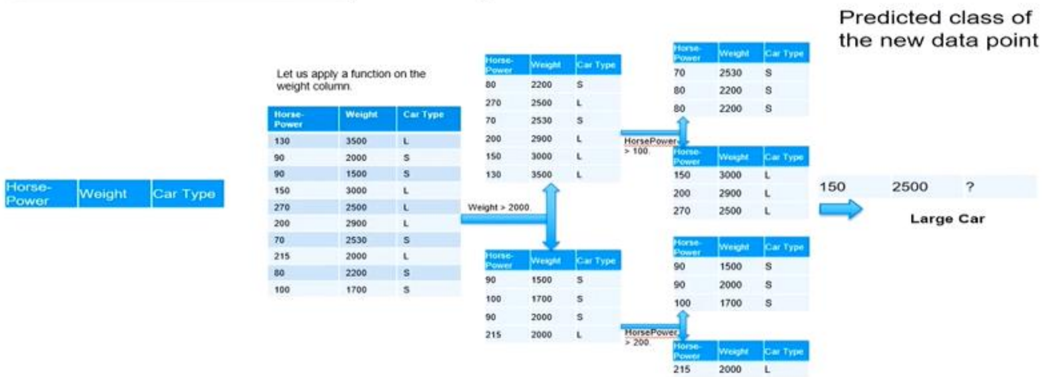
Note: The CART algorithm employed by scikitlearn creates only binary tree i.e. each node is split into two subnodes

Decision Trees Structure - Predicting

Horse-Power	Weight	Car Type
150	2500	?

Predicted class of the new data point

Decision Trees Structure - Predicting



Decision Trees Structure – Training Errors

1. Suppose we come across a combination of "Weight" and "Horse-Power" for any of the classes which was not available in the training data on which the decision tree was built. For e.g. a Small Car with HorsePower of 250 and Weight is 2000. The Decision Tree will classify it as Large Car.

2. Such classification errors can occur both during the training and testing. They are called training errors and testing errors. This is true for any algorithm

3. The decision tree algorithm by default will try to build a tree where the smallest child nodes are perfectly homogenous in the target columns

4. To achieve perfect homogeneity in the target column, the algorithm may build a large tree where each leaf has only 1 record!!! Such models are overfit models. They give zero errors on training but perform poorly on test data

Decision Trees – Posterior Probability

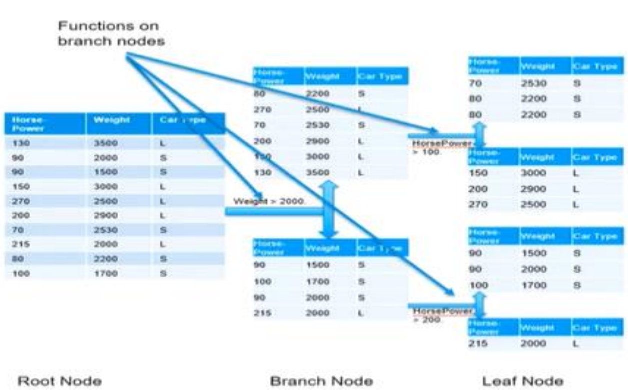
5. Sometimes when the algorithm runs out of independent attributes to use to break a node into smaller nodes or it is forced to stop, we may find nodes where the target column is not homogenous

Small Cars

Horse-Power	Weight	Car Type
90	1500	S
100	1700	S
90	2000	S
215	2000	L

6. In such case, the label assigned to the node is based on majority class and the ratio of the classes indicates the posterior probability of the two classes at that node. P(s) = ¾ and P(L) = ¼

Decision Trees – Structure & Node types



7. 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:

8. 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)

How does decision tree algorithm select the column and the breakpoint in the column to create the tree?

