upGrad

#LifeKoKaroLift

→ i) Homogenity M → Gini index, Information gain

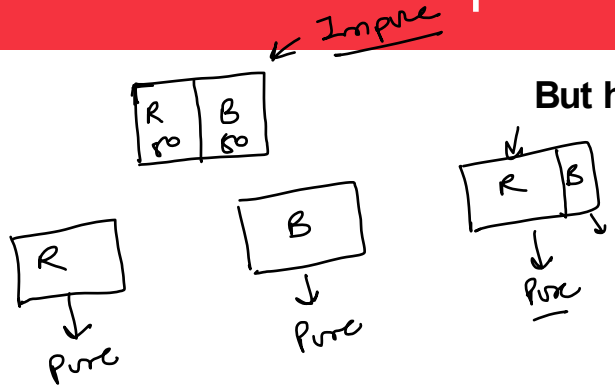
→ ii) Random forest → Bagging, OOB Error

Tree Models: Concepts and Doubts

Tree Models: Concepts

upGrad

But how the split happens?



Homogenity measure → quantifying → Impurity

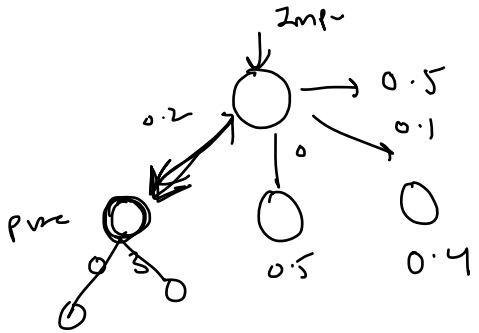
Gini | Entropy

Impure → NO clarity / no Rules

Purity (max. purity)

→ Impure → achieve maximum purity

Classification Split — Gini index / Entropy gain



Tree Models: Concepts

upGrad

Yes	9
No	5

$\sum p_i^2$

$\Delta G_B - G_A =$ maximum

	Yes	No	
Rainy	2	3	5
overcast	4	0	4
sunny	3	2	5

Gini Index

Befor split

$$\boxed{\sum p_i(1-p_i)} = \boxed{1-\sum_{i=1}^{n} p_i^2}$$

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$$G(\text{play Golf}) = 1-\left[\left(\frac{9}{14}\right)^2+\left(\frac{5}{14}\right)^2\right]$$
$$= 1-[0.41+0.12]$$
$$= 0.47$$

$$G(\text{outlook}) = \frac{5}{14}\left[1-\left\{\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right\}\right]+\frac{4}{14}\left[1-\left\{\left(\frac{4}{4}\right)^2\right\}\right]+\frac{5}{14}\left[1-\left\{\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right\}\right]$$

$$\Rightarrow \frac{5}{14}[1-0.52]+\frac{4}{14}\times 0+\frac{5}{14}\times[1-0.52]$$

$$\Rightarrow 0.171+0+0.171 = 0.342$$

$$\Delta(0.47-0.34) = 0.13$$ max info.

Poll Question

upGrad

	C_0	C_1	
Family	1	3	= 4
Sports	8	0	= 8
Luxury	1	7	= 8

	C_0	C_1	
Male	6	4	= 10
Female	4	6	= 10

In the following questions we will try to calculate Gini Index

Question-1: Compute the Gini Index of the overall training example.

