Exercise 8

Task 1. a) first, we draw the correspondence graph of each attribute and target:

Professional Status

civil servant	Employee	self-employed	non-employed
yes	yes	yes	no
no	no	no	no
	no		

Contract Duration

Low	medium	high
yes	no	~~no~~ yes
yes	no	no
	no	no
	no	

Sencond, we compute the information gain for each attribute:

Professional Status: $\frac{2}{9}\cdot \text{entropy}([1,1]) + \frac{3}{9}\cdot \text{entropy}([1,2]) + \frac{2}{9}\cdot \text{entropy}([1,1]) + 0$
$= \frac{4}{9} + 0.3061 = 0.7505$ bits

Contract Duration: $\frac{2}{9}\cdot 0 + \frac{4}{9}\cdot 0 + \frac{1}{3}\cdot(0.5283+0.39) = 0.3061$ bits

Information gain$_P$ = entropy([3,6]) − 0.7505 = 0.9183 − 0.7505 = 0.1678 bits

Information gain$_C$ = 0.9183 − 0.3061 = 0.6122 bits

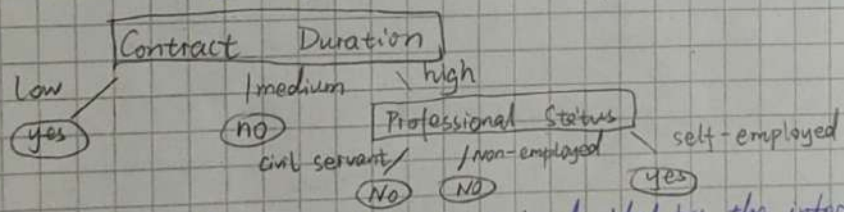
Therefore, we chose the variable „Contract Duration" as the root:

Since we have only „yes" for „low", „no" for „medium" of variabe „Contract Duration", we complete to split for them. Now we only need to compute the information gain for „high":

Entropy(Contract Duration = high) = Entropy([1,2]) = 0.9183

Entropy(Contract Duration = high | Professional Status) = 0

the resulting Tree:



The order of the attributs of the resulting tree is decided by the information gain. We pick the attribute with max information gain for each split.

The leaf nodes is yes/no.

b) We clasfy Number 11 & 12 by the tree from (a):

No. 11 with „Low" of „Contract Duration", it is classified to „yes" of termination

No. 12 with „medium" of „Contract Duration", it is classified to „no".

Hence Customer with number 11 is at risk of termination.

c) Customer with low contract duration tend to terminate; with medium duration tend to not terminate. For customer with high duration, the ones ~~with~~ who are self-employed at risk of termination. Unexpectively, customer with high duration but ~~with~~ non-employed tend to not terminate.

d) Assume dataset X and attribute V with n possible different values ($v_1, v_2, \cdots v_n$).

The information gain of splitting X with ~~V~~ is:

$$\text{Gain}(X, v) = \text{Entropy}(X) - \cancel{\sum_{i=1}^{n} \frac{|X_i|}{|X|}} \text{Entropy}(X|v_1, v_2, \cdots, v_n)$$

The information gain ratio is:

$$\text{Gainratio}(X, v) = \frac{\text{Gain}(X, v)}{-\sum_{i=1}^{n} \frac{|X_i|}{|X|} \log_2 \frac{|X_i|}{|X|}} \text{ - splitinformation}$$

From the definition of Gain(X, v), the attribute with more possible values (higher possibility with pure classification) might has larger information gain and so be chosen. Therefore we introduce ~~Gainration~~ to reduce the bias caused by higher-branching attribute (larger n means larger splitinformation)

The definition of Gini index:

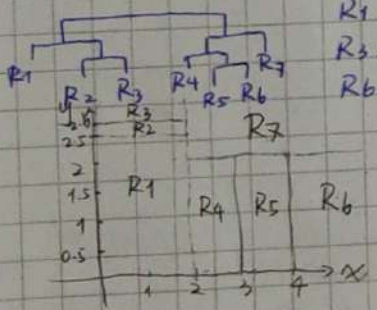
$$\text{Gini}(X, v) = \sum_{i=1}^{n} \frac{|X_i|}{|X|} \text{Gini}(X_i) \overset{?}{=} 1 - P^2 - (1-P)^2 = 2P(1-P)$$

, assume there are 2 possible outcomes A, B. P is the possibility of event A.

e) $\text{Gini (Contract Duration)} = \cancel{\frac{3}{4}(\frac{2}{}} \frac{2}{9} \cdot 0 + \frac{4}{9} \cdot 0 + \frac{3}{9} \cdot 2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{27}$

Task 2.

a)



R_1: $x \le 2$, $y \le 2.5$; □ R_2: $x \le 2$, $2.5 < y \le 2.6$; ○

R_3: $x \le 2$, $y > 2.6$; □ R_4: $2 < x \le 3$, $y \le 2$; □ R_5: $3 < x \le 4$, $y \le 2$; ○

R_6: $x > 4$, $y \le 2$; □ R_7: $x > 2$, $y > 2$. ○

b) Patition R_2 and R_5 can be described as overfitted. The subtrees of R_2 R_3 and R_5 R_6 should be removed.

c) The generalization capability of a classifier mean the classification not only works well on the training data, but also on the testing data. (new data)