Decision Trees – Learning Process / Loss function & impurity

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 reduction in impurity of the target column i.e. increase in the homogeneity of the target column at every split of the given data

3. To understand the loss function we need to understand how it computes the impurity / purity of the target column. There are two measures of impurity viz Entropy and Gini…

Decision Trees – Learning Process / Loss function & impurity

Decision Trees – Learning Process / Loss function & impurity

1. There is a bag of 50 balls of orange and white respectively
2. You have to pull out one ball from the bag with closed eyes. If the ball is -
   a. orange, explain the linear regression model
   b. white, nothing to be done
3. This state where you have to decide and your decision can result in multiple outcomes with equal probability is said to be state of maximum uncertainty

4. If you have a bag full of balls of only one colour, then there is no uncertainty. You know what is going to happen. Uncertainty is zero.
5. Thus, the more the homogeneity, lesser the uncertainty and vice versa
6. Uncertainty is expressed as entropy or Gini index

Entropy … a measure of uncertainty

a. Given that there are two possible outcomes (orange or white) for a given action (picking the ball)

b. Given that we know the probability of getting each outcome (P(Orange) = .5 and P(White) = 1 - .5 = .5)

c. We also know that when only one outcome (White or Orange) is possible, there is no uncertainty. P(White = 1) and P(Orange = 1-1 = 0)

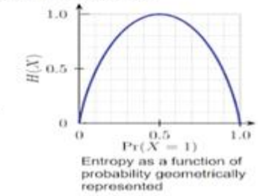
d. We can express the relation between probability and impurity of target column in a mathematical form

Decision Trees – Shannon's Entropy

a. Imagine a bag contains 6 Orange and 4 White balls. Let the two classes Orange -> class 0 and White -> class 1. The probability ranges from 0 to 1

b. The impurity of the bag can be represented as a log to base 2 of probability of a class (pi). Impurity ranges from 0 to 1 (for binary classification). The relation between probability and impurity can be expressed as -

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



Entropy as a function of probability geometrically represented

c. H(X), entropy of the bag (X) will be calculated as per the formula
   a. The index i refers to the number of classes possible
   b. H(X) = - (0.6 * log2(0.6)) - (0.4 * log2(0.4)) = 0.9709506

d. Suppose we remove all Orange 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%

Decision Trees – Shannon's Entropy

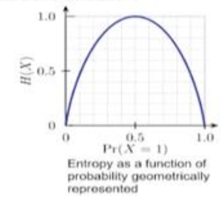
a. Imagine a bag contains 6 Orange and 4 White balls. Let the two classes Orange -> class 0 and White -> class 1. The probability ranges from 0 to 1

b. The impurity of the bag can be represented as a log to base 2 of probability of a class (pi). Impurity ranges from 0 to 1 (for binary classification). The relation between probability and impurity can be expressed as -

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

$\log_2(1) = 0$

$\log_2(.5) = 1$



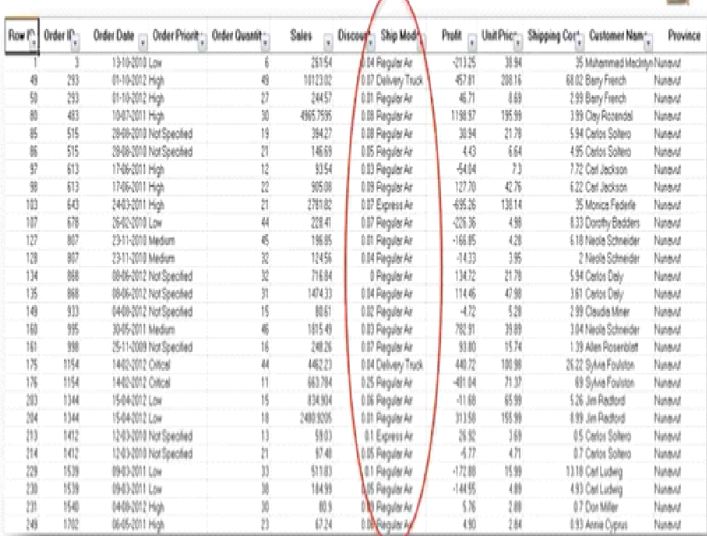
Entropy as a function of probability geometrically represented

c. H(X), entropy of the bag (X) will be calculated as per the formula
   a. The index i refers to the number of classes possible
   b. H(X) = - (0.6 * log2(0.6)) - (0.4 * log2(0.4)) = 0.9709506

d. Suppose we remove all Orange 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%