Question-2: Calculate the Gini Index of the gender attribute

Question-3: Calculate the Gini Index of the car_type attribute

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

$$G(\text{class}) = 1 - \left[\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right] = 0.5$$

$$G(\text{Gender}) = \frac{10}{20}\left[1 - \left\{\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right\}\right] + \frac{10}{20}\left[1 - \left\{\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right\}\right]$$

$$= 0.48$$

	C_0	C_1	
Small	3	2	5
medium	3	4	7
Large	2	2	4
XL	2	2	4

Solution

upGrad

$$G(\text{car-type}) = \frac{4}{20}\left[1-\left\{\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right\}\right] + \frac{8}{20}\left[1-\left(\frac{8}{8}\right)^2\right] + \frac{8}{20}\left[1-\left\{\left(\frac{1}{8}\right)^2+\left(\frac{7}{8}\right)^2\right\}\right]$$

$$= 0.1625$$

$$G(\text{shirt size}) =$$

$$\frac{5}{20}\left[1-\left\{\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right\}\right] + \frac{7}{20}\left[1-\left\{\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right\}\right] + \frac{4}{20}\left[1-\left\{\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right\}\right]$$

$$+ \frac{4}{20}\left[1-\left\{\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right\}\right]$$

$$= 0.49$$

$$\left.\begin{aligned} \Delta(0.5-0.48) &= 0.02 \\ \Delta(0.5-0.1625) &= 0.33 \\ \Delta(0.5-0.49) &= 0.01 \end{aligned}\right\} \rightarrow \checkmark$$

Poll Question

upGrad

$$\text{Entropy} = \left(-p_i \log_2 p_i\right)$$

Question-4: Calculate the Gini Index of the shirt_size attribute

Question-5: Among all the attributes, which attribute is better to use for the first split in te formation of the decision tree?

- 0.52
- 0.55
- 0.54

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Tree Models: Concepts

upGrad

→ Use multiple trees
→ Bagging
→ Add bias at each tree and at each node.

Decision Tree ← overfit your training data

→ High variance model
→ Low bias

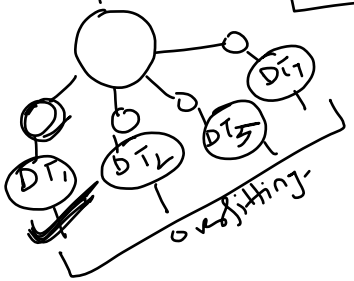
Random Forest

Bagging → Bootstrapped → Creating samples with repetition
→ Aggregation → Combining.

Bootstrapping.

~~Bagging~~ – Level-1 Adding bias

Train



100

9

Train

1	2	3
4	5	6
7	8	9

(1 1 2 6 8 9 9 5 5) B.S1
4 5 5 4 3 2 1 1 9 B.S2
7 6 5 4 7 2 1 3 1 B.S3
BS50

BS1 → DT1
BS2 → DT2
BS3 → DT3
BS7 → DT7
BS50 → DT50

weak learner

i) Level-2 Bias

ii)

A → 80 → Race

A A A A

50 vm

25

I will use those trees for which, it was not the part of the B.Sample

→ At each node a random sample of feature are taken for creating the split

M → M/100 → Total # of features in your data

Agg. [25 < 20 — true, 5 → -ve] → 1 → true.

Tree Models: Concepts

oob error = $\frac{\text{\# of incorrect class}}{\text{Total no of data}}$ $\frac{4}{10} = 0.4$

upGrad

Training data

Random Forest (OOB error)

Prediction	Actual		BS_1 DT_1	BS_2 DT_2	BS_3 DT_3	BS_4 DT_4	BS_5 DT_5	BS_6 DT_6
+ve	+ve	1	×	+ve	×	+ve	−ve	×
−ve	−ve	2	−ve	×	−ve	×	×	+ve
+ve	−ve	3	×	×	+ve	×	+ve	+ve
−ve	+ve	4	−ve	×	−ve	−ve	×	×
+ve	+ve	5	×	+ve	×	+ve	×	+ve
−ve	−ve	6	−ve	×	+ve	×	−ve	×
−ve	+ve	7	+ve	×	×	×	−ve	−ve
−ve	−ve	8	×	−ve	×	−ve	−ve	×
−ve	+ve	9	−ve	×	+ve	×	−ve	×
−ve	+ve	10	×	+ve	×	×	−ve	−ve

upGrad

#LifeKoKaroLift

Thank You!