Decision Tree Classification

AI42001

31 July 2019

Data Representation

- Each data-point represented by D-dimensional feature vector X_i
- Animal classification: [#legs, #tail, colour, size, weight]
- Some of these features more useful for classification
- Sometimes, a single feature is enough to classify

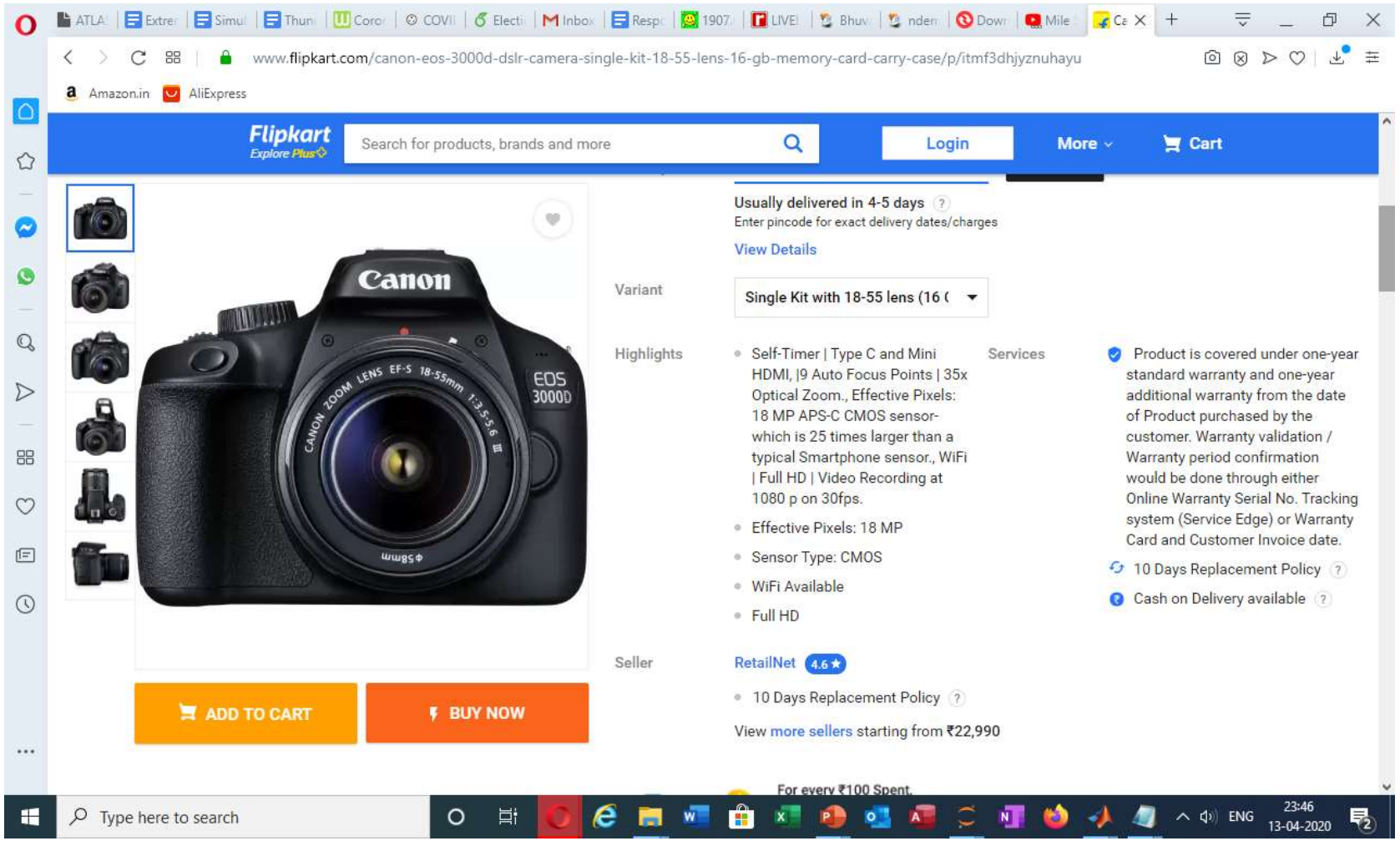
Feature Selection

- Cat vs Snake classification
- “#legs” feature is sufficient!

