Roll no.
2019111019

Date:
Page:

SMAI

# For parent node D:

No. of positive review = 5

No. " negative " = 5

$$\text{Parent entropy} = H(D) = -\sum_{i \in \text{class}} P_i \log(P_i)$$

$$= -2 \times 1/2 \times \log(1/2)$$

$$= 1$$

# If we divide along smell

→ For node "woody"

No. of positive = 2

" " negative = 3

$$P(D_i) = (2+3)/10 = 1/2$$

$$H(D_i) = \frac{-2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.97$$

→ For node "fruity"

No. of positive = 3

No. of negative = 2

$$P(D_i) = (3+2)/10 = 1/2$$

$$H(D_i) = \frac{-3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.97$$

$$\text{Info gain} = 1 - \frac{1}{2} \times 0.97 - \frac{1}{2} \times 0.97 = 0.03$$

Date:
Page:

# If we divide along taste

→ For node 'sweet'

No. of positive = 0

No. of negative = 3

$P(Di) = 3/10$

$H(Di) = \frac{-3}{3} \log\left(\frac{3}{3}\right) = 0$

→ For node 'sour'

No. of positive = 3

No. of negative = 0

$P(Di) = 3/10$

$H(Di) = \frac{-3}{3} \log\left(\frac{3}{3}\right) = 0$

→ For node 'salty'

No. of positive = 2

" " negative = 2

$P(Di) = 4/10$

$H(Di) = 2 \times \frac{-2}{4} \log\left(\frac{2}{4}\right) = 1.$

Info gain $= 1 - \frac{4}{10} = 0.6$

Date:
Page:

# If we divide along portion

- For node 'small'

No. of negative = 1

No. of positive = 4

$$P(Di) = (1+4)/10 = 1/2$$

$$H(Di) = -\frac{1}{5}\log(1/5) - \frac{4}{5}\log(4/5) = 0.72$$

- For node 'large'

No. of negative = 4

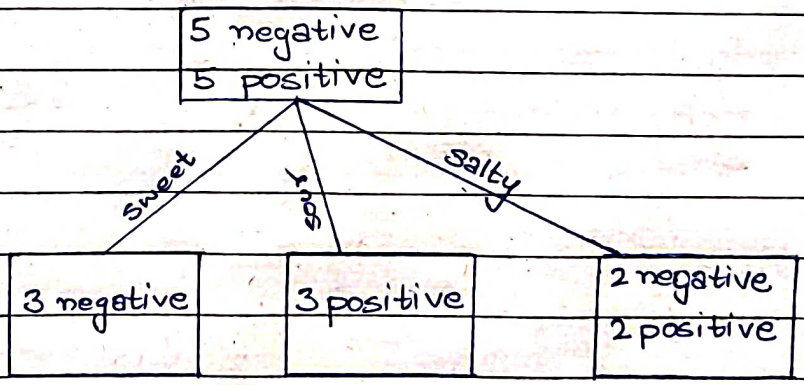
No. of positive = 1

$$P(Di) = (4+1)/10 = 1/2$$

$$H(Di) = -\frac{4}{5}\log(4/5) - \frac{1}{5}\log(1/5) = 0.72$$

$$\text{Info gain} = 1 - 2\times\frac{1}{2}\times 0.72 = 0.28$$

As info gain is maximum for taste, we will split along taste.



Now if the taste is salty, the table will be -

| Review | Smell | Taste | Portion |
|---|---|---|---|
| Negative | Fruity | Salty | Large |
| Negative | Fruity | Salty | Large |
| Positive | Fruity | Salty | Small |
| Positive | Fruity | Salty | Small |

From the above table it is clear that portion is the only feature with positive info gain and we can directly classify after splitting along portion.

So the decision tree obtained is -