Food review using decision trees - SMAI - Avneesh Mishra

2021701032

Review	Smell	Taste	Portion
Negative	Woody N	Sweet N	Small N
Negative	Fruity N	Salty N	Large N
Negative	Fruity N	Salty N	Large N
Positive	Fruity P	Sour P	Small P
Positive	Woody P	Sour P	Small P
Negative	Woody N	Sweet N	Large N
Positive	Woody P	Sour P	Large P
Positive	Fruity P	Salty P	Small P
Positive	Fruity P	Salty P	Small P
Negative	Woody N	Sweet N	Large N

Impurity measure: Entropy : $H(Q_m) = -\sum_k P_{mk}\log(P_{mk})$ $k \to$ {Review class: positive = 1, negative = 0}

Node 0 (entropy in the data) → $N_0 = 10$ datapoints (5 positive, 5 negative) P_{mk} = proportion of class k in node m

$P_{0P} = \frac{1}{10}\times 5 = 0.5$; $P_{0N} = \frac{5}{10} = 0.5$ → Entropy = $-(0.5\log(0.5)+0.5\log(0.5)) = 0.6931$ (data in the raw form) → θ_1

A: Entropy of each rule in first stage

→ Potential Rules : 5 possible decisions for split

Smell in {Woody, fruity} → 1 $\Delta i = 0.0201$

Taste in {Sweet, Salty, Sour} → 3 $\Delta i_1 = 0.2784$ ✓ $\Delta i_2 = 0$ $\Delta i_3 = 0.2784$

Portion in {Small, Large} → 1 $\Delta i = 0.1927$

Potential Rule 1 : $\theta_1^{left} = (Smell == Woody)$, $\theta_1^{right} = (Smell == fruity)$ $N^{left} = 5$; $N^{right} = 5$ $P_{LN} = \frac{3}{5}$ $P_{LP} = \frac{2}{5}$ $P_{RN} = \frac{2}{5}$ $P_{RP} = \frac{3}{5}$

$H(\theta_1^{left}) = -\left[\frac{3}{5}\log\left(\frac{3}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)\right] = 0.6730$ $H(\theta_1^{right}) = -\left(\frac{2}{5}\log\left(\frac{2}{5}\right)+\frac{3}{5}\log\left(\frac{3}{5}\right)\right) = 0.6730$ $G(\theta_1) = \frac{5}{10}\times 0.6730 + \frac{5}{10}\times 0.6730 = 0.6730$ $\Delta i = 0.6931 - 0.6730 = 0.0201$

Potential Rule 2 : $\theta_1^{left} = (Taste == Sweet)$, $\theta_1^{right} = (Taste == Salty \text{ or } Sour)$ $N^{left} = 3$ (3N+0P) ; $N^{right} = 7$ (2N+5P) $P_{LN} = \frac{3}{3} = 1$ $P_{LP} = \frac{0}{3} = 0$ $P_{RN} = \frac{2}{7}$ $P_{RP} = \frac{5}{7}$

$H(\theta_1^{left}) = -\left[\frac{3}{3}\ln\left(\frac{3}{3}\right)+0\times\ln(0)\right] = 0$ $H(\theta_1^{right}) = -\left(\frac{2}{7}\ln\left(\frac{2}{7}\right)+\frac{5}{7}\ln\left(\frac{5}{7}\right)\right) = 0.5982$ $G(\theta_1) = \frac{3}{10}\times 0 + \frac{7}{10}\times 0.5982 = 0.4187$ $\Delta i = 0.6931 - 0.4187 = 0.2744$

Potential Rule 3 : $\theta_1^{left} = (Taste == Salty)$, $\theta_1^{right} = (Taste == Sweet \text{ or } Sour)$ $N^{left} = 4$ (2N+2P); $N^{right} = 6$ (3P+3N) $P_{LN} = \frac{2}{4} = 0.5$ $P_{LP} = \frac{2}{4} = 0.5$ $P_{RN} = \frac{3}{6} = 0.5$ $P_{RP} = \frac{3}{6} = 0.5$

$H(\theta_1^{left}) = -\left[0.5\ln(0.5)+0.5\ln(0.5)\right] = 0.6931$ $H(\theta_1^{right}) = -\left[0.5\ln(0.5)+0.5\ln(0.5)\right] = 0.6931$ $G(\theta_1) = \frac{4}{10}\times 0.6931 + \frac{6}{10}\times 0.6931 = 0.6931$ $\Delta i = 0.6931 - 0.6931 = 0$

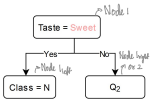
Potential Rule 4 : $\theta_1^{left} = (Taste == Sour)$, $\theta_1^{right} = (Taste == Sweet \text{ or } Salty)$ $N^{left} = 3$ (0N+3P); $N^{right} = 7$ (5N+2P) $P_{LN} = \frac{0}{3} = 0$ $P_{LP} = \frac{3}{3} = 1$ $P_{RN} = \frac{5}{7}$ $P_{RP} = \frac{2}{7}$

$H(\theta_1^{left}) = -\left[0\ln(0)+1\ln(1)\right] = 0$ $H(\theta_1^{right}) = -\left[\frac{5}{7}\ln\left(\frac{5}{7}\right)+\frac{2}{7}\ln\left(\frac{2}{7}\right)\right] = 0.5982$ $G(\theta_1) = \frac{3}{10}\times 0 + \frac{7}{10}\times 0.5982 = 0.4187$ $\Delta i = 0.6931 - 0.4187 = 0.2744$

