CSCC11 Assignment 2

Jefferson Li (jeffersonli.li@mail.utoronto.ca)

October 18, 2020

Question 1.

our goal is too maximize $P(f = 3)$

$$P(f = 3) = P(S + D = 3) \tag{1}$$

$$= P(S = s, D = 3 - s) = P(S = s)P(D = 3 - s) \quad \text{as they are independent} \tag{2}$$

$$P(S = s)P(D = 3 - s) = (\frac{1}{\sqrt{2\pi}3} exp(\frac{-1}{2}(\frac{s-0}{9})^2)(\frac{1}{\sqrt{2\pi}2} exp(\frac{-1}{2}(\frac{(3-s)-0}{4})^2) \quad \text{Replace with PDFs} \tag{3}$$

$$ln(P(f = 3)) = ln(\frac{1}{\sqrt{2\pi}3}) + \frac{-1}{2}(\frac{s-0}{9})^2 + ln(\frac{1}{\sqrt{2\pi}2}) + \frac{-1}{2}(\frac{(3-s)-0}{4})^2 \quad \text{take ln for easier differentiation} \tag{4}$$

$$\frac{\partial ln(P(f = 3))}{\partial s} = -1(\frac{s}{9}) - 1(\frac{3-s}{4}) * -1 \quad \text{differentiate the probability function} \tag{5}$$

$$\frac{\partial ln(P(f = 3))}{\partial s} = \frac{-s}{9} + \frac{3}{4} - \frac{s}{4} = \frac{-13s}{36} + \frac{3}{4} \tag{6}$$

$$\frac{\partial ln(P(f = 3))}{\partial s} = 0 \implies s = \frac{27}{13} \quad \text{find the most probable } s \tag{7}$$

$$\frac{\partial^2 ln(P(f = 3))}{\partial^2 s} = \frac{-13}{36} < 0 \quad 2_{nd} \text{ derivative check} \tag{8}$$

Since 2_{nd} derivative is negative, $s = \frac{27}{13}$ is a global maximum and therefore the most probable starting location

Question 2.

- a) Using Uni-variate splits, we need to consider 4 splits. As there are 4 features, and each feature has 2 unique values. So each feature has 1 split, and therefore 4 splits in total
- b)

$$H(D_0) = P_{true} log_2(\frac{1}{P_{true}}) + P_{False} log_2(\frac{1}{P_{False}}) = \frac{7}{16} log_2(\frac{16}{7}) + \frac{9}{16} log_2(\frac{16}{9}) \approx 0.98869$$

  - For the color split
    let t_0 be -1 if color is Red, and 1 if color is Green

$$IG(D_0, t_0) = H(D_0) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{9}$$
$$= H(D_0) - \frac{8}{16}(\frac{5}{8}log_2(\frac{8}{6}) + \frac{3}{8}log_2(\frac{8}{3})) - \frac{8}{16}(\frac{2}{8}log_2(\frac{8}{2}) + \frac{6}{8}log_2(\frac{8}{6})) \tag{10}$$
$$\approx 0.105833 \tag{11}$$

- For the size split
  let t_0 be -1 if size is Small, and 1 if size is Large

$$IG(D_0, t_0) = H(D_0) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{12}$$
$$= H(D_0) - \frac{8}{16}(\frac{5}{8}log_2(\frac{8}{6}) + \frac{3}{8}log_2(\frac{8}{3})) - \frac{8}{16}(\frac{2}{8}log_2(\frac{8}{2}) + \frac{6}{8}log_2(\frac{8}{6})) \tag{13}$$
$$\approx 0.105833 \tag{14}$$

- For the shape split
  let t_0 be -1 if shape is Sphere, and 1 if shape is Cube

$$IG(D_0, t_0) = H(D_0) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{15}$$
$$= H(D_0) - \frac{8}{16}(\frac{5}{8}log_2(\frac{8}{6}) + \frac{3}{8}log_2(\frac{8}{3})) - \frac{8}{16}(\frac{2}{8}log_2(\frac{8}{2}) + \frac{6}{8}log_2(\frac{8}{6})) \tag{16}$$
$$\approx 0.105833 \tag{17}$$