- Classifier function: #legs = 4: cat; #legs = 0: snake
- Decision function!

- For Cat vs Dog classification, “#legs” is certainly not sufficient
- It is not a “discriminative feature!”

Product Ratings based on Features



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?

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	????

Feature Selection

- The user has 5 exactly 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	$350	4

Feature Selection

- The user has 5 exactly 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	$350	4

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

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?

- Company ={C1, C2}, Price = real number, Y = {LOW (1-3), HIGH (4-5)}

	COMPANY=C1	COMPANY=C2	
#(Y=LOW)	43	20	63
#(Y=HIGH)	11	26	37
Total	54	46	100

What's a discriminative feature?

- Company ={C1, C2}, Price = real number, Y = {LOW (1-3), HIGH (4-5)}

	Price<300	**Price >=300**	
#(Y=LOW)	45	18	63
#(Y=HIGH)	25	12	37
Total	70	30	100

What's a discriminative feature?

- Company ={C1, C2}, Price = real number, Y = {LOW (1-3), HIGH (4-5)}

	Price<500	Price >=500	
#(Y=LOW)	55	8	63
#(Y=HIGH)	35	2	37
Total	90	10	100

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

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

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

What's a discriminative feature?

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

$$H = -\sum_i p_i (\log_2 p_i)$$

- Case 1: H = 1
- Case 2: H = 0.47

Feature selection based on entropy

- Before split: #(Y=cat) = 100, #(Y=dog) = 100. Entropy = 1.

	X1=YELLOW	X1=WHITE	
#(Y=CAT)	52	48	100
#(Y=DOG)	47	53	100
Total	99 (Entropy ~ 1)	101 (Entropy ~ 1)	200

- Information gain =

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

1 – (99/200*1 + 101/200*1) ~ 0!

Feature selection based on entropy

- Before split: #(Y=cat) = 100, #(Y=dog) = 100. Entropy = 1.

	X2 < 15	X2 > 15	
#(Y=CAT)	95	5	100
#(Y=DOG)	10	90	100
Total	105 (Entropy = 0.45)	95 (Entropy = 0.30)	200

- Information gain =

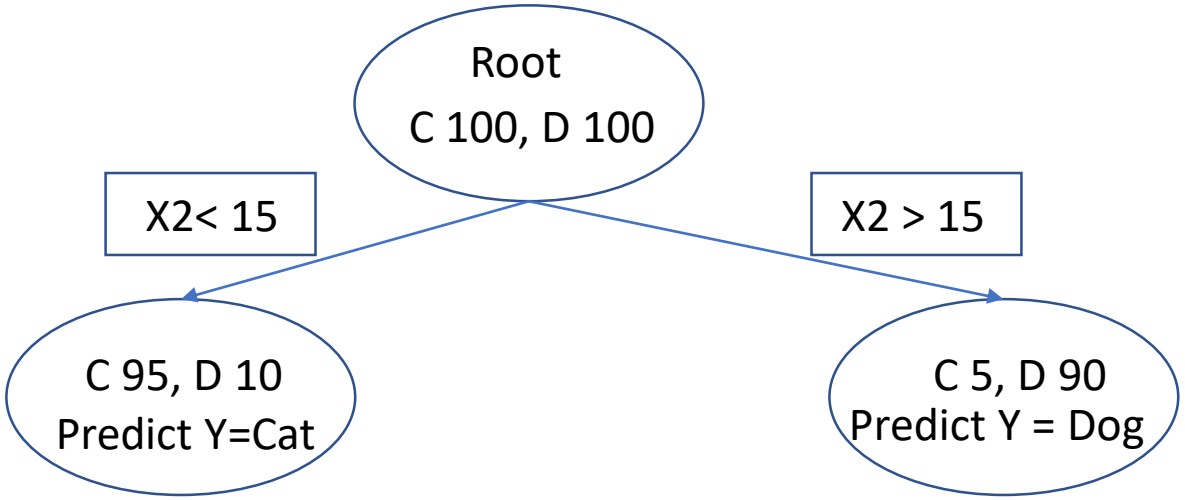
Original Entropy – (Split1_size*Split1_ entropy + Split2_size*Split2_ entropy)

1 – (105/200*0.45 + 95/200*0.3) = 0.62!!