Potential Rule 5 : $\theta_1^{left} = (Portion == Small)$, $\theta_1^{right} = (Portion == Large)$ $N^{left} = 5$ (1N+4P); $N^{right} = 5$ (4N+1P) $P_{LN} = \frac{1}{5}$ $P_{LP} = \frac{4}{5}$ $P_{RN} = \frac{4}{5}$ $P_{RP} = \frac{1}{5}$

$H(\theta_1^{left}) = -\left[\frac{1}{5}\ln\left(\frac{1}{5}\right)+\frac{4}{5}\ln\left(\frac{4}{5}\right)\right] = 0.5004$ $H(\theta_1^{right}) = -\left[\frac{4}{5}\ln\left(\frac{4}{5}\right)+\frac{1}{5}\ln\left(\frac{1}{5}\right)\right] = 0.5004$ $G(\theta_1) = \frac{5}{10}\times 0.5004 + \frac{5}{10}\times 0.5004 = 0.5004$ $\Delta i = 0.6931 - 0.5004 = 0.1927$

Potential Rule 2 can be chosen for best split (highest gain in information)



Node 2 : All nodes with Taste = Salty or Sour $N_2 = 7$ (2N+5P) $H(Q_2) = -\left(\frac{2}{7}\ln\left(\frac{2}{7}\right)+\frac{5}{7}\ln\left(\frac{5}{7}\right)\right) = 0.5982$

3 Possible rules:
- Smell in {Woody, fruity} $\Delta i = 0.1175$
- Taste in {Salty, Sour} $\Delta i = 0.2022$
- Portion in {Small, Large} $\Delta i = 0.3255$ ✓

Potential Rule 1 : $\theta_2^{left} = (Smell == Woody)$, $\theta_2^{right} = (Smell == fruity)$ $N^{left} = 2$ (2P+0N) $N^{right} = 5$ (2N+3P)

$H(\theta_2^{left}) = -(1\ln(1)+0\ln(0)) = 0$ $H(\theta_2^{right}) = -\left(\frac{2}{5}\ln\left(\frac{2}{5}\right)+\frac{3}{5}\ln\left(\frac{3}{5}\right)\right) = 0.6730$ $G(\theta_2) = \frac{2}{7}\times 0 + \frac{5}{7}\times 0.6730 = 0.4807$ $\Delta i = 0.5982 - 0.4807 = 0.1175$

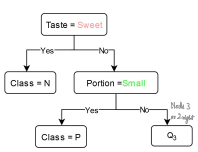
Potential Rule 2 : $\theta_2^{left} = (Taste == Salty)$, $\theta_2^{right} = (Taste == Sour)$ $N^{left} = 4$ (2N+2P) $N^{right} = 3$ (0N+3P)

$H(\theta_2^{left}) = -\left(\frac{2}{4}\ln\left(\frac{2}{4}\right)+\frac{2}{4}\ln\left(\frac{2}{4}\right)\right) = 0.6931$ $H(\theta_2^{right}) = -\left[0\ln(0)+1\ln(1)\right] = 0$ $G(\theta_2) = \frac{4}{7}\times 0.6931 + \frac{3}{7}\times 0 = 0.3960$ $\Delta i = 0.5982 - 0.3960 = 0.2022$

Potential Rule 3 : $\theta_2^{left} = (Portion == Small)$, $\theta_2^{right} = (Portion == Large)$ $N^{left} = 4$ (0N+4P) $N^{right} = 3$ (2N+1P)

$H(\theta_2^{left}) = -\left[0\ln(0)+1\ln(1)\right] = 0$ $H(\theta_2^{right}) = -\left[\frac{2}{3}\ln\left(\frac{2}{3}\right)+\frac{1}{3}\ln\left(\frac{1}{3}\right)\right] = 0.6365$ $G(\theta_2) = \frac{4}{7}\times 0 + \frac{3}{7}\times 0.6365 = 0.2727$ $\Delta i = 0.5982 - 0.2727 = 0.3255$

Potential Rule 3 can be chosen for best split (Highest Δi)



Node 3 : All nodes with (Taste = Salty or Sour) and (Portion = Large) $N_3 = 3$ (2N+1P) $H(Q_3) = -\left[\frac{2}{3}\ln\left(\frac{2}{3}\right)+\frac{1}{3}\ln\left(\frac{1}{3}\right)\right] = 0.6365$

Potential Rule 1 : $\theta_3^{left} = (Taste == Salty)$, $\theta_3^{right} = (Taste == Sour)$ $N^{left} = 2$ (2N+0P) $N^{right} = 1$ (0N+1P) $H(\theta_3^{left}) = -[1\ln(1)+0\ln(0)] = 0$; $H(\theta_3^{right}) = -[0\ln(0)+1\ln(1)] = 0$; $G(\theta_3) = 0$; $\Delta i = 0.6365 - 0 = 0.6365$

Potential Rule 2: $\theta_3^{left} = (Smell == Woody)$, $\theta_3^{right} = (Smell == fruity)$ $N^{left} = 1$ (0N+1P) $N^{right} = 2$ (2N+0P) $H(\theta_3^{left}) = -[0\ln(0)+1\ln(1)] = 0$; $H(\theta_3^{right}) = -[1\ln(1)+0\ln(0)] = 0$; $G(\theta_3) = 0$; $\Delta i = 0.6365 - 0 = 0.6365$

Potential Rule 2 (or Small) can be used as the splitting criteria (as criteria of taste has already been used, we'd like to factor in more fields)