COMP4432 Machine Learning

Tutorial Questions on Decision Tree

1. Given the following table.

Parcel ID	*Origin*	*Destination*	*Type*	*Weight*
1	HK	HK	Parcel	Light
2	Kln	Kln	Letter	Light
3	NT	Kln	Letter	Light
4	HK	HK	Parcel	Heavy
5	Kln	Kln	Parcel	Light
6	NT	NT	Letter	Light
7	HK	HK	Letter	Light
8	Kln	Kln	Parcel	Heavy
9	Kln	Kln	Letter	Light
10	HK	HK	Letter	Light
11	HK	HK	Parcel	Heavy
12	Kln	Kln	Letter	Light
13	HK	HK	Letter	Light
14	Kln	Kln	Parcel	Light
15	HK	NT	Parcel	Heavy
16	NT	Kln	Letter	Light
17	HK	NT	Letter	Light
18	Kln	HK	Parcel	Light
19	HK	NT	Parcel	Heavy
20	HK	HK	Parcel	Light
21	Kln	Kln	Letter	Light
22	Kln	HK	Parcel	Heavy
23	Kln	Kln	Letter	Light
24	Kln	Kln	Letter	Light
25	HK	HK	Parcel	Light

Construct a decision tree, based on information gain, to classify the type of courier services (cf. column *Type*). You may assume that the first 20 records are available for model construction and the remaining 5 records are used to validate your answer.

Answer.

1. Determining the root attribute:

I(p,n)=(10,10)=1

Entropy for Origin

Origin	p_i	n_i	$I(p_i , n_i)$
HK	6	4	0.97
Kln	4	3	0.985
NT	0	3	0

Entropy=(10/20)*I(6,4) + (7/20)*I(4,3) + (3/20)*I(0,3)=0.83

Information_Gain(Origin)=1-0.83=0.17

Entropy for Destination

Dest.	p_i	n_i	$I(p_i , n_i)$
HK	5	3	0.954
Kln	3	5	0.954
NT	2	2	1

Information_Gain(Dest.)=1-(8/20)*0.954-(8/20)*0.954-(4/20)*1=0.0368

Entropy for Weight

Weight	p_i	N_i	$I(p_i , n_i)$
Light	5	10	0.918
Heavy	5	0	0

Information_Gain(Weight)=1-(15/20)*0.918-(5/20)*0=0.312

Hence, *Weight* is selected as the decision attribute for the root node. The above steps will be repeated to build the sub-trees.

---

2. Determining the internal node attributes:

Since we have two branches from the root and one (Weight=Heavy) can be terminated, there will only have one sub-tree under the root. Then, it is needed to determine the node attribute when (Weight=Light):

