Decision Tree

20 September 2024 10:23

Gini index → measure of inequality earlier. 0 no inequality, 1 inequality 100%.

→ Lorenz curve — index = area b/w 45% line ($y=x$) and the curve.

- less area → curve aligns with $y=x$ → less inequality
- more area → curve no align w/ $y=x$ → more inequality

area ∝ inequality

area estimated by trapezoidal rule

Decision Tree:

dataset — S → n classes.

→ $S = n_1 + n_2$ (binary classific^n)

$$G(S) = \frac{n_1}{S} \cdot G(S_1) + \frac{n_2}{S} \cdot G(S_2)$$

→ $G(s) = 1 - \sum p_i^2$ (relative freq of class i)

Ex: ① Unbiased coin:

$P_{head} = 0.5 \quad P_{tail} = 0.5$

$$G(S) = 1 - \sum p_i^2 = 1 - [(0.5)^2 + (0.5)^2] = 0.5$$

② Biased coin:

$P_{head} = 0.3 \quad P_{tail} = 0.7$

$$G(S) = 1 - \sum p_i^2 = 1 - [(0.3)^2 + (0.7)^2] = 0.42$$

	home owns	Married	Gender	Employed?	Credit Rate	Risk class
1	Yes	Yes	M	Yes	A	B ✓
2	NO	NO	F	Yes	A	A ✓
3	Yes	Yes	F	Yes	B	C
4	Yes	NO	M	NO	B	B ✓
5	NO	Yes	F	Yes	B	C
6	NO	NO	F	Yes	B	A ✓
7	NO	NO	M	NO	B	B ✓
8	Yes	NO	F	Yes	A	A ✓
9	NO	Yes	F	Yes	A	C
10	Yes	Yes	F	Yes	A	C

class

$S = 10 \mid A=3 \quad B=3 \quad C=4$

① $G(S)$ (gini index for overall dist.) $= 1 - \sum p_i^2$ (here, final class layer.) $= 1 - \left[\left(\frac{3}{10}\right)^2 + \left(\frac{3}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right]$

$G(S) = 0.66$

→ Go attribute by attribute now.

② Attribute: Home

n_1 ← (5) Y: A = 1 B = 2 C = 2

n_2 ← (5) N: A = 2 B = 1 C = 2

$$G(Yes) = 1 - \sum p_i = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.64$$

$$G(No) = 1 - \sum p_i = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.64$$

$$G(home) = \frac{n_1}{S} G(S_1)^{Yes} + \frac{n_2}{S} G(S_2)^{No}$$

$= 5 \ldots = 0.64$

③ Attribute: Married

(5) Y: A = 0 B = 1 C = 4

③ Attribute: Married

(5) Y: A=0 B=1 C=4

(5) N: A=3 B=2 C=0

$G(yes) = 1 - \sum p_i = 1 - \left[\left(\frac{0}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right] = 0.32$

$G(no) = 1 - \sum p_i = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(\frac{0}{5}\right)^2\right] = 0.48$

$G(married) = \frac{n_1}{S} G(S_1) + \frac{n_2}{S} G(S_2) = \frac{5}{10}(0.32) + \frac{5}{10}(0.48)$

$= 0.4$

⑤ Attribute: Employed

(8) Y: A=3 B=1 C=4

(2) N: A=0 B=2 C=0

$G(Y) = 1 - \sum p_i^2 = 1 - \left[\left(\frac{3}{8}\right)^2 + \left(\frac{1}{8}\right)^2 + \left(\frac{4}{8}\right)^2\right] = 0.594$

$G(N) = 1 - \sum p_i^2 = 1 - \left[\left(\frac{2}{2}\right)^2\right] = 0$

$G(employed) = \frac{8}{10}(0.594) + \frac{2}{10}(0) = 0.475$

$G(home) = \frac{n_1}{S} G(S_1) + \frac{n_2}{S} G(S_2)$

$= \frac{5}{10}(0.64) + \frac{5}{10}(0.64) = 0.64$

④ Attribute: Gender

(3) M: A=0 B=3 B=0

(7) F: A=3 B=0 C=4

$G(male) = 1 - \sum p_i = 1 - \left[0^2 + \left(\frac{3}{3}\right)^2 + 0\right] = 0$

