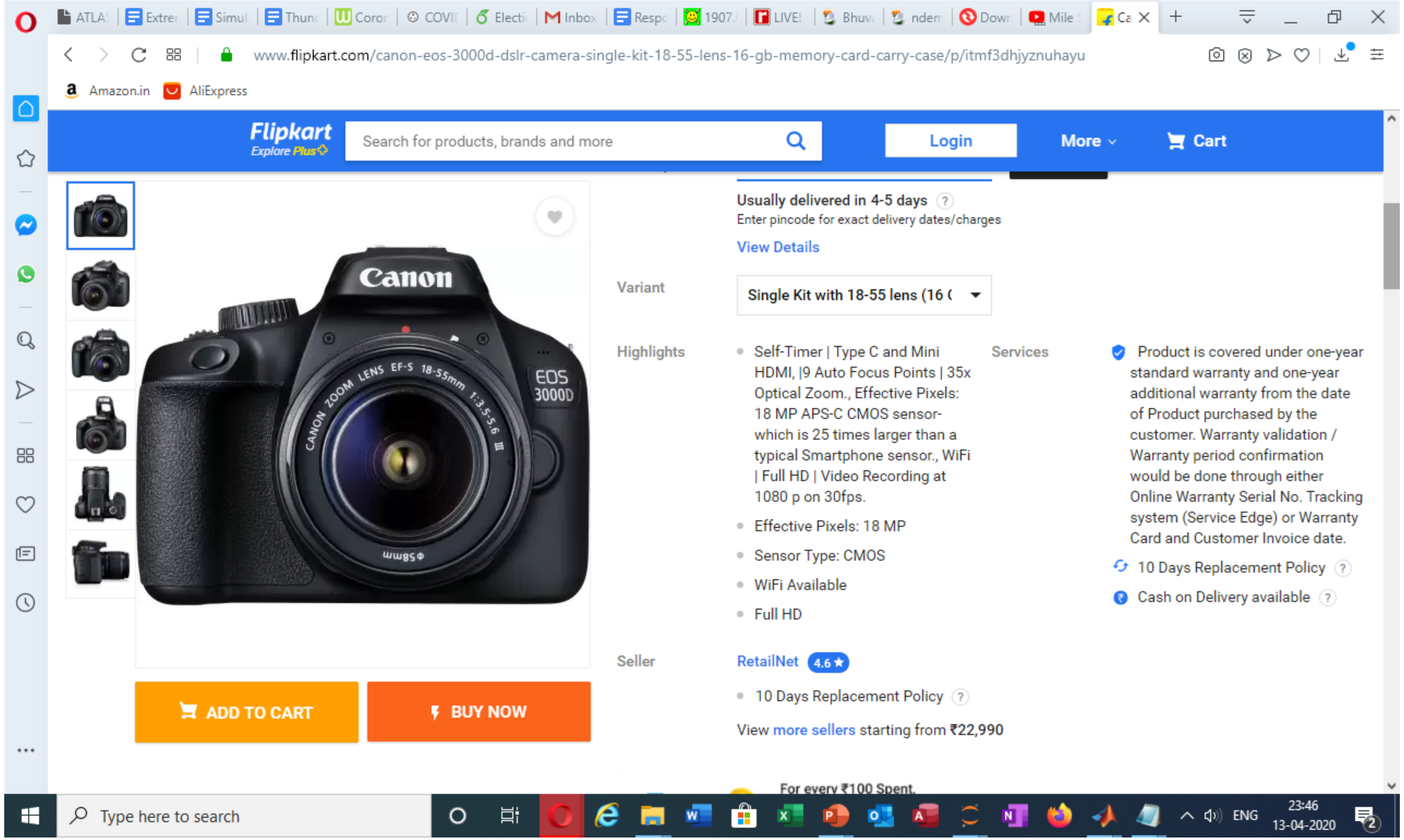
Decision Tree Classification for Rating Prediction

Machine Learning Unit 8

Product Ratings based on Features



www.flipkart.com/canon-eos-3000d-dslr-camera-single-kit-18-55-lens-16-gb-memory-card-carry-case/p/itmf3dhjyznuhayu

Usually delivered in 4-5 days ?

