

Ques: 01

The formula for entropy is:

$$H(S) = -P_1 \log_2 P_1 - P_2 \log_2 P_2$$

where  $P_1$  &  $P_2$  are the proportion of each

Low Risk: 4 instances (D: 1, 3, 5, 7)

High Risk: 4 instances (D: 2, 4, 6, 8)

$$H(S) = -\left(\frac{4}{8} \log_2 \frac{4}{8} + \frac{4}{8} \log_2 \frac{4}{8}\right)$$

$$H(S) = 1.0$$

Credit Score  $> 650$

Low Risk: 4

High Risk: 0

$$H(S_1) = -\left(\frac{4}{4} \log_2 \frac{4}{4} + \frac{0}{4} \log_2 \frac{0}{4}\right) = 0$$

Credit Score  $\leq 650$

Low Risk: 0

High Risk: 4

$$H(S_2) = -\left(\frac{0}{4} \log_2 \frac{0}{4} + \frac{4}{4} \log_2 \frac{4}{4}\right) = 1.0$$

Gain:  $I_G = 1.0 - 0 = 1.0$

Ques: 2

Lg Age  $\leq 35$  ID: {1, 2, 4}

Rg Age  $> 35$  ID: {3, 5, 6, 7, 8}

$$\text{Variance} = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\bar{x} = \frac{720 + 650 + 750 + 600 + 780 + 630 + 710 + 640}{8} = \frac{5480}{8} = 685$$

$$\sigma^2 = \frac{(720-685)^2 + (650-685)^2 + (750-685)^2 + (600-685)^2 + (780-685)^2 + (630-685)^2 + (710-685)^2 + (640-685)^2}{8}$$

$$\sigma^2 = \frac{28600}{8} = 3575$$

Compute the variance

Lg Age  $\leq 35$

Subset: {1, 2, 4}

Credit scores: {720, 630, 600}

$$\bar{x}_L = \frac{720 + 650 + 600}{3} = \frac{1970}{3} = 656.67$$

$$\sigma_L^2 = \frac{(720-656.67)^2 + (650-656.67)^2 + (600-656.67)^2}{3}$$

$$= \frac{4000.89 + 44.49 + 3211.69}{3} = \frac{7252.07}{3} = 2117.02$$

Compute Variance.

Eg Age > 35

Subset {3, 5, 6, 7, 8}

Credit Scores {750, 780, 630, 710, 640}

$$\bar{x}_P = \frac{750 + 780 + 630 + 710 + 640}{5} = 702$$

Variance

$$\sigma_P^2 = \frac{(750 - 702)^2 + (780 - 702)^2 + (630 - 702)^2 + (710 - 702)^2 + (640 - 702)^2}{5}$$
$$= \frac{(48)^2 + (78)^2 + (-72)^2 + (8)^2 + (-62)^2}{5}$$
$$= 3496$$

Weighted Variance:

$$\text{Weighted Variance} = 0.375 \times 2419.02 + 0.625 \times 3496$$
$$= 0.375 \times 2419.02 + 0.625 \times 3496$$
$$= 3092.13$$

Compute Variance Reduction:

$$\text{Variance Reduction} = 3575 - 3092.13$$
$$= 482.87$$

Ques: 03

Age	CreditScore	Education	RiskLevel
28	650	14	High
31	600	12	High
29	630	14	High
33	640	12	High

Age Range: 28 - 33

CreditScore: Between 600 - 650

Education: Mostly 12 or 14

Age	CreditScore	Education (missing)
30	645	missing

T<sub>2</sub> has a high probability of being High Risk

Ques: 04

using linear regression

$$\hat{y} = \theta_0 + \theta_1 x$$

where

- $\hat{y}$  is the predicted CreditScore
- $x$  is the input feature (Age)
- $\theta_0 = 500$  (initial intercept)
- $\theta_1 = 5$  (initial slope)

Predicted,  $\hat{y}_i = \theta_0 + \theta_1 \times \text{Age}_i$

Cost of linear regression using MSE

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y_i)^2 \quad \left[ \begin{array}{l} m = 8 \text{ (number of training examples)} \\ y_i \text{ is actual credit score} \end{array} \right]$$

ID	Age	Y	$h_{\theta}(x)$
1	35	720	$500 + 5(35) = 675$
2	28	650	$500 + 5(28) = 640$
3	45	750	$500 + 5(45) = 725$
4	31	600	$500 + 5(31) = 655$
5	52	780	$500 + 5(52) = 760$
6	29	630	$500 + 5(29) = 645$
7	42	710	$500 + 5(42) = 710$
8	33	640	$500 + 5(33) = 665$