Decision Trees – Entropy & Objective

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



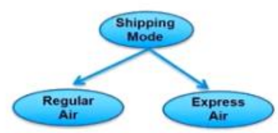
Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
1	3	13-10-2010	Low	6	261.54	0.04	Regular Air	-213.25	38.94	35	Muhammed MacIntyre	Nunavut
49	293	01-10-2012	High	49	10123.02	0.07	Delivery Truck	457.81	208.16	68.02	Barry French	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
103	643	24-03-2011	High	21	2781.82	0.07	Express Air	-695.26	138.14	35	Monica Federle	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
175	1154	14-02-2012	Critical	44	4462.23	0.04	Delivery Truck	440.72	100.98	26.22	Sylvia Foulston	Nunavut
176	1154	14-02-2012	Critical	11	663.784	0.25	Regular Air	-481.04	71.37	69	Sylvia Foulston	Nunavut
203	1344	15-04-2012	Low	15	834.904	0.06	Regular Air	-11.68	65.99	5.26	Jim Radford	Nunavut
204	1344	15-04-2012	Low	18	2480.9205	0.01	Regular Air	313.58	155.99	8.99	Jim Radford	Nunavut
213	1412	12-03-2010	Not Specified	13	59.03	0.1	Express Air	26.92	3.69	0.5	Carlos Soltero	Nunavut
214	1412	12-03-2010	Not Specified	21	97.48	0.05	Regular Air	-5.77	4.71	0.7	Carlos Soltero	Nunavut
229	1539	09-03-2011	Low	33	511.83	0.1	Regular Air	-172.88	15.99	13.18	Carl Ludwig	Nunavut
230	1539	09-03-2011	Low	38	184.99	0.05	Regular Air	-144.55	4.89	4.93	Carl Ludwig	Nunavut
231	1540	04-08-2012	High	30	80.9	[illegible]	Regular Air	5.76	2.88	0.7	Don Miller	Nunavut
249	1702	06-05-2011	High	23	67.24	[illegible]	Regular Air	4.90	2.84	0.93	Annie Cyprus	Nunavut

Sales Data

Decision Trees – Entropy & Objective

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 Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
7933	56706	10-11-2009	High	7	754.92	0	Regular Air	-129.57	99.99	19.99	Adrian Shami	Alberta
8007	57216	29-07-2010	High	48	109.78	0.03	Regular Air	4.48	2.2	1.2	Ruben Dartt	Alberta
8017	57281	20-04-2010	High	26	561.66	0.07	Regular Air	-111.33	22.38	15.1	Andy Yotov	Alberta
806	5767	28-04-2012	High	31	2390.54	0.08	Regular Air	756.10	82.99	5.5	Jessica Myrick	Alberta
1	3	13-10-2010	Low	6	261.54	0.04	Regular Air	-213.25	38.94	35	Muhammed MacIntyre	Nunavut
107	679	26-02-2010	Low	44	228.41	0.07	Regular Air	-226.36	4.98	8.33	Dorothy Badders	Nunavut
203	1344	15-04-2012	Low	15	834.904	0.06	Regular Air	-11.68	65.99	5.26	Jim Radford	Nunavut
204	1344	15-04-2012	Low	18	2480.9205	0.01	Regular Air	313.58	155.99	8.99	Jim Radford	Nunavut
229	1539	09-03-2011	Low	33	511.83	0.1	Regular Air	-172.88	15.99	13.18	Carl Ludwig	Nunavut
230	1539	09-03-2011	Low	38	184.99	0.05	Regular Air	-144.55	4.89	4.93	Carl Ludwig	Nunavut
256	1792	08-11-2010	Low	28	370.48	0.04	Regular Air	-5.45	13.48	4.51	Carlos Soltero	Nunavut
381	2631	23-09-2010	Low	27	1078.49	0.08	Regular Air	252.66	40.96	1.99	Jack Garza	Nunavut
755	5409	08-01-2012	Low	11	48.91	0.01	Regular Air	-7.04	3.98	2.97	Don Jones	Nunavut
821	5925	12-11-2011	Low	44	3822.42	0.08	Regular Air	-354.90	92.23	39.61	Joy Bell	Nunavut



Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
6555	46599	13-09-2010	Not Specified	4	284	0.01	Express Air	208.21	62.18	10.84	Victoria Brennan	Alberta
7118	50726	31-10-2011	Low	21	98.88	0.07	Express Air	36.01	4.13	0.99	Ruben Dartt	Alberta
7288	51872	23-04-2012	Medium	11	812.498	0	Express Air	128.95	65.99	1.25	Shui Tom	Alberta
7658	54913	03-07-2010	High	11	54.61	0.06	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
7888	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

Decision Trees – Entropy & Objective

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 Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
7933	56706	10-11-2009	High	7	754.92	0	Regular Air	-129.57	99.99	19.99	Adrian Shami	Alberta
8007	57216	29-07-2010	High	48	109.78	0.03	Regular Air	4.48	2.2	1.2	Ruben Dartt	Alberta
8017	57281	20-04-2010	High	26	561.66	0.07	Regular Air	-111.33	22.38	15.1	Andy Yotov	Alberta
806	5767	28-04-2012	High	31	2390.54	0.08	Regular Air	756.10	82.99	5.5	Jessica Myrick	Alberta

Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
1	3	13-10-2010	Low	6	261.54	0.04	Regular Air	-213.25	38.94	35	Muhammed MacIntyre	Nunavut
107	679	26-02-2010	Low	44	228.41	0.07	Regular Air	-226.36	4.98	8.33	Dorothy Badders	Nunavut
203	1344	15-04-2012	Low	15	834.904	0.06	Regular Air	-11.68	65.99	5.26	Jim Radford	Nunavut
204	1344	15-04-2012	Low	18	2480.9205	0.01	Regular Air	313.58	155.99	8.99	Jim Radford	Nunavut
229	1539	09-03-2011	Low	33	511.83	0.1	Regular Air	-172.88	15.99	13.18	Carl Ludwig	Nunavut
230	1539	09-03-2011	Low	38	184.99	0.05	Regular Air	-144.55	4.89	4.93	Carl Ludwig	Nunavut
256	1792	08-11-2010	Low	28	370.48	0.04	Regular Air	-5.45	13.48	4.51	Carlos Soltero	Nunavut
381	2631	23-09-2010	Low	27	1078.49	0.08	Regular Air	252.66	40.96	1.99	Jack Garza	Nunavut
755	5409	08-01-2012	Low	11	48.91	0.01	Regular Air	-7.04	3.98	2.97	Don Jones	Nunavut

Decision Trees – Entropy & Objective

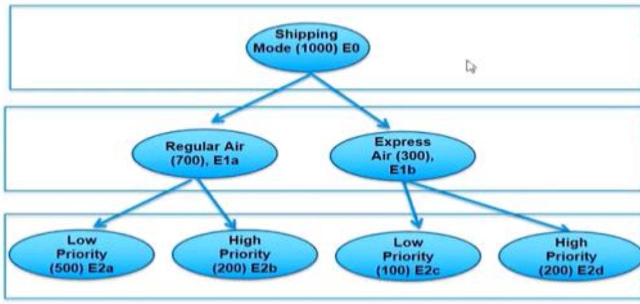
Suppose we wish to find if there was any influence of shipping mode, order priority on customer location. Customer location is target column and like the bag of coloured balls

Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
7933	56706	10-11-2009	High	7	754.92	0	Regular Air	-129.57	99.99	19.99	Adrian Shami	Alberta
8007	57216	29-07-2010	High	48	109.78	0.03	Regular Air	4.48	2.2	1.2	Ruben Dartt	Alberta
8017	57281	20-04-2010	High	26	561.66	0.07	Regular Air	-111.33	22.38	15.1	Andy Yotov	Alberta
806	5767	28-04-2012	High	31	2390.54	0.08	Regular Air	756.10	82.99	5.5	Jessica Myrick	Alberta

Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	Province
1	3	13-10-2010	Low	6	261.54	0.04	Regular Air	-213.25	38.94	35	Muhammed MacIntyre	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

Decision Trees – Entropy & Objective



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

Decision Trees – Loss functions

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

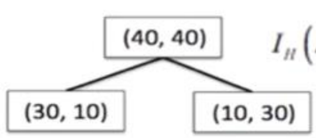
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

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

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

Decision Trees - Information Gain using Entropy

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



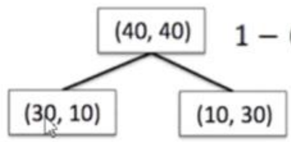
$$I_H(D_p) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = 1$$

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

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

Decision Trees - Information Gain using Gini index

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

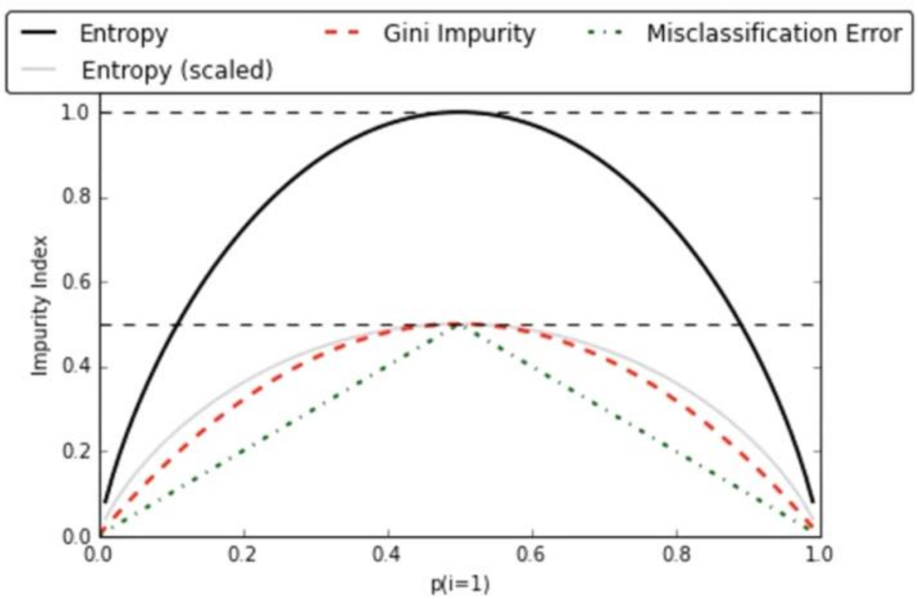


$$1 - (0.5^2 + 0.5^2) = 0.5$$

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

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

Decision Trees – Gini , Entropy , Misclassification Error



Note: Misclassification Error is not used in Decision Trees

Decision Trees -

Advantages -

1. Simple , Fast in processing and effective
2. Can work with missing data
3. Handles numeric and categorical variables
4. Interpretation of results is easier when represented as rules

Dis-advantages -

1. Often biased towards 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. Decision trees tend to become overfit by default creating large complex trees
5. Large trees can be difficult to interpret

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

Help

Friday, August 23, 2019 9:21 PM

https://scikit-learn.org/stable/modules/tree.html