Enter pincode for exact delivery dates/charges

View Details

Variant: Single Kit with 18-55 lens (16 (

Highlights

- Self-Timer | Type C and Mini HDMI, |9 Auto Focus Points | 35x Optical Zoom., Effective Pixels: 18 MP APS-C CMOS sensor-which is 25 times larger than a typical Smartphone sensor., WiFi | Full HD | Video Recording at 1080 p on 30fps.
- Effective Pixels: 18 MP
- Sensor Type: CMOS
- WiFi Available
- Full HD

Services

- Product is covered under one-year standard warranty and one-year additional warranty from the date of Product purchased by the customer. Warranty validation / Warranty period confirmation would be done through either Online Warranty Serial No. Tracking system (Service Edge) or Warranty Card and Customer Invoice date.
- 10 Days Replacement Policy ?
- Cash on Delivery available ?

Seller: RetailNet 4.6 ★

- 10 Days Replacement Policy ?

View more sellers starting from ₹22,990

For every ₹100 Spent,

How much rating will a particular user give this camera out of 5?

Probably depends on features!

Which features does the user like?

Source: Flipkart website

Feature Selection

- The user has exactly 5 options: 1, 2, 3, 4 or 5 stars!
- Her choice depends on the different features of the product!
- But she may consider some features to be more important than others !
- Which features determine her vote?

Company	Color	Resolution	Video Rate	Price	Her Rating
C1	Black	10 MP	25 fps	$200	2
C1	White	15 MP	25 fps	$250	2
C2	White	12 MP	30 fps	$250	4
C1	Black	15 MP	30 fps	$300	3
C2	Black	20 MP	25 fps	$400	3
C2	White	12 MP	50 fps	$500	5
C2	Black	15 MP	30 fps	$250	????

Decision Tree for Feature Selection

- Which features does she consider as important while rating?
- Let’s look at her history of rating 100 cameras!

Rating	Count
1	21
2	24
3	18
4	20
5	17

Overall,
Count=100

Rating	Count
1	15
2	18
3	10
4	5
5	6

Company = C1,
Count=54

Rating	Count
1	6
2	6
3	8
4	15
5	11

Company = C2,
Count=46

Rating	Count
1	15
2	20
3	13
4	12
5	10

Color=Black,
Count=70

Rating	Count
1	6
2	4
3	5
4	8
5	7

Color=White,
Count=30

Which feature is more important for ratings - company or color???

What’s a discriminative feature?

- Prob(Y = HIGH | COMPANY = C1) = 11/54 ~ 0.2 [Easy to decide]
- Prob(Y = HIGH | COMPANY = C2) = 26/46 ~ 0.55

COMPANY: good feature

- Prob(Y = HIGH | PRICE < 300) = 25/70 ~ 0.36
- Prob(Y = HIGH | PRICE >= 300) = 12/30 = 0.4

PRICE<300: bad feature

- Prob(Y = HIGH | PRICE < 500) = 35/90 ~ 0.4
- Prob(Y = HIGH | PRICE >= 500) = 2/10 ~ 0.2 [Easy to decide][Very few examples]

PRICE<500: doubtful feature

Decision Tree Algorithm

- Idea: identify the “most discriminative” feature, use it to classify!

- Problem 1: How to quantify “discriminative-ness”?

- Problem 2: What if no feature is very discriminative?

Decision Tree Algorithm

- Idea: identify the “most discriminative” feature, use it to classify!

- Problem 1: How to quantify “discriminative-ness”?

  - entropy!

- Problem 2: What if no feature is very discriminative?

  - try a sequence of features!