$G(female) = 1 - \sum p_i^2 = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.511$

$G(gender) = \frac{n_1}{S} G(S_1) + \frac{n_2}{S} G(S_2)$

$= \frac{3}{7}(0) + \frac{4}{7}(0.511) = 0.358$

⑥ Attribute: Credit rate

(5) a: A=2 B=2 C=2

(5) b: A=1 B=1 C=2

Same as Home

G(credit) = 0.64

⑦ Gini Index Gain:

Attribute	Gini index before split	Gini index after split	Gain
Home	0.66	G(home)=0.64	0.02
Married	0.66	G(married)=0.40	0.26
Gender	0.66	G(gender)=0.358	0.302
Employed	0.66	G(employed)=0.475	0.185
Credit	0.66	G(credit)=0.64	0.02

G(S): overall gini index of full dataset S.

⑧ Highest info gain @ gender → split there first.

Now another example, but w/ info gain + entropy.

NO	Age	Income	Profession	Watch See Movies
1	L ~~<30~~	LOW	Busines	Yes
2	H ~~>30~~	HIGH	Engg	NO
3	L ~~<30~~	MED	Engg	Yes
4	L ~~<30~~	LOW	Agri	NO
5	H ~~>30~~	HIGH	Busines	Yes
6	H ~~>30~~	MED	Agri	NO

① Overall entropy:

S = 6 | Y = 3 N = 3

$H(S) = -\sum p_i \log p_i = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right) = 1$

② Attribute: Age

(3) Low: Y = 2 N = 1

(3) High: Y = 1 Y = 2

$H(low) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.918$

5	~~H~~	High	Business	Yes
6	~~H~~	MED	Agri	NO

Engineer : 23y, low income — movie?

② Attribute: Income

(2) Low : Y = 1 N = 1
(2) Med : Y = 1 N = 1
(2) High : Y = 1 N = 1

$H(low) = H(med) = H(high)$

$= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$

$= 1 \times -\frac{1}{2}\log\left(\frac{1}{2}\right) = -1\log_2(2)^{-1} = 1$

Take weighted : (weights are frac. of L, M, H)

$H(income) = \frac{2}{6}(1) + \frac{2}{6}(1) + \frac{2}{6}(1) = 1$

⇒ Information gain = $1 - H(income) = 0$

$H(low) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.918$

$H(high) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.918$.

Now take weighted
(weight is fraction of L, H)

$H(Age) = \frac{3}{6}(H(low)) + \frac{3}{6}(H(low)) = 0.918$

Entropy for age
↳ Information gain = $1 - 0.918 = 0.082$

③ Attribute : Profession

(2) Busi : Y = 2 N = 0
(2) Agri : Y = 0 N = 2
(2) Eng : Y = 1 N = 1

$H(busi) = H(agri) = \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$

$H(eng) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$

Take weighted : weights from B, A, Σ

$H(prof) = \frac{2}{6}(0) + \frac{2}{6}(0) + \frac{2}{6}(1) = 0.33$

⇒ Info gain = $1 - 0.33 = 0.67$

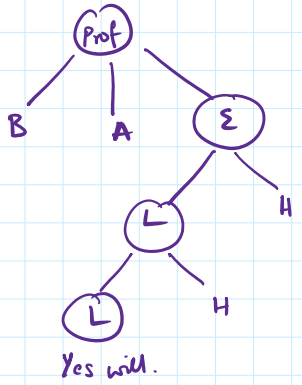
④

Attribute	Entropy	Information
Age	0.918	0.082
Income	1	0
Profession	0.33	0.67

↳ Least entropy, Most info gain] 1st split

Next is age.