I(p,n)=(5,10)=0.918

Entropy for Origin

Origin	p_i	n_i	$I(p_i , n_i)$
HK	2	4	0.918
Kln	3	3	1
NT	0	3	0

Entropy=(6/15)*I(2,4) + (6/15)*I(3,3) + (3/15)*I(0,3)=0.767

Information_Gain(Origin)=0.918-0.767=0.151

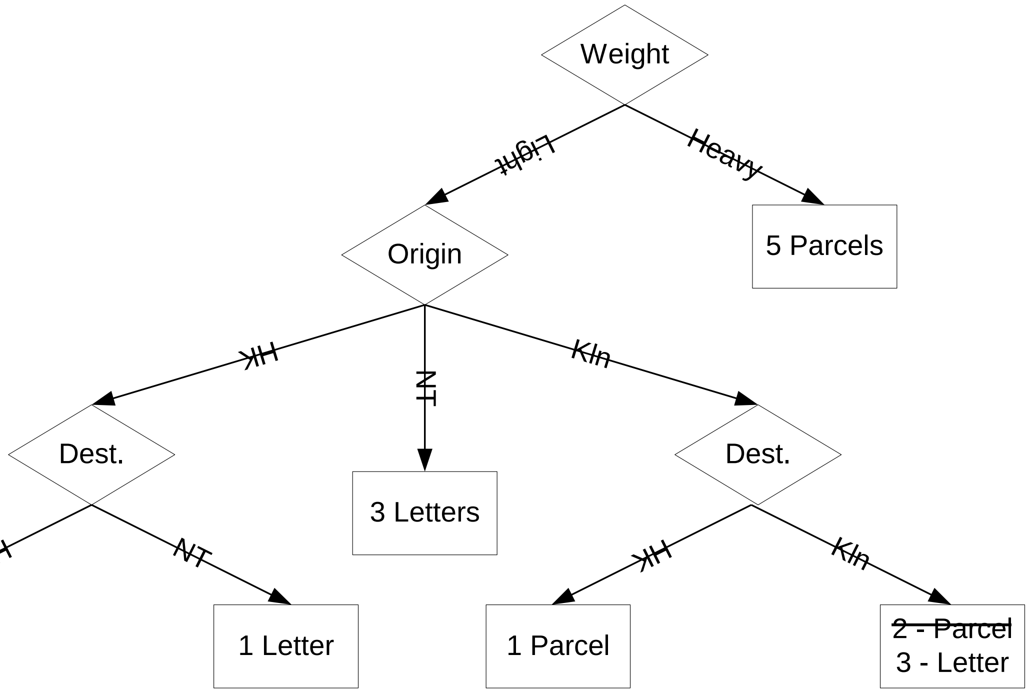
Entropy for Destination

Dest.	p_i	n_i	$I(p_i , n_i)$
HK	3	3	1
Kln	2	5	0.863
NT	0	2	0

Entropy=(6/15)*I(3,3) + (7/15)*I(2,5) + (2/15)*I(0,2)=0.803

Information_Gain(Dest.)= 0.918-0.803=0.115

This time, *Origin* is selected as the decision attribute for this node. The final decision tree can then be built by taking the last available attribute into considerations and it is shown below.



Except the last record, i.e., parcel ID 25, all testing records can be classified correctly. The classification rate on the testing data is then equal to 80%.

2. You are given the following 40 2-D points belonging to two classes, i.e., blue class and orange class. They are integer-aligned, i.e., (0,0; blue), (2,0; blue), (4,0; blue), (6,0; blue), (8,0; blue), (10,0; orange), (12,0; orange), (14,0; orange), (0,2; blue), … , (14,8; orange).

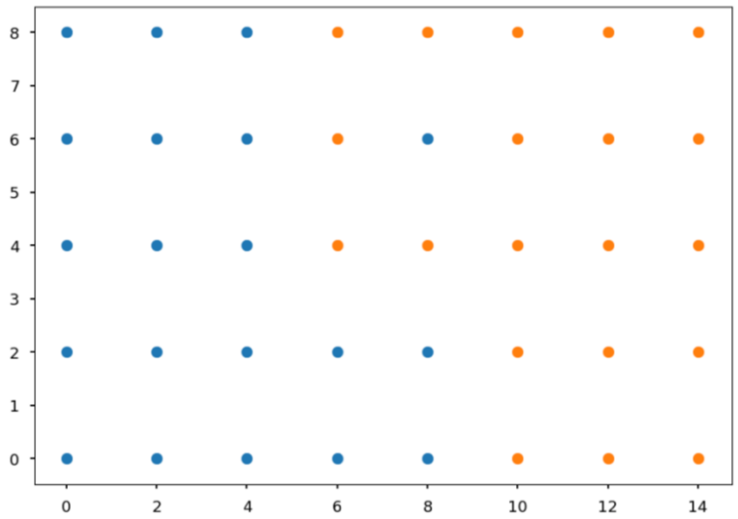


Fig.2 2D Dataset for Q.2

a) According to the idea of ID3 (and referring to slide 31 of lecture notes-DT), use the available 40 points to build a decision tree with classification accuracy (training)≥95%. Comprehensive steps of your computation are NOT required but indicative ones are expected.

Answer:
(binary tree assumption; it is only an illustrative solution):
For the two dimensions say *x* and *y* here, we need to compute the entropy of the following splits:

- *x*=1, *x*=3, *x*=5, *x*=7, *x*=9, *x*=11, *x*=13 and
- *y*=1, *y*=3, *y*=5, *y*=7

Obviously, some splits are meaningless, including *x*=1, *x*=3, *x*=11, *x*=13. Some splits are redundant, including *x*=5 and *x*=9 (choose either one) and *y*=1 and *y*=7 (choose either one). It is also obvious that splitting *x* will have smaller entropy than splitting y.
So for the root node of the tree, you may just show the entropy computation of splitting at *x*=5, *x*=7. Hence, we have

Expected Entropy (Split at *x*=5)$= \frac{15}{40}\cdot I(15,0) + \frac{25}{40} I(5,20)$

