Decision Trees

Consider the following data set. Use information theory to design a decision tree that classifies people according to whether they are a “cat-person” or a “dog-person”.

#	Introvert	Stylish	Has-Mice	Category
1	T	T	T	Cat-person
2	T	T	F	Cat-person
3	F	F	T	Dog-person
4	T	F	F	Dog-person
5	T	F	T	Cat-person
6	F	T	F	Dog-person
7	T	T	F	Cat-person
8	F	F	T	Dog-person

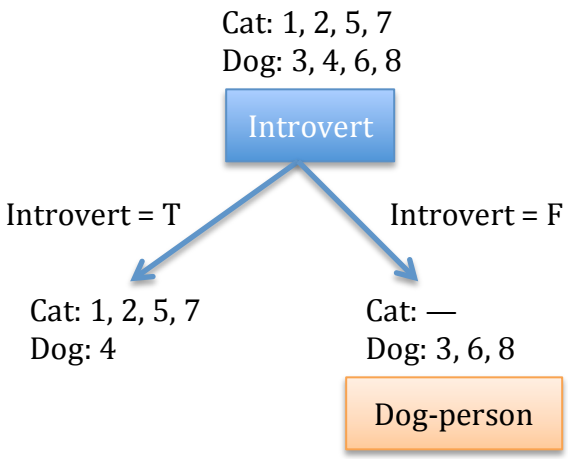
Solution:

The root:

Gain(introvert) = I(4/8, 4/8) - [5/8 * I(4/5, 1/5) + 3/8 * I(0/3, 3/3)]
= 1 – 0.4512050593046014
= 0.5487949406953986

Gain(stylish) = I(4/8, 4/8) - [4/8 * I(3/4, 1/4) + 4/8 * I(1/4, 3/4)]
= 1 - 0.8112781244591328
= 0.18872187554086717

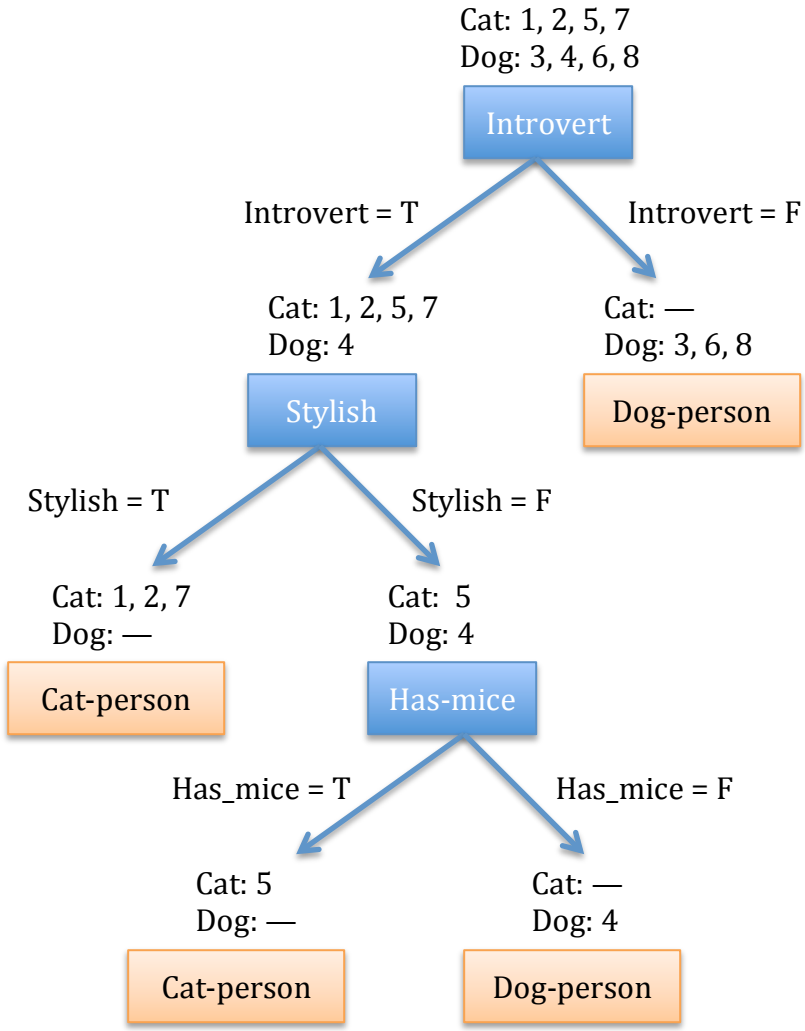
Gain(has-mice) = I(4/8, 4/8) – [4/8 * I(2/4, 2/4) + 4/8 * I(2/4, 2/4)]
= 1 - 1
= 0



The introvert=T branch:

Gain(stylish) = I(4/5, 1/5) – [3/5 * I(3/3, 0/3) + 2/5 * I(1/2, 1/2)]
= 0.7219280948873623 - 0.4
= 0.3219280948873623

Gain(has-mice) = I(4/5, 1/5) – [2/5 * I(2/2, 0/2) + 3/5 * I(2/3, 1/3)]
= 0.7219280948873623 - 0.5509775004326937
= 0.17095059445466865



By process of elimination the introvert=T, stylist=F branch must use the has-mice attribute since it is the only attribute that remains.