Feature selection based on entropy

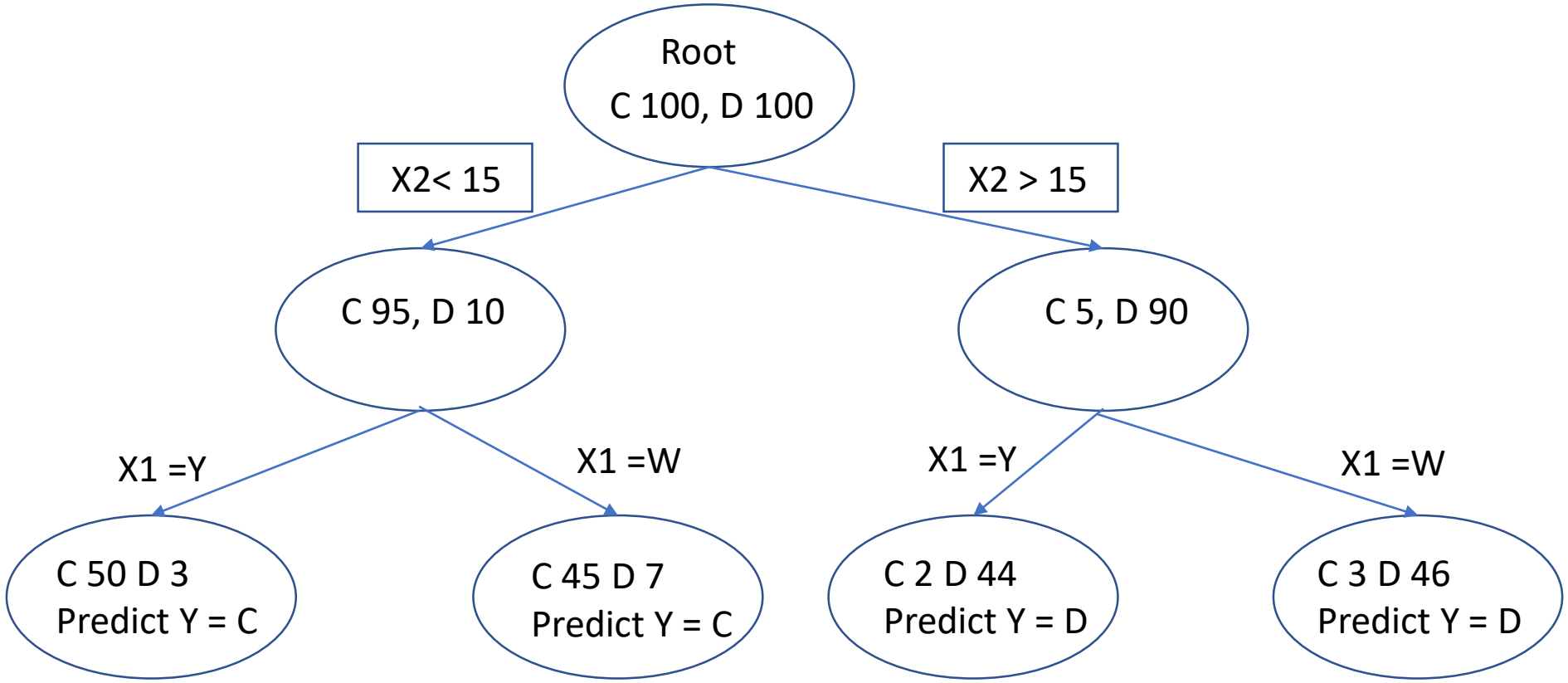
- Each discrete feature splits the dataset
- Continuous features can always be converted to discrete
- “Pure” dataset: - disbalanced class distribution
  - low entropy
  - high information gain
- Choose that feature which provides most information gain!