Entropy: measure of discriminativeness

- P(Y=1) = 0.5, p(Y=2) = 0.5 : low discriminative ability
- P(Y=1) = 0.9, p(Y=2) = 0.1 : high discriminative ability

$$\boxed{H = -\sum_{i} p_i(\log_2 p_i)}$$

- Case 1: H = 1 (0.69)
- Case 2: H = 0.5 (0.33)

Feature selection based on entropy

	COMPANY=C1	COMPANY=C2	NO SPLIT
#(Y=LOW)	43	20	63
#(Y=HIGH)	11	26	37
Entropy	0.51	0.68	0.66

- Information gain =

Original Entropy – (Split1_size*Split1_ entropy + Split2_size*Split2_ entropy)

0.66 – (54/100*0.51 + 46/100*0.68) ~ 0.07

Feature selection based on entropy

	PRICE<300	PRICE>=300	NO SPLIT
#(Y=LOW)	45	18	63
#(Y=HIGH)	25	12	37
Entropy	0.65	0.67	0.66

- Information gain =

Original Entropy – (Split1_size*Split1_ entropy + Split2_size*Split2_ entropy)

0.66 – (70/100*0.65 + 30/100*0.67) ~ 0!!

Feature selection based on entropy

	PRICE<500	PRICE>=500	NO SPLIT
#(Y=LOW)	55	8	63
#(Y=HIGH)	35	2	37
Entropy	0.67	0.5	0.66

- Information gain =

Original Entropy – (Split1_size*Split1_ entropy + Split2_size*Split2_ entropy)

0.66 – (90/100*0.67 + 10/100*0.5) ~ 0.01!!

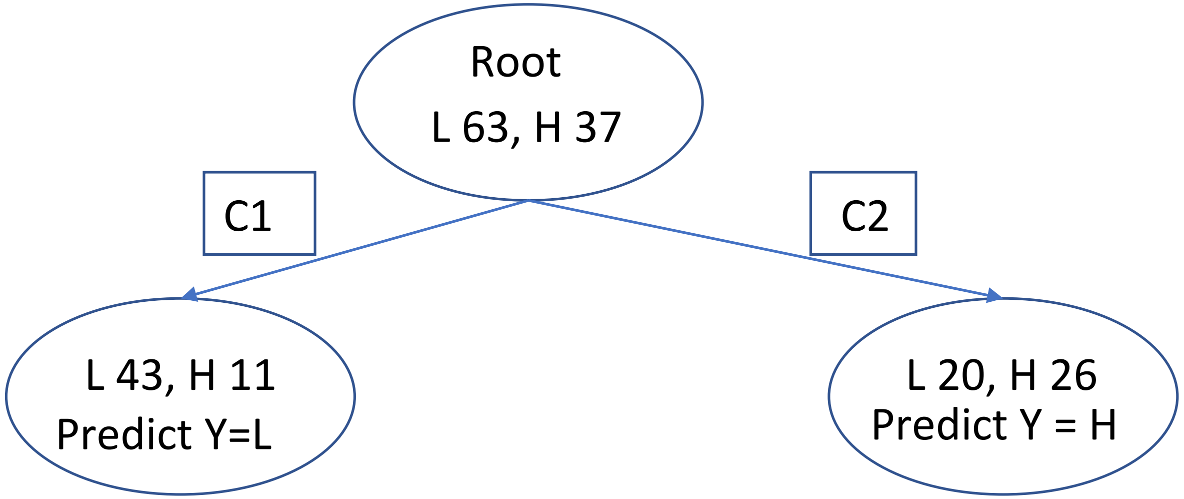
Feature selection based on entropy

- Each discrete feature splits the dataset
- Continuous features can always be converted to discrete
- “Pure” dataset: - disbalanced class distribution