$= \frac{15}{40}\cdot\left(-\frac{15}{15}\log_2\frac{15}{15} - \frac{0}{15}\log_2\frac{0}{15}\right) + \frac{25}{40}\cdot\left(-\frac{5}{25}\log_2\frac{5}{25} - \frac{20}{25}\log_2\frac{20}{25}\right)$

=0+$\frac{25}{40}\cdot 0.722$=0.451

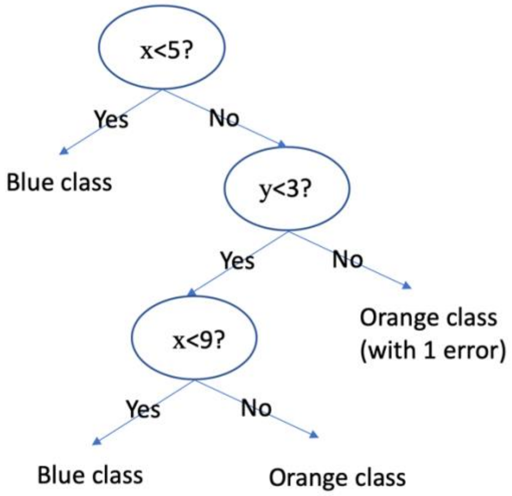
Expected Entropy (Split at *x*=7)$= \frac{20}{40}\cdot I(17,3) + \frac{20}{40} I(3,17) = 0.610$

So, the root node of the tree is "*x*<5?" (or "*x*<9?").

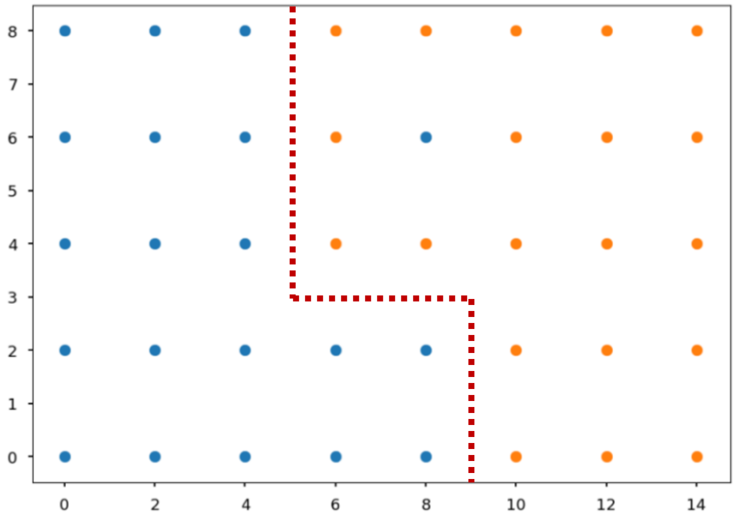
After that, for "*x*<5?=Yes", terminate and create a leaf node.
For "*x*<5?=No", we need to grow a sub-tree as follows.
We need to consider the following splits:

- *x*=7, *x*=9, *x*=11, *x*=13 and
- *y*=1, *y*=3, *y*=5, *y*=7

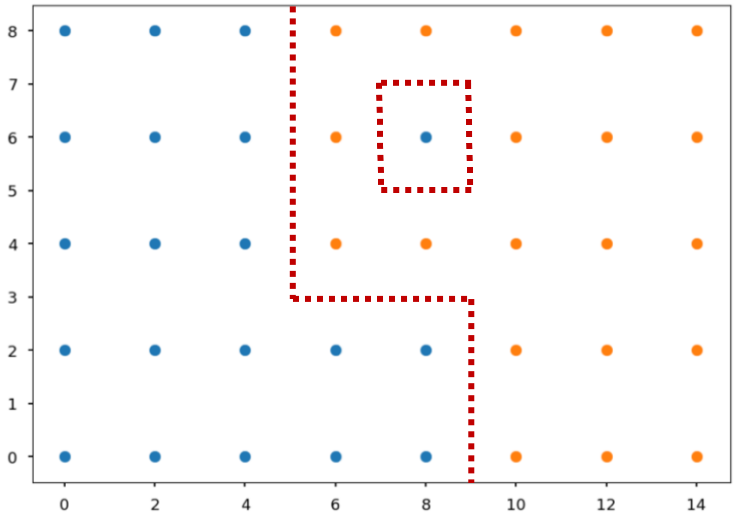
It seems that splitting at *y*=3 should be chosen to minimize impurity (lower entropy). You may want to justify it by showing entropy calculations as exemplified above. Repeat such steps and wrt classification accuracy (training)≥95%, the following DT is obtained.

