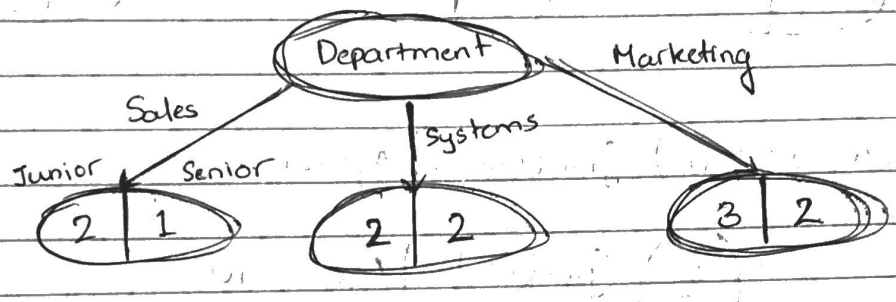
Machine Learning HW2, Anshuman Dikhit

① a Construct Decision Tree w/ multiway split

class: Status {Junior, Senior}

n=12



$$Gini(Dept) = \left(\frac{3}{12}\right) \cdot Gini(Status \mid Dept = Sales) + \left(\frac{4}{12}\right) \cdot Gini(Status \mid Dept = Systems) + \left(\frac{5}{12}\right) \cdot Gini(Status \mid Dept = Marketing)$$

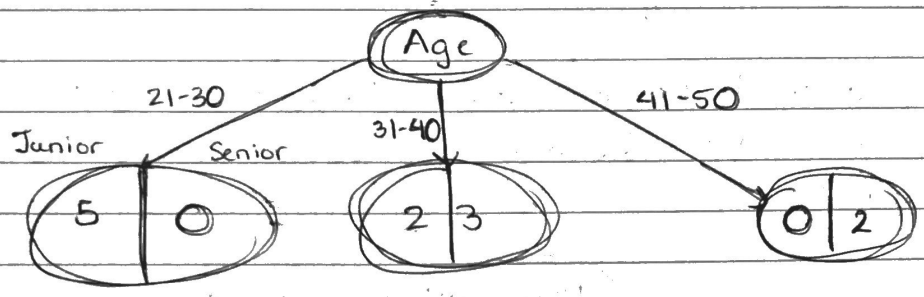
$$Gini(Status \mid Dept = Sales) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.44$$

$$Gini(Status \mid Dept = Systems) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini(Status \mid Dept = Marketing) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$GINI(Dept) = \left(\frac{3}{12}\right) 0.44 + \left(\frac{4}{12}\right) 0.5 + \left(\frac{5}{12}\right) 0.48 = 0.47667$$

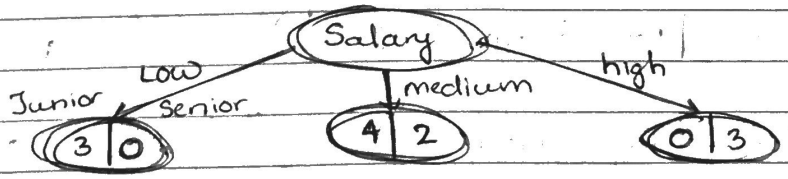
0.47667



$$Gini(Age) = \left(\frac{5}{12}\right) \cdot Gini(Status \mid Age = 21\text{-}30) + \left(\frac{5}{12}\right) \cdot Gini(Status \mid Age = 31\text{-}40) + \left(\frac{2}{12}\right) \cdot Gini(Status \mid Age = 41\text{-}50)$$

$$Gini(Status \mid Age = 21\text{-}30) = 0$$
$$Gini(Status \mid Age = 31\text{-}40) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$
$$Gini(Status \mid Age = 41\text{-}50) = 0$$