- For the texture split
  let t_0 be -1 if texture is Soft, and 1 if texture is Hard

$$IG(D_0, t_0) = H(D_0) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{18}$$
$$= H(D_0) - \frac{8}{16}(\frac{5}{8}log_2(\frac{8}{6}) + \frac{3}{8}log_2(\frac{8}{3})) - \frac{8}{16}(\frac{2}{8}log_2(\frac{8}{2}) + \frac{6}{8}log_2(\frac{8}{6})) \tag{19}$$
$$\approx 0.105833 \tag{20}$$

- c) NOTE : please use the part d) diagram to follow the node and level

Since all splits result in the same information gain, we can choose any one. Lets choose to split by color

- level 1 left node

$$H(D_1) = P_{true}log_2(\frac{1}{P_{true}}) + P_{False}log_2(\frac{1}{P_{False}}) = \frac{5}{8}log_2(\frac{8}{5}) + \frac{3}{8}log_2(\frac{8}{3}) \approx 0.954434$$

  * For the size split
    let t_1 be -1 if size is Small, and 1 if size is Large

$$IG(D_1, t_1) = H(D_1) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{21}$$
$$= H(D_1) - \frac{1}{2}(\frac{1}{1}log_2(\frac{1}{1}) + \frac{0}{0}log_2(0)) - \frac{1}{2}(\frac{1}{4}log_2(\frac{4}{1}) + \frac{3}{4}log_2(\frac{4}{3})) \tag{22}$$
$$\approx 0.68 \tag{23}$$

* For the shape split
  let t_1 be -1 if shape is Sphere, and 1 if shape is Cube

$$IG(D_1, t_1) = H(D_1) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{24}$$
$$= H(D_1) - \frac{1}{2}(\frac{3}{4}log_2(\frac{4}{3}) + \frac{1}{4}log_2(\frac{4}{1})) - \frac{1}{2}(\frac{2}{4}log_2(\frac{4}{2}) + \frac{2}{4}log_2(\frac{4}{2})) \tag{25}$$
$$\approx 0.36 \tag{26}$$

* For the texture split
  let t_1 be -1 if texture is Soft, and 1 if texture is Hard

$$IG(D_1, t_1) = H(D_1) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{27}$$
$$= H(D_1) - \frac{1}{2}(\frac{3}{4}log_2(\frac{4}{3}) + \frac{1}{4}log_2(\frac{4}{1})) - \frac{1}{2}(\frac{2}{4}log_2(\frac{4}{2}) + \frac{2}{4}log_2(\frac{4}{2})) \tag{28}$$
$$\approx 0.36 \tag{29}$$

Using the largest IG, the split function we will use for level 1 left node is splitting by size

– level 1 right node

$$H(D_2) = P_{true}log_2(\frac{1}{P_{true}}) + P_{False}log_2(\frac{1}{P_{False}}) = \frac{2}{8}log_2(\frac{8}{2}) + \frac{6}{8}log_2(\frac{8}{6}) \approx .81127$$

* For the size split
  let t_2 be -1 if size is Small, and 1 if size is Large

$$IG(D_2, t_2) = H(D_2) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{30}$$
$$= H(D_2) - \frac{1}{2}(\frac{1}{4}log_2(\frac{4}{1}) + \frac{3}{4}log_2(\frac{4}{3})) - \frac{1}{2}(\frac{1}{4}log_2(\frac{4}{1}) + \frac{3}{4}log_2(\frac{4}{3})) \tag{31}$$
$$\approx 0.311 \tag{32}$$

* For the shape split
  let t_2 be -1 if shape is Sphere, and 1 if shape is Cube

$$IG(D_2, t_2) = H(D_2) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{33}$$
$$= H(D_2) - \frac{1}{2}(\frac{2}{4}log_2(\frac{4}{2}) + \frac{2}{4}log_2(\frac{4}{2})) - \frac{1}{2}(0) \tag{34}$$
$$\approx 0.56127 \tag{35}$$

* For the texture split
  let t_2 be -1 if texture is Soft, and 1 if texture is Hard

$$IG(D_2, t_2) = H(D_2) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{36}$$
$$= H(D_2) - \frac{1}{2}(\frac{2}{4}log_2(\frac{4}{2}) + \frac{2}{4}log_2(\frac{4}{2})) - \frac{1}{2}(0) \tag{37}$$
$$\approx 0.56127 \tag{38}$$

Using the largest IG, the split function we will use for level 1 right node is splitting by shape.