Last split is income
↳ doesn't really matter.



Yes will.

Next Example.

1. Pizza Dataset : using this dataset, draw possible trees without any heuristics

(a) how many decision trees are at least possible,

(b) what is the hypothesis space ?

	Example	Crust Size	Shapes	Filling Size	Class
1	e_1	Big	Circle	Small	Pos
2	e_2	Small	Circle	Small	Pos
3	e_3	Big	Square	Small	Neg
4	e_4	Big	Triangle	Small	Neg
5	e_5	Big	Square	Big	Pos
6	e_6	Small	Square	Small	Neg
7	e_7	Small	Square	Big	Pos
8	e_8	Big	Circle	Big	Pos

AOML (Decision Tree) -Mathematics for Machine Learning Bhaskarjyoti Das (Dept of CSE in AI&ML) PES UNIVERSITY

(c) using the above dataset, use Entropy Gain

N binary attributes → 2^N trees

M n-ary attributes → n^m trees (n = any of attr., m = no. of)

→ Entire set of possible trees.

(a) Overall entropy (P : 5, N : 3) $-\frac{5}{8}\log\left(\frac{5}{8}\right) - \frac{3}{8}\log\left(\frac{3}{8}\right) = H(S)$

Rest is same as prev. example.

ⓒ using the above dataset, use Entropy Gain and find out which attribute to split first.

One more, no table this time.

Q2 Consider Training Data Set S with 100 instances and 2 attributes — Height & Weight. S has 70 +ve and 30 -ve examples

Values (Height) = {Med, Tall}
Values (Weight) = {Normal, Fat}

$S_{Med} = [0+, 60-]$ $S_{Tall} = [20+, 20-]$
$S_{Normal} = [35+, 35-]$ $S_{Fat} = [30+, 0-]$

Decide which attribute partitions Sample S better.

100 instances $\Rightarrow S = 100$
$P = 70$ $N = 30$

Height / Med, Tall Weight / Normal, Fat

① Attribute: Height
(60) Med : $P = 0$ $N = 60$
(40) Tall : $P = 20$ $N = 20$

$H(med) = -\frac{0}{60}\log(0) - \frac{60}{60}\log(1) = 0$

$H(tall) = -\frac{20}{40}\log\left(\frac{20}{40}\right) - \frac{20}{40}\log\left(\frac{20}{40}\right) = 1$

Weighted: (weight is fract. of M,T)

$H(height) = \frac{60}{100}[H(med)] + \frac{40}{100}[H(tall)] = 0.4$

Info gain (height) = 0.6

② Attribute: weight:
(70) Normal: $P = 35$ $N = 35$
(30) Fat: $P = 30$ $N = 0$

$H(Norm) = -\frac{35}{70}\log\left(\frac{35}{70}\right) - \frac{35}{70}\log\left(\frac{35}{70}\right) = 1$

$H(Fat) = -\frac{30}{30}\log\left(\frac{30}{30}\right) - \frac{0}{30}\log\left(\frac{0}{30}\right) = 0$

Weighted: (weight is fraction of N, F)

$H(weight) = \frac{70}{100}[H(Norm)] + \frac{30}{100}[H(fat)] = 0.7$

Info gain (weight) = 0.3

③

Attribute	Entropy	Info gain
Height	0.4	0.6
Weight	0.7	0.3

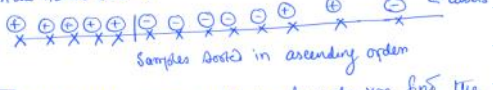
Most info gain + least entropy.

