Due Date: Webcampus
How to submit: Webcampus

HW2-1: Consider the training example shown in Table 1, for a binary classification of mammals.

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	yes
whale	warm-blooded	yes	no	no	yes
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no
python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

a- What is the entropy of this collection of training set?
b- What are the Information Gains of splitting "**Body temperature**" and "**Give Birth**"?
c- Between "**Give Birth**", and "**four legged**" what is the best split according to the **classification error rate**?
d- Between "**Give Birth**", and "**four legged**" what is the best split according to the **Gini index**?
e- Use "**Give Birth**" as the first split parameter, and "**four legged**" as the second split parameter. Draw the tree then calculate information gain.

HW2-2. Model Overfitting

Consider the decision trees (T_1 and T_2) shown in Figures 1 and 2 respectively.

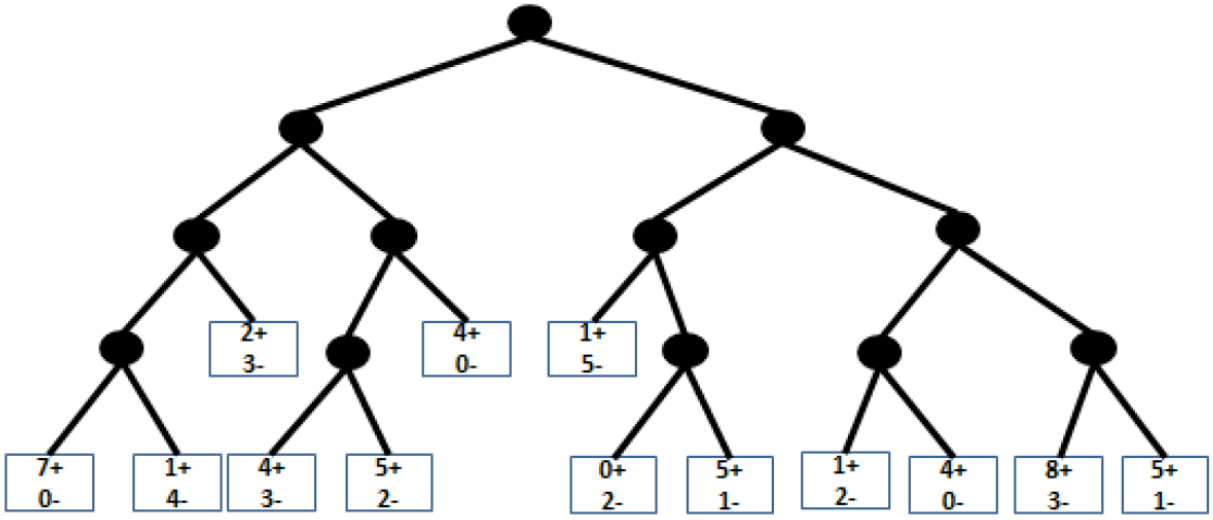

Figure 1: T_1

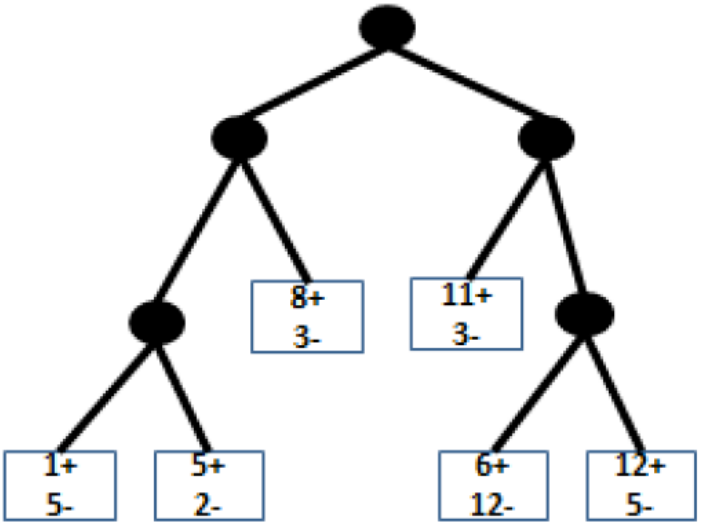

Figure 2: T_2

a) Compute the generalization error rate of the trees using optimistic approach.

b) Compute the generalization error rate of the trees using pessimistic approach. Use penalty term = 0.5, 0.75, and 1.

c) Based on the generalization error from part (a) and (b), which tree is preferred and why?

d) Based on Occam's Razor, which tree is better and why?

HW2-3. Model Overfitting

Generate the dataset as in slide 56 in Chapter 3. 10 % of the data is used for training and 90% of the data used for testing. Reproduce the right figure in slide 61 in Chapter 3 and submit your python code in ipynb format.