- level 2 left most node
  - Since every data in this partition is True, we can terminate splitting and have a leaf node for $True$.
- level 2 left middle node

$$H(D_3) = P_{true}log_2(\frac{1}{P_{true}}) + P_{False}log_2(\frac{1}{P_{False}}) = \frac{1}{4}log_2(4) + \frac{3}{4}log_2(\frac{4}{3}) \approx 0.81127$$

  - For the shape split
    let t_3 be -1 if shape is Sphere, and 1 if shape is Cube

$$IG(D_3, t_3) = H(D_3) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{39}$$
$$= H(D_3) - \frac{1}{2}(\frac{1}{2}log_2(\frac{2}{1}) + \frac{1}{2}log_2(\frac{2}{1})) - \frac{1}{2}(0) \tag{40}$$
$$\approx 0.56127 \tag{41}$$

  - For the texture split
    let t_3 be -1 if texture is Soft, and 1 if texture is Hard

$$IG(D_3, t_3) = H(D_3) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{42}$$
$$= H(D_3) - \frac{1}{2}(\frac{1}{2}log_2(\frac{2}{1}) + \frac{1}{2}log_2(\frac{2}{1})) - \frac{1}{2}(0) \tag{43}$$
$$\approx 0.56127 \tag{44}$$

Using largest IG, we will split on shape

- level 2 right middle node

$$H(D_4) = P_{true}log_2(\frac{1}{P_{true}}) + P_{False}log_2(\frac{1}{P_{False}}) = \frac{1}{2}log_2(2) + \frac{1}{2}log_2(\frac{2}{1}) = 1$$

  - For the size split
    let t_4 be -1 if size is Small, and 1 if size is Large

$$IG(D_4, t_4) = H(D_4) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{45}$$
$$= H(D_4) - \frac{1}{2}(\frac{1}{2}log_2(\frac{2}{1}) + \frac{1}{2}log_2(\frac{2}{1})) - \frac{1}{2}(\frac{1}{2}log_2(\frac{2}{1}) + \frac{1}{2}log_2(\frac{2}{1})) \tag{46}$$
$$= 1 \tag{47}$$

  - For the texture split
    let t_4 be -1 if texture is Soft, and 1 if texture is Hard

$$IG(D_4, t_4) = H(D_4) - \frac{N_L}{N_j}H(D_L) - \frac{N_R}{N_j}H(D_R) \tag{48}$$
$$= H(D_4) - \frac{1}{2}(\frac{1}{1}log_2(\frac{1}{1}) + \frac{0}{0}log_2(\frac{0}{0})) - \frac{1}{2}(\frac{0}{0}log_2(\frac{0}{0}) + \frac{1}{1}log_2(\frac{1}{1})) \tag{49}$$
$$\approx 0.68627 \tag{50}$$

Using largest IG, we will split on texture

- level 2 right most node
  - Since every data in this partition is False, we can terminate splitting and have a leaf node for $False$.
- level 3 left most node

$$H(D_5) = P_{true} log_2(\frac{1}{P_{true}}) + P_{False} log_2(\frac{1}{P_{False}}) = \frac{1}{2} log_2(2) + \frac{1}{2} log_2(\frac{2}{1}) = 1$$

  - For the texture split
    let t_5 be -1 if texture is Soft, and 1 if texture is Hard

$$IG(D_5, t_5) = H(D_5) - \frac{N_L}{N_j} H(D_L) - \frac{N_R}{N_j} H(D_R) \tag{51}$$