Decision Stump



- Training accuracy: 95/100 for cats, 90/100 for dogs

Decision Tree



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

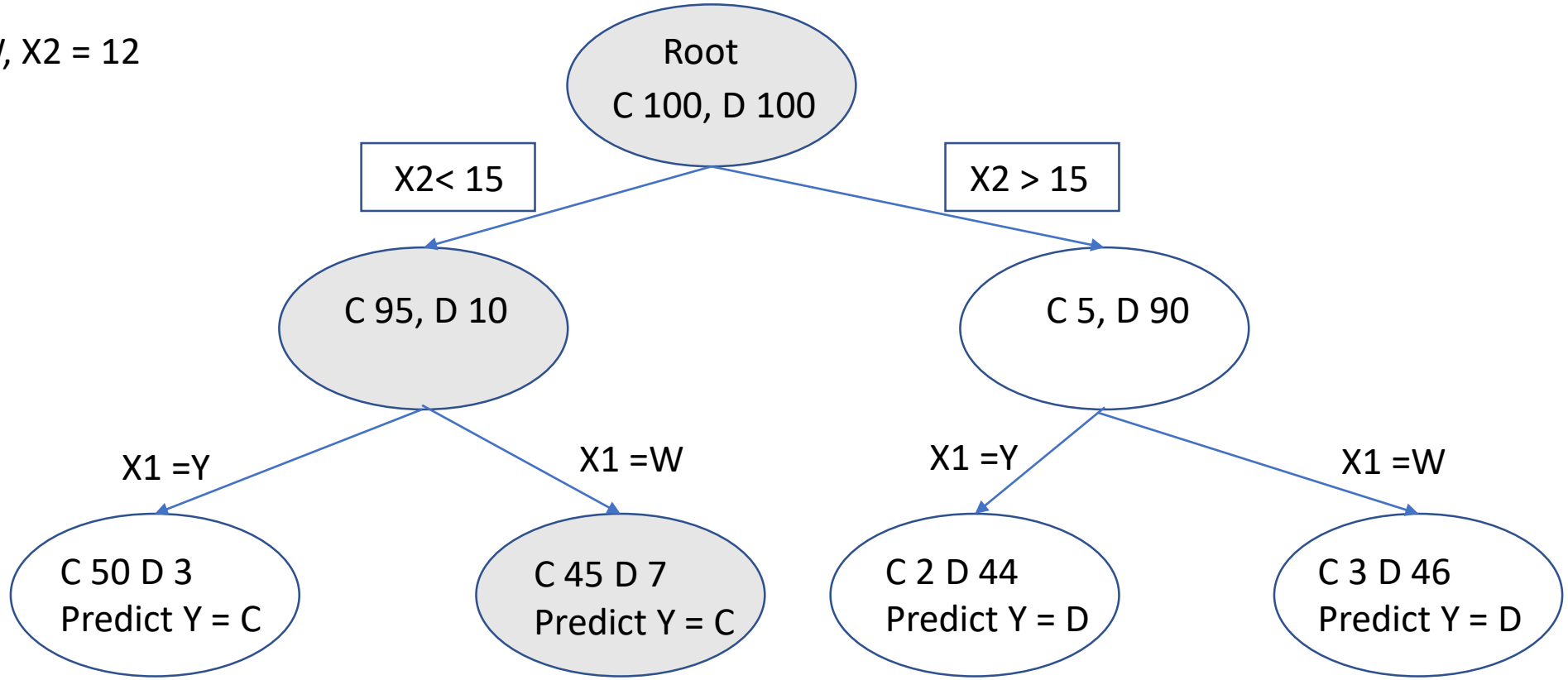
Decision Tree algorithm

- 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 for Testing

X1 = W, X2 = 12



- Prediction: Y= C

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

- 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