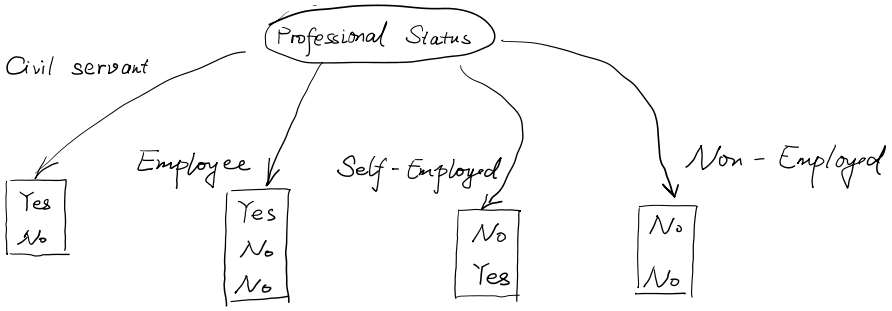
Exercise 08

Wednesday, December 20, 2023 2:16 PM

1(a) Unclassified entropy:

entropy([3, 6])

$= \frac{1}{3}(\log_2(\frac{1}{3}) + \frac{1}{6}\log_2(\frac{1}{6}) = 0.959$ bits



Entropies: $[1,1]_2$, $[1,2]_3$, $[1,1]_2$, $[2,0]_2$

entropy([1,1]) $= -\frac{1}{2} \cdot \log(\frac{1}{2}) \cdot 2$

$= 1$ bits

entropy([1,2]) $= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$

$= 0.918$ bits

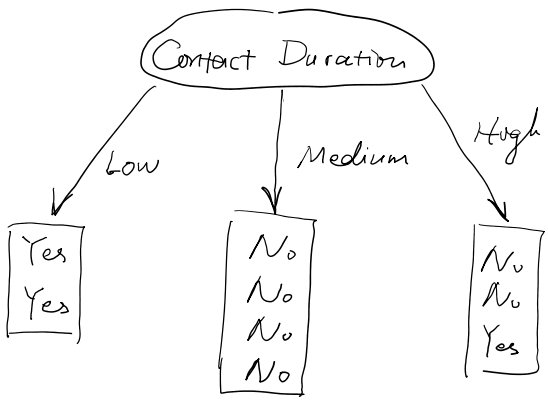
entropy([1,1]) = 1 bits

entropy([2,0]) = 0 bits

Average: $\frac{2}{9} \cdot 1 + \frac{3}{9} \cdot 0.918 + \frac{2}{9} \cdot 1 + \frac{2}{9} \cdot 0$