$$Gini(Age) = \frac{5}{12}(0.48) = 0.2$$

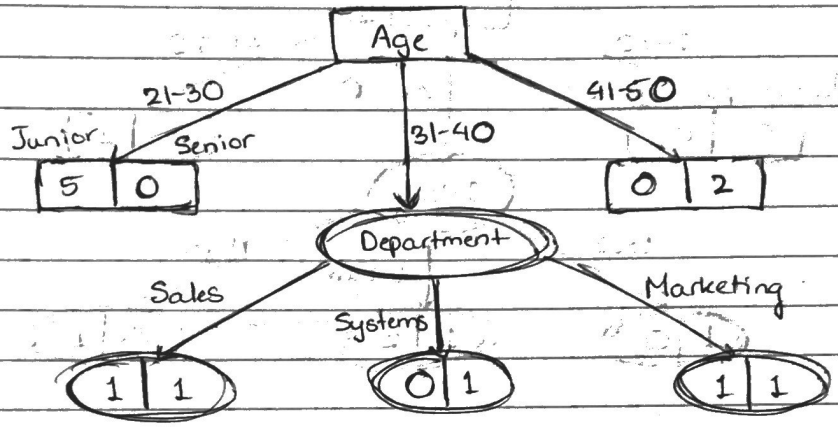


$$Gini(Salary) = \left(\frac{3}{12}\right) \cdot Gini(Status \mid Low) + \left(\frac{1}{2}\right) \cdot Gini(Status \mid medium) + \left(\frac{3}{12}\right) \cdot Gini(Status \mid high)$$

$$Gini(Status \mid medium) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.44$$

$$Gini(Salary) = \frac{1}{2}(0.44) = 0.22$$

Root Node: Age

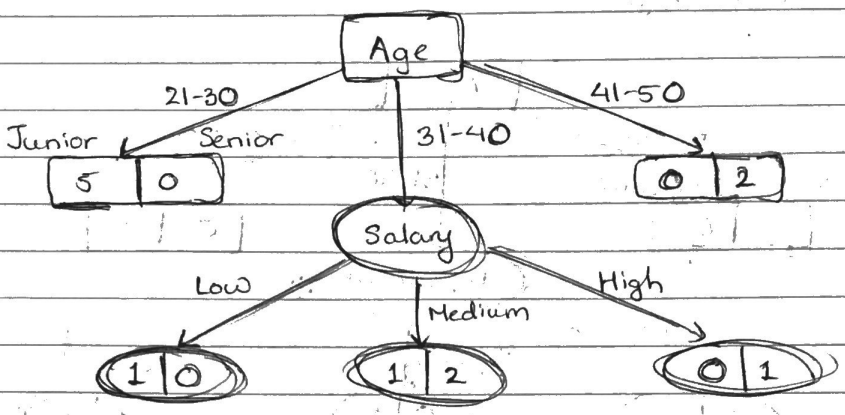


$$\text{Gini}(\text{Department}) = \left(\frac{2}{5}\right) \cdot \text{Gini}(\text{Status} \mid \text{Dept} = \text{Sales}) + \left(\frac{1}{5}\right) \cdot \text{Gini}(\text{Status} \mid \text{Dept} = \text{Systems}) + \left(\frac{2}{5}\right) \cdot \text{Gini}(\text{Status} \mid \text{Dept} = \text{Marketing})$$

$$\text{Gini}(\text{Status} \mid \text{Dept} = \text{Sales}) = 1 - (0.5)^2 - (0.5)^2 = 0.5$$

$$\text{Gini}(\text{Status} \mid \text{Dept} = \text{Marketing}) = " = 0.5$$

$$\text{Gini}(\text{Department}) = \left(\frac{2}{5}\right)\left(\frac{1}{2}\right) + 0 + \left(\frac{2}{5}\right)\left(\frac{1}{2}\right) = 0.4$$

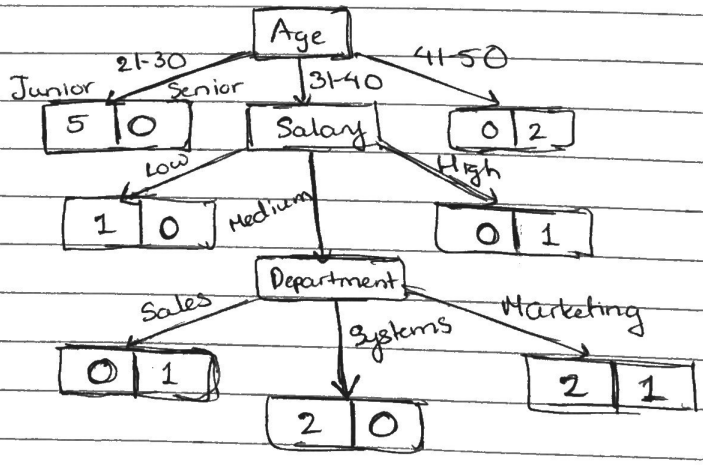


$$Gini(Salary) = \left(\frac{1}{5}\right) \cdot Gini(Status \mid Salary = Low) + \left(\frac{3}{5}\right) \cdot Gini(Salary \mid Salary = Medium) + \left(\frac{1}{5}\right) \cdot Gini(Salary \mid Salary = High)$$