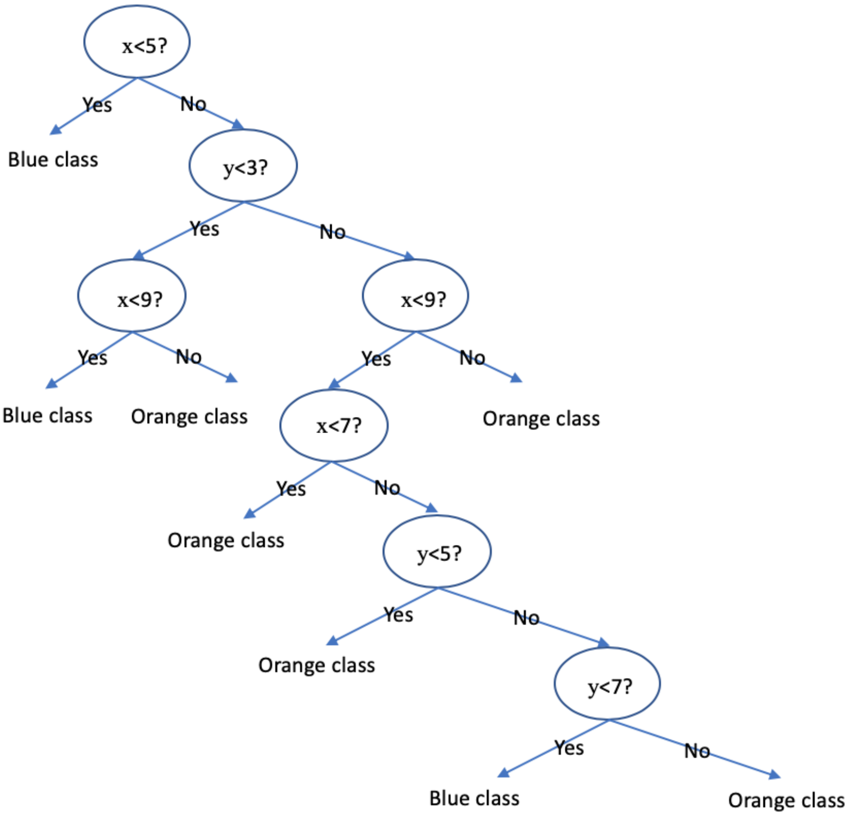


b) Draw the decision surface of the decision tree you build in part (a).



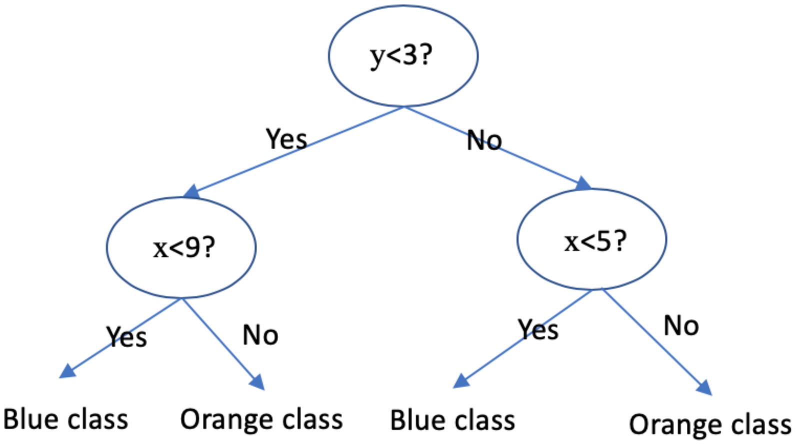
c) If the classification accuracy (training) is now lifted to 100%, build a corresponding decision tree and draw its decision surface. Again, a concise answer is not needed.





d) As a second thought, can we obtain a decision tree with 2 levels only (root→level_1→ leaves) and axis-parallel decision nodes for the criterion of classification accuracy (training) ≥95%?

Answer: Yes, a decision tree like the following can satisfy the requirement.


y<3?
Yes
No
x<9?
x<5?
Yes
No
Yes
No
Blue class
Orange class
Blue class
Orange class