$= 0.750$ bits

Information gain: $0.959 - 0.750 = 0.209$ bits



Entropies: [2,0] [0,4], [1,2]
2 4 3

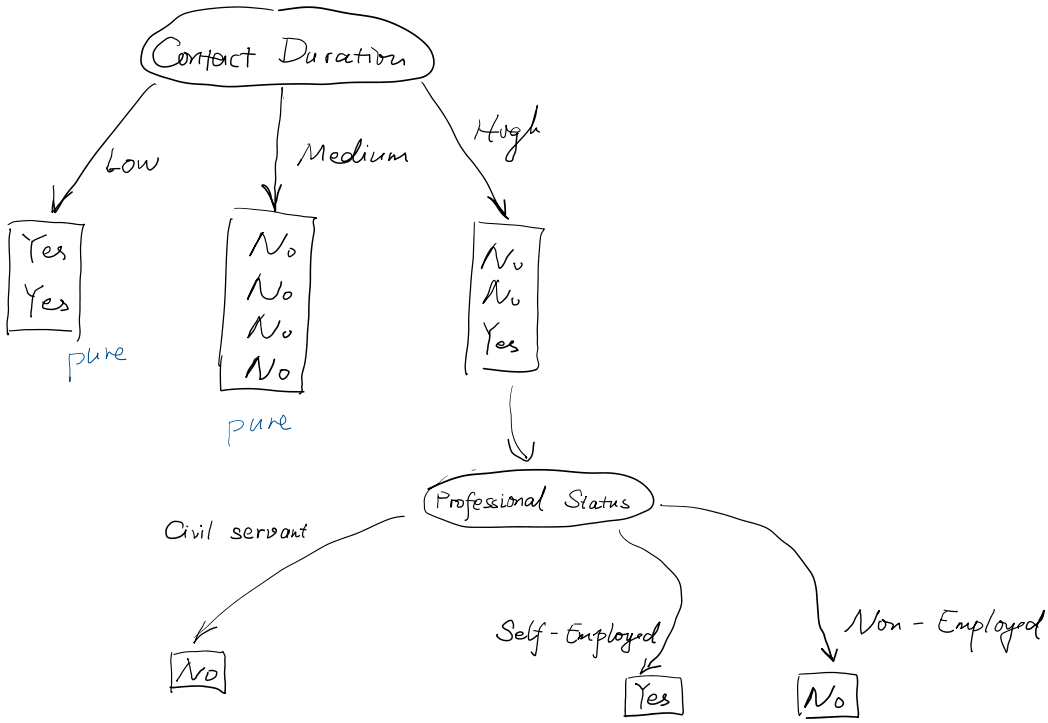
entropy([2,0]) = entropy([0,4]) = 0

entropy([1,2]) = 0.918 bits

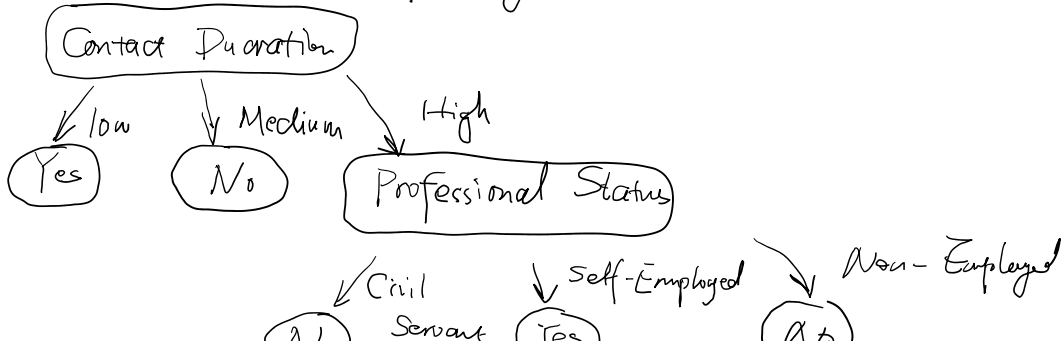
average: $\frac{2}{9} \cdot 0 + \frac{4}{9} \cdot 0 + \frac{3}{9} \cdot 0.918 = 0.306$ bits

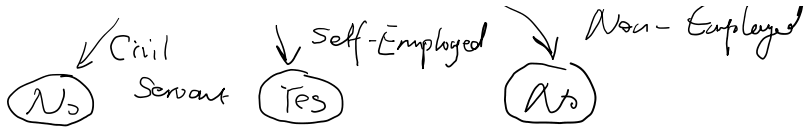
Information gain: 0.959 − 0.306 = 0.653 bits > 0.209 = gain(Professional Status)

⇒ choose contact duration as root.



⇒ The decision tree is the following





b) I1: Low contact duration → Yes

I2: Medium " " → No

c)

IF Contact_Duration = Low THEN Termination = Yes
IF Contact_Duration = Medium THEN Termination = No
IF Contact_Duration = High AND Professional_Status = Civil Servant THEN Termination = No
IF Contact_Duration = High AND Professional_Status = Self-Employed THEN Termination = Yes
IF Contact_Duration = High AND Professional_Status = Non-Employed THEN Termination = No

d) All the three methods can be used for decision tree splitting.

The information gain and gain ratio find the best attributes with highest purity in splitting the tree by comparing the entropy loss. The higher the loss is, the better the attribute splits the data. The information gain is especially suited for categorical attributes and is susceptible to attribute with higher number of categories. Gain ratio method considers the size of the subsets s.t it overcomes the disadvantage of information gain method. The Gini index approach, on the other hand, applies different formula in finding the impurity of the attribute. The lower the impurity is, the better.

e) 2 Low: [2,0] . $Gini = 1 - 1^2 = 0$

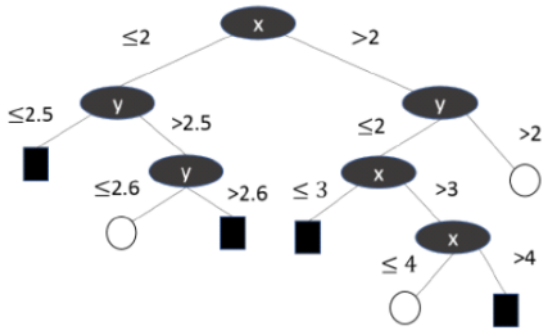
4 Medium: [0,4] $Gini = 1 - 1^2 = 0$

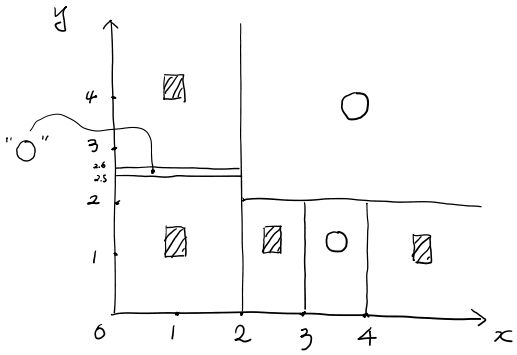
3 High: [1,2] $Gini = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = \frac{5}{9}$

Average: $0 \cdot \frac{2}{9} + 0 \cdot \frac{4}{9} + \frac{5}{9} \cdot \frac{3}{9} = \frac{5}{27}$

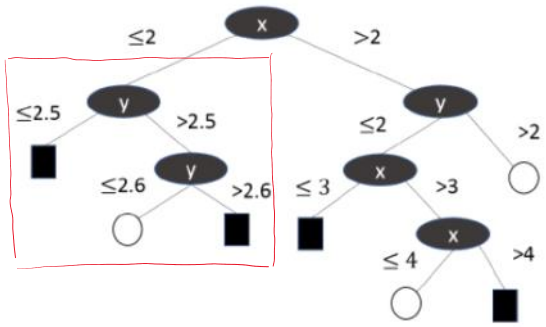
Task 2: Decision trees (3 Points)

a) Specify the partitioning for the following decision tree. Draw a coordinate system for this. Make sure that the partition identifiers match the tree leaves. (2 P.)





b) Which partition can be described as overfitted? In the decision tree, mark the nodes or subtrees that must be removed to avoid overfitting. (0,5 P.)



c) What does the generalization capability of a classifier mean? (0,5 P.)

The generalization capability is the ability to keep the accuracy of classifying unseen data relatively close to that of classifying the trained data, meaning the model does not "memorize" the pattern in the training dataset.