  - low entropy

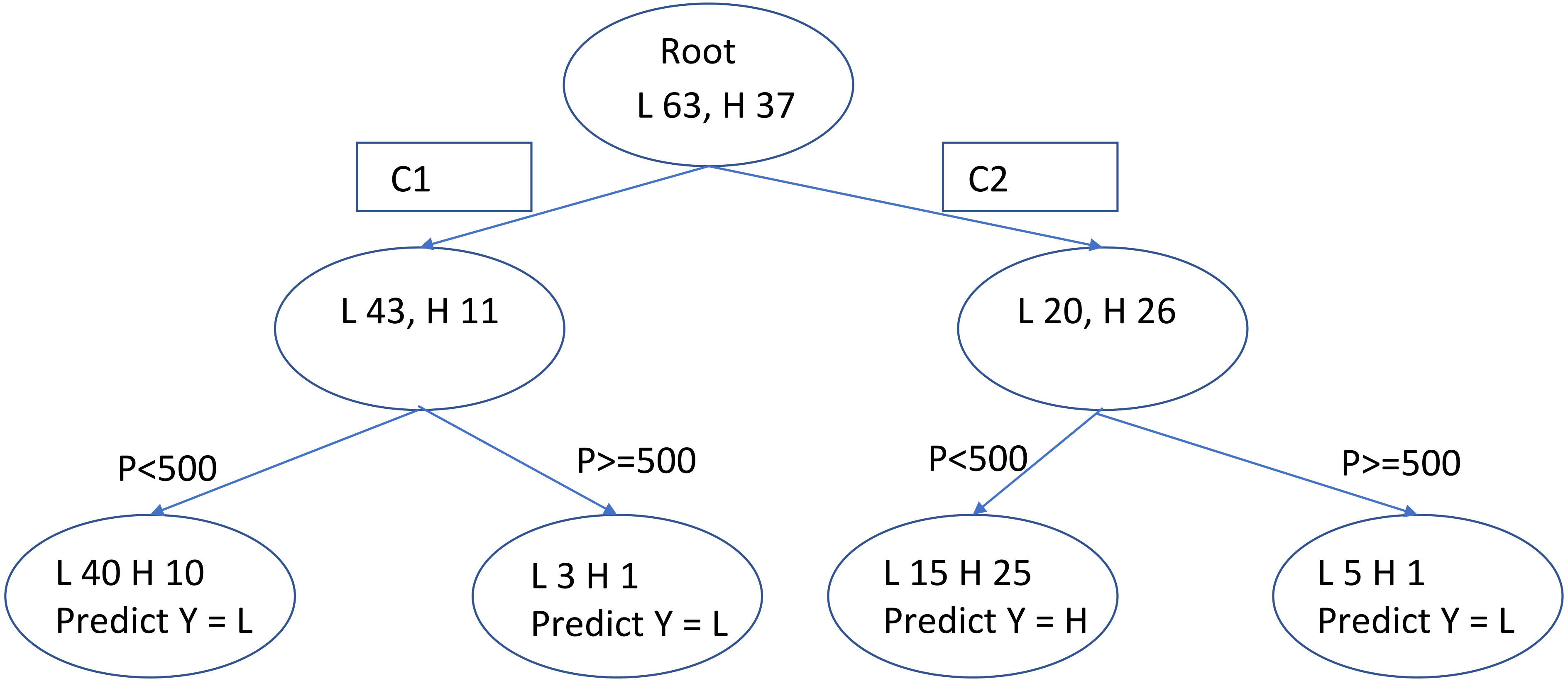
  - high information gain
- Choose that feature which provides most information gain!

Decision Stump



Training accuracy: 43/63 for LOW, 26/37 for HIGH, 69/100 OVERALL

Decision Tree



- Does this split provide “information gain”???
- If yes, split. If no, stop at previous step