$$Gini(Salary \mid Salary = Medium) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44$$

$$Gini(Salary) = \frac{3}{5} \cdot (0.44) = \underline{0.264}$$

Final Decision Tree



$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{T.P. + F.P.}}$$

b

		Predicted Class	
	Class	Junior	Senior
Actual	Junior	7	0
	Senior	1	4

recall = $\frac{4}{5}$

precision = $\frac{4}{4}$

$$\text{F-measure} = \frac{2 \cdot \left(\frac{4}{5} \cdot \frac{4}{4}\right)}{\left(\frac{4}{5} + \frac{4}{4}\right)} = \frac{\frac{8}{5}}{\frac{16+20}{20}} = \frac{2}{\cancel{8}} \cdot \frac{\cancel{20}}{\cancel{36}_{9}} = \boxed{\frac{8}{9}}$$

c

Sales	21-30	High	Junior
Systems	21-30	Medium	Junior
Marketing	41-50	High	Senior
Marketing	31-40	Low	Junior

② Naive Bayes Classification

a We have to assume independence between all attributes X_i where a label is given:

$$P(X_1 X_2 \dots X_d \mid Y_j) = P(X_1 \mid Y_j) \cdot P(X_2 \mid Y_j) \dots P(X_d \mid Y_j)$$

If this stands, then we can estimate $P(X_i \mid Y_j)$, $\forall X_i, Y_j$ combinators for all training data.

b $P(A=0 \mid +) = \frac{2}{5}$ $\quad P(A=0 \mid -) = \frac{3}{5}$

$P(A=1 \mid +) = \frac{3}{5}$ $\quad P(A=1 \mid -) = \frac{2}{5}$

$P(B=0 \mid +) = \frac{4}{5}$ $\quad P(B=0 \mid -) = \frac{3}{5}$

$P(B=1 \mid +) = \frac{1}{5}$ $\quad P(B=1 \mid -) = \frac{2}{5}$

$P(C=0 \mid +) = \frac{1}{5}$ $\quad P(C=0 \mid -) = \frac{0}{5}$

$P(C=1 \mid +) = \frac{4}{5}$ $\quad P(C=1 \mid -) = \frac{5}{5}$

c

$$P(+ \mid K) = \frac{\left(\frac{2}{5}\right)\left(\frac{1}{5}\right)\left(\frac{1}{5}\right)\left(\frac{1}{2}\right)}{K} = \frac{0.008}{K}$$

$$P(- \mid K) = \frac{0}{K} = 0 \Rightarrow$$

+ is the class label

d $P(A=0 \mid +) = \frac{4}{9}$ $P(A=0 \mid -) = \frac{5}{9}$

$P(A=1 \mid +) = \frac{5}{9}$ $P(A=1 \mid -) = \frac{4}{9}$

$P(B=0 \mid +) = \frac{6}{9}$ $P(B=0 \mid -) = \frac{5}{9}$

$P(B=1 \mid +) = \frac{3}{9}$ $P(B=1 \mid -) = \frac{4}{9}$

$P(C=0 \mid +) = \frac{3}{9}$ $P(C=0 \mid -) = \frac{2}{9}$

$P(C=1 \mid +) = \frac{6}{9}$ $P(C=1 \mid -) = \frac{7}{9}$

e

$$P(+ \mid K) = \frac{(4/9)(6/9)(6/9)(1/2)}{K} = \frac{0.247}{K}$$

$$P(- \mid K) = \frac{(5/9)(4/9)(3/9)(1/2)}{K} = \frac{0.274}{K} \Rightarrow$$

- is class label

f The second method, if one of the conditional probabilities is 0, then the calculation of the class probability skews to 0. Via m-estimates we can ensure that doesn't happen