HW-2
Ehsan Mosadegh

Question-1

A-

Entropy= $-(0.4*\log_2 0.4) - (0.6*\log_2 0.6) = +0.970$

B-

Information gain based on Entropy

Feature= Body Temperature
Entropy(cold)== purest split = 0
Entropy(warm)= $-(0.8 \log_2 0.8) – (0.2 \log_2 0.2) = +0.721$
Information Gain= 0.970 – (0) – (0.5*0.721) = 0.609

Feature= Give Birth
Entropy(no)== pure dataset = 0
Entropy(yes)= $-(0.2 \log_2 0.2) – (0.8\log_2 0.8) = 0.721$
Information Gain= 0.970 – (0) – (0.5*0.721) = 0.6095

C-

Error rate= 1-max P_i

Error rate of total dataset/parent node= 1 – (0.6) = 0.4

Feature= Give Birth
Error rate(no) = 0
Error rate(yes) = 1- (max {4/5 , 1/5} = 0.8) = 0.2
Information gain= 0.4 – (0*weight) – (0.2*weight=(5/10)) = 0.3

Feature= Four Legged
Error rate(split= no)= 1 - (max{2/6 , 4/6}= 4/6)= 1/3
Error rate(split= yes)= 1 – (max{1/2 , 1/2}= 1/2)= 0.5
Information gain= 0.4 – (1/3*0.6) – (5/10*4/10) = 0.0
According to classification error rate and information gain, the split that gives us higher information gain is the best one to start split, so we select Give birth as the split because it has higher information gain = 0.3

D-

Gini index= 1 - (sum P_c^2)

Gini of parent dataset= 1 – $(0.4)^2$ – $(0.6)^2$ = 0.48

Feature= Give Birth
Gini(split= no)= 1 – 0 – 1 = 0
Gini(split= yes)= 1 – $(4/5)^2$ – $(1/5)^2$ = 8/25 = 0.32
Information Gain= 0.48 – (0) – (0.32*0.5) = 0.32

Feature= Four Legged
Gini(split= no)= 1 – $(2/6)^2$ – $(4/6)^2$ = 16/36 = 0.44
Gini(split= yes)= 0.5
Information Gain= 0.48 – (0.44*0.6) – (0.5*0.4) = 0.016

Gini index is a measure of impurity and we use it with the information gain to find which feature has higher information gain. In this example, between these two features the best split is the one that has the highest information gain which is "Give Birth"= 0.32

E-

No --- Give Birth --- Yes
Yes=0 Yes=4
No=5 No=1
--- no ------------ yes ---
Name Indexes=3,4,10 1,2

We calculate Information Gain based on Gini index similar to previous question
Gini(split= no) = 1 – $(2/3)^2$ – $(1/3)^2$ = 4/9
Gini(split= yes)= 1 – 0 - 1 = 0
Gini(parent)= 1 – $(4/5)^2$ – $(1/5)^2$ = 0.32
Information Gain= 0.32 – (4/9*3/5) – (0*2/5) = 0.053
So information gain for the tree at "Four Legged" node is 0.053.

Question-2
A- Generalization error using optimistic approach == training error

Error-g(T1)= 15/73= 0.21
Error-g(T2)= 20/73= 0.27

B- Generalization using pessimistic approach == training error+ penalty term * (no. of leaf nodes/ total no. of samples)

E-g(T1)= 0.21+0.5(13/73)= 0.299
E-g(T2)= 0.27+0.5(6/73)= 0.31

E-g(T1)= 0.21+0.75(13/73)= 0.344
E-g(T2)= 0.27+0.75(6/73)= 0.33

E-g(T1)= 0.21+(13/73)= 0.388
E-g(T2)= 0.27+(6/73)= 0.352

C- the test error for the bigger tree is smaller, but we see that the generalization error for the smaller tree will get smaller by increasing the penalty term. So we prefer the smaller tree (T2) because it has smaller generalization error compared to T1. It shows that T1 is overfit to the data and is memorizing the train data set, but T2 can generalize better to the data.

D- The shorter tree (T2) because although it is shorter and with less depth than T1, it can generalize better to test data. The simpler model works better than a complex model which is probable to memorize the data.