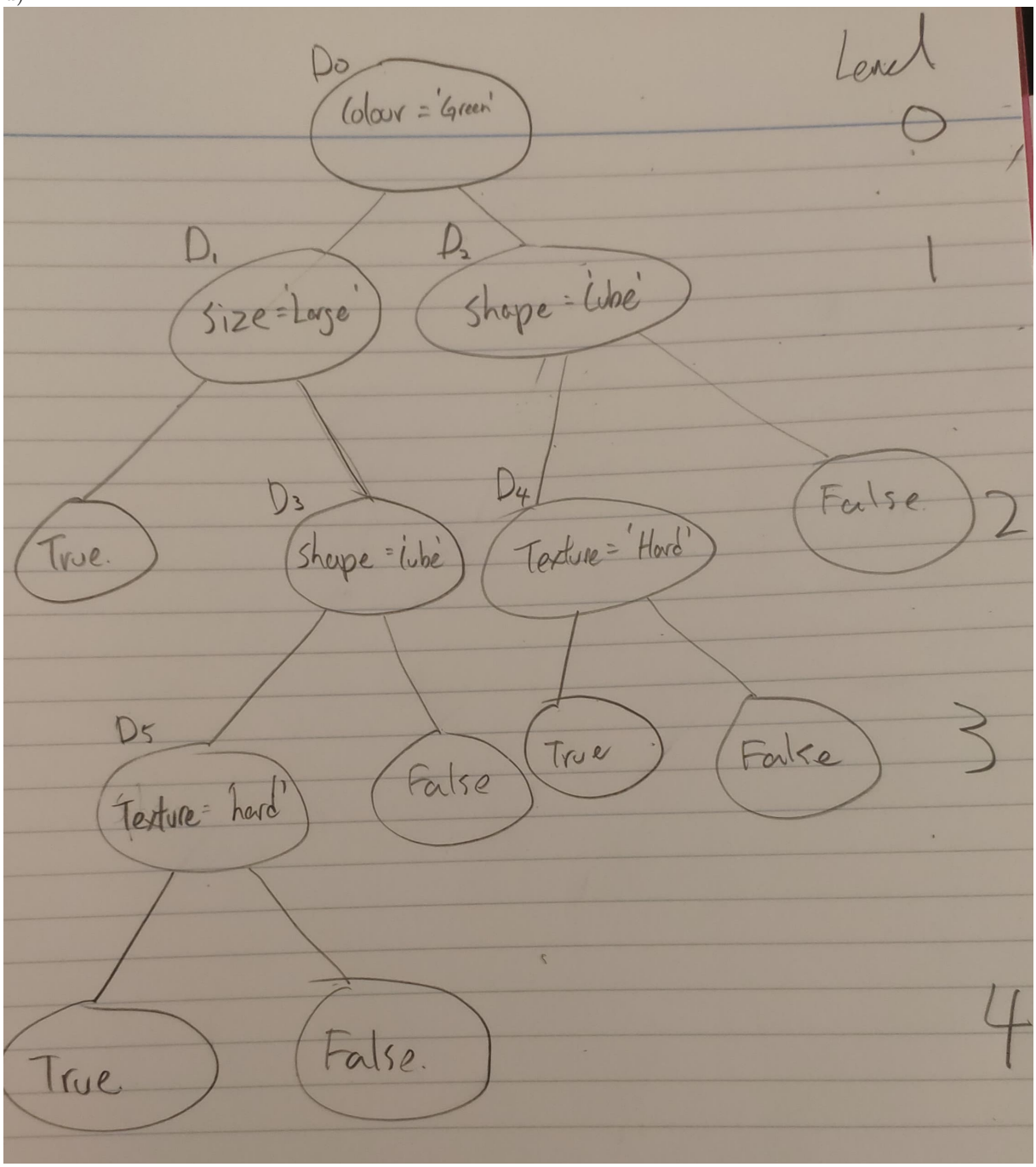
$$= H(D_5) - \frac{1}{2}(\frac{1}{1} log_2(\frac{1}{1}) + \frac{0}{0} log_2(\frac{0}{0})) - \frac{1}{2}(\frac{0}{0} log_2(\frac{0}{0}) + \frac{1}{1} log_2(\frac{1}{1})) \tag{52}$$

$$= 1 \tag{53}$$

Using largest IG, we will split on texture

- level 3 left middle node
  - Since every data in this partition is False, we can terminate splitting and have a leaf node for $False$.
- level 3 right middle node
  - Since every data in this partition is True, we can terminate splitting and have a leaf node for $True$.
- level 3 right most node
  - Since every data in this partition is False, we can terminate splitting and have a leaf node for $False$.
- level 4 left node
  - Since every data in this partition is True, we can terminate splitting and have a leaf node for $True$.
- level 4 right node
  - Since every data in this partition is False, we can terminate splitting and have a leaf node for $False$.

- d)


D0
Colour = 'Green'
Level
0
D1
Size = 'Large'
D2
Shape = 'Cube'
1
D3
Shape = 'Cube'
D4
Texture = 'Hard'
False
2
True.
D5
Texture = 'hard'
False
True
False
3
True
False.
4