Next example:

Q3 Decision Tree with Numerical Attribute that is Continuous

In a dataset, if the continuous attribute is sorted in ascending order, it looks like below where the labels are ⊕ and ⊖

On the number line ⇒



Q. using this numeric attribute, how do you find the best split points ?

Key: * Since continuous number, first sort in ascending order
* Then check where the 'label' flips
for example, between 5th & 6th sample,
10th & 11th sample
12th & 13th sample

* Here we have 3 split points possible
$\theta_1, \theta_2, \theta_3$
* using Information Gain, evaluate these $\theta_1, \theta_2, \theta_3$

8) Numerical + categorical Attribute Mix

↳ Find "Best Split" using Gini Impurity

Dataset

Age	Like Fruits	Like KFC	Going to be Healthy
51	1	1	1
42	0	0	0
44	1	1	1
30	1	1	1
24	0	0	0
36	0	1	1
36	0	0	0
47	1	1	1
46	1	0	0
47	0	1	0

Hints
1. Numerical attribute → Age. So, ascending order arrange
2. Find the possible split points of every adjacent pair
   ↳ For each calculate Gini heuristic
3. Do the same for Categorical

Effectively convert Numerical → Categorical or Boolean.

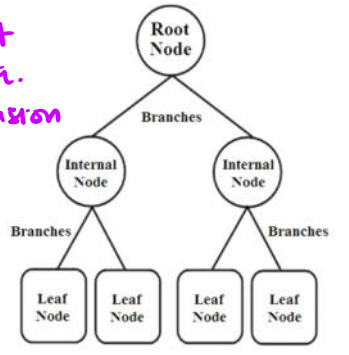
4. Choose the Best Split

Some stuff to remember:

1. Small tree >> Big tree — interpretability; large trees overfit
   → Some choices lead to smaller trees
   ↳ attributes differ in info
   ↓
   ∴ Info gain / Gini index
2. Decision tree philosophy: split training data into small subsets.
3. Hypothesis space = space of trees — Hill climbing search.
   ↓ Heuristic
4. DISJUNCTION (∪) OF CONJUNCTIONS (∩)

Final takeaways:

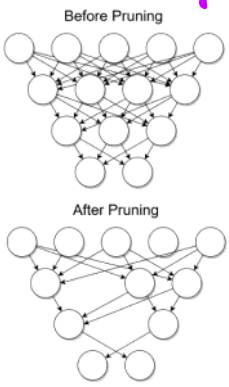
1. DT = flowchart-like struct
   - root node = entire dataset
   - internal node = tests on attr.
   - branch = outcome of decision
   - leaf node = final outcome.



2. Adv:
   a. highly interpretable
   b. little data prep.
   c. classificⁿ + regression = versatile.
3. Disadv:
   a. more levels dominate — bias.
   b. unstable — small variations = diff tree entirely.
   c. overfit — hard to generalize

Before Pruning

b. unstable — small variations = diff tree entirely.
c. overfit — hard to generalize →



Pruning
- reduce no. of connections + nodes.
  a. Prepruning (early stopping)
     · stop tree from growing once criteria met.
  b. Post pruning
     · remove branches from fully grown tree

④ ID3 algorithm (Iterative dichotomizer 3) — what we did w/ entropy + info gain.
  a. Inductive bias
     - describe basis by which it chooses one hypothesis
       · simple to complex heuristic search (use gain)
       · shorter trees >> longer trees — occam's razor.
       · attr. of higher gain preferred.
       · attr. w/ many values preferred

  b. Hypothesis space = all possible DT that can be made.

  c. Hyperparameter = maximum depth of the tree.
       · too low = generalization
       · too high = memorization.

  d. Advantages = · simple to understand
       · requires little training data.
       · discrete + continuous attr.

  e. Disadvantages = · overfitting
       · not effective w/ many attr.

class of target — predict cont. variables value

⑤ CART (Classifn + Regression Trees) → Binary tree — gini index.
  a. Advantages = · simple
       · numerical + categorical data
       · missing values — impute w/ surrogate splits
       · multiclass classification

  b. Disadvantages = · overfitting
       · greedy (may not find most optimal)
       · biased towards attr. w/ many categories