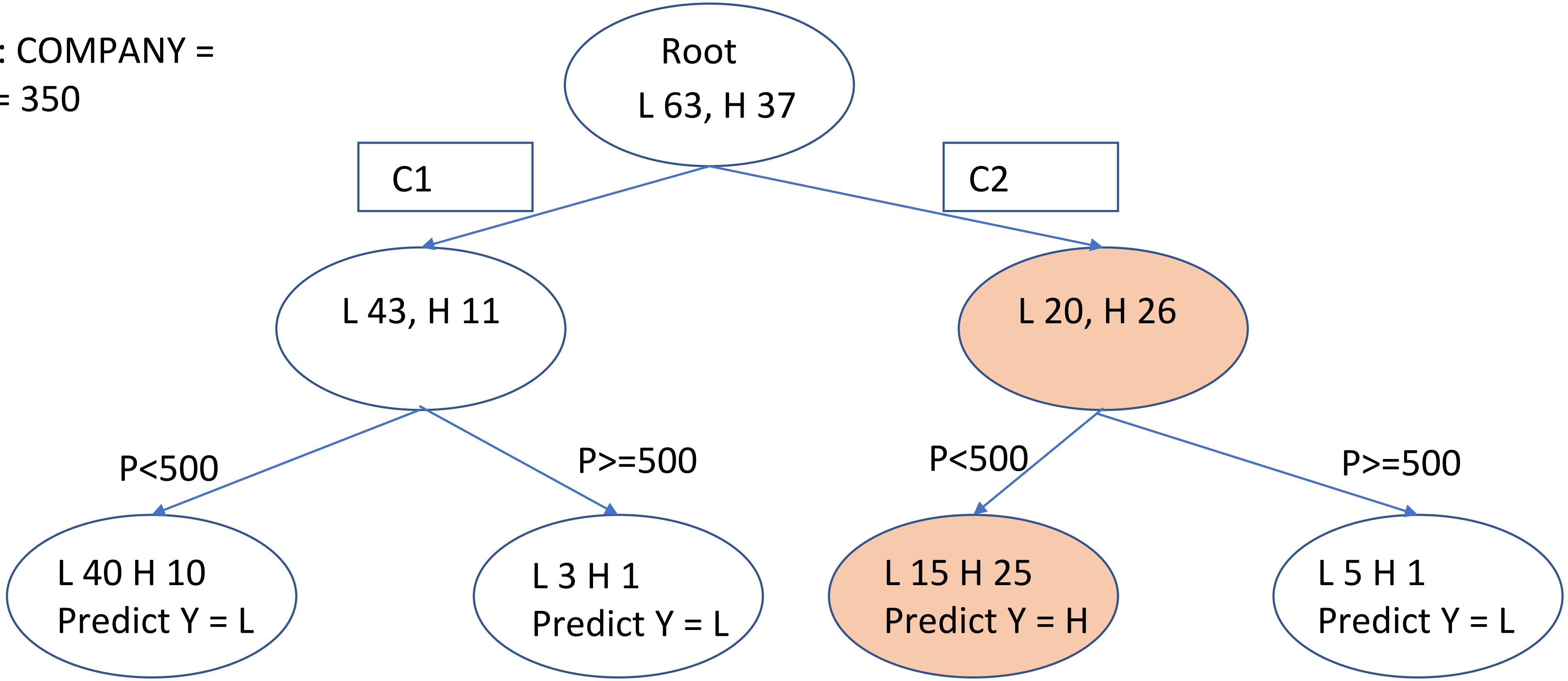
Decision Tree algorithm

- 1. Identify the feature that results in maximum information gain
- 2. Split the dataset accordingly
- 3. Identify if any feature can result in further information gain on the split sets
- 4. If yes, split further. If no, stop.
- 5. Goto 3
- 6. At each leaf, the prediction is the mode label

- Test:
- Follow the sequence of decisions based on the features of test example
- Make prediction according to leaf

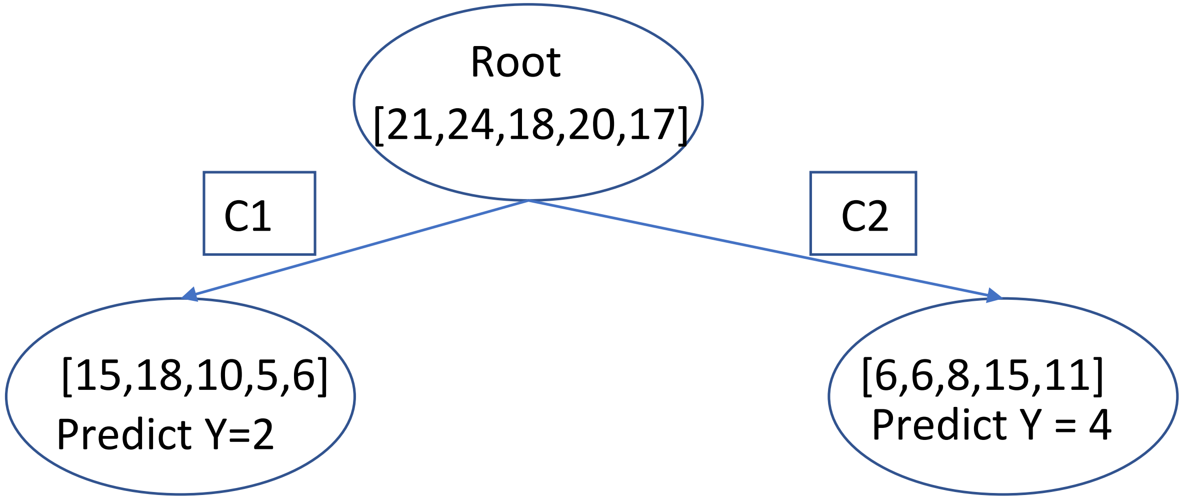
Decision Tree

TEST CASE: COMPANY = C2, PRICE = 350



- Prediction: Y= H

Multi-class Decision Stump



Training accuracy: 18/24 for CLASS 2, 15/20 for CLASS 4, 33/100 OVERALL

Advantages and Disadvantages

Advantage:

- Easy to interpret
- Easy to classify at test time
- Provides a ranking of features (according to usefulness)

Disadvantages:

- No optimal solution known, IG is just heuristic, can create many small branches
- Can cause overfitting if tree grows deep (need to stop growing)

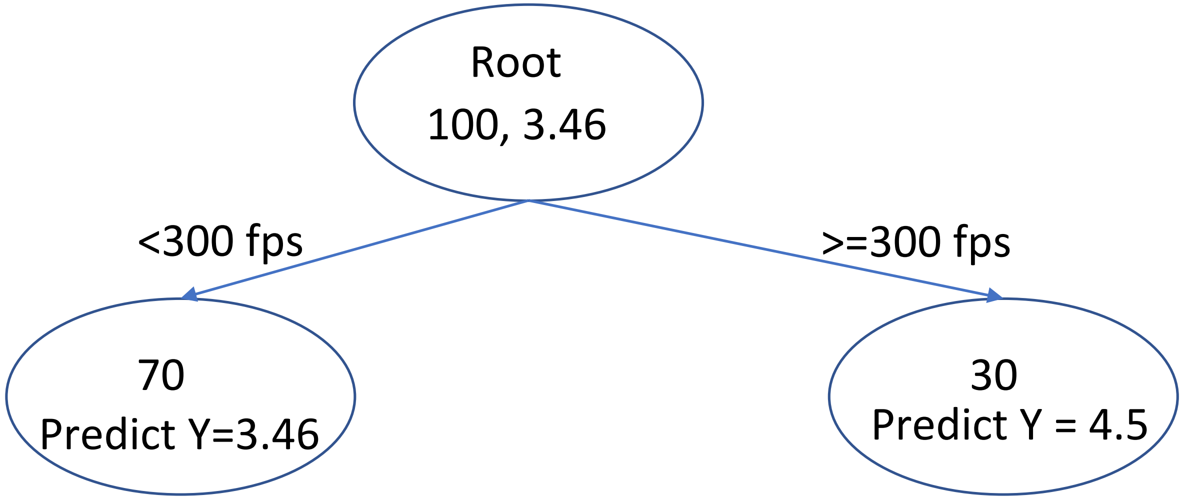
Regression Trees

- What if we want to predict Average Rating of a product?
- Real number between 1 and 5!
- Decision trees can also be used for regression
- Measure of homogeneity at each node: variance of labels (instead of entropy)
- Split criteria: reduction in total variance (instead of information gain)
- Final prediction: Mean label in the leaf node (instead of mode)

	COMPANY=C1	COMPANY=C2	NO SPLIT
COUNT	54	46	100
MEAN of RATINGS	3.0	4.0	3.46
VARIANCE of RATINGS	1.5	0.5	1.1
Reduction in Variance			1.1-(0.54*1.5+0.46*0.5)= 0.06

	VIDEO RATE <30 fps	Video RATE >=30fps	NO SPLIT
COUNT	70	30	100
MEAN of RATINGS	3.1	4.5	3.46
VARIANCE of RATINGS	1.2	0.4	1.1
Reduction in Variance			1.1-(0.7*1.2+0.3*0.4)= 0.14

Regression Tree



Root

100, 3.46

<300 fps

>=300 fps

70

Predict Y=3.46

30

Predict Y = 4.5