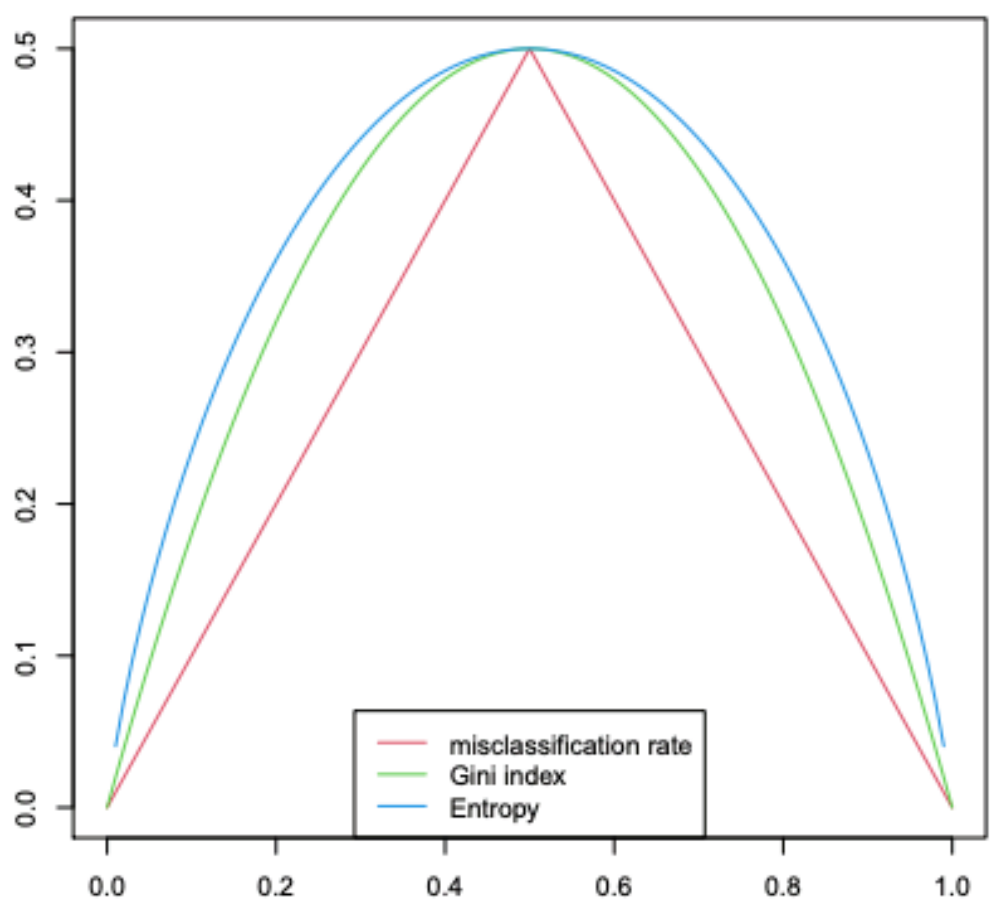
T4

Tuesday, June 22, 2021 10:57 PM

Conceptual

1. Consider the Gini index, classification error, and cross-entropy in a simple classification setting with two classes. Create a single plot that displays each of these quantities as a function of $\hat{p}_{m1}$. The x-axis should display $\hat{p}_{m1}$, ranging from 0 to 1, and the y-axis should display the value of the Gini index, classification error, and entropy.

```
p1 <- seq(0,1,by=0.01)
p2 <- 1-p1
P  <- cbind(p1,p2)
misclassification <- 1-apply(P,1,max)
   #apply(...): row-wise maximum
gini <- p1*(1-p1) + p2*(1-p2)
entropy <- -p1*log(p1) - p2*log(p2)
entropy <- entropy*0.5/max(entropy,na.rm=TRUE)
   #scaled to pass through (0.5,0.5)
plot(p1,misclassification,type="l",col=2,ylim=c(0,0.5),ylab="",xlab="p")
lines(p1,gini,col=3)
lines(p1,entropy,col=4)
legend("bottom",legend=c("misclassification rate","Gini index",
                         "Entropy"),col=2:4,lty=1)
```



2. Biologists are interested in distinguishing between two different species of fleas: Haltica oleracea (H.o) and Haltica carduorum(H.c).

   Four variables are recorded for each observation:
   X_1: distance of the transverse groove to the posterior border of the protho- rax;
   X_2: length of the elytra;
   X_3: distance of the second antennal joint;
   X_4: length of the third antennal joint.

   A sample of 39 fleas were used to construct a classification tree.

   i. Sketch a classification tree which could be used to identify the speciestype of a new observation, which reflects the following facts about the original data.

      a) The 13 largest values of second antennal joint distance are all of species H.c and are greater than 1.52mm.

      b) Excluding the corresponding 13 observations, the largest 16 values of the distance of the transverse groove to the posterior border of the prothorax are all greater than 185mm and all but one of these are of species H.o.

      c) Of the remaining 10 observations not addressed above, 5 of the observations have elytra lengths greater than 260mm and these are all of species H.c (while the others were of species H.o).

   Be sure to include the following in your classification tree:

   - all root, internal and terminal nodes;
   - the binary split at each (non-terminal) node;
   - the predicted class at each node;
   - the probabilities of both classes at each node.

   ii. Calculate the overall misclassification rates for the tree.

2. Based on the information provided in the question, we can first construct a tree like Figure 2. The root node and decision nodes are left blank as the question does not explicitly state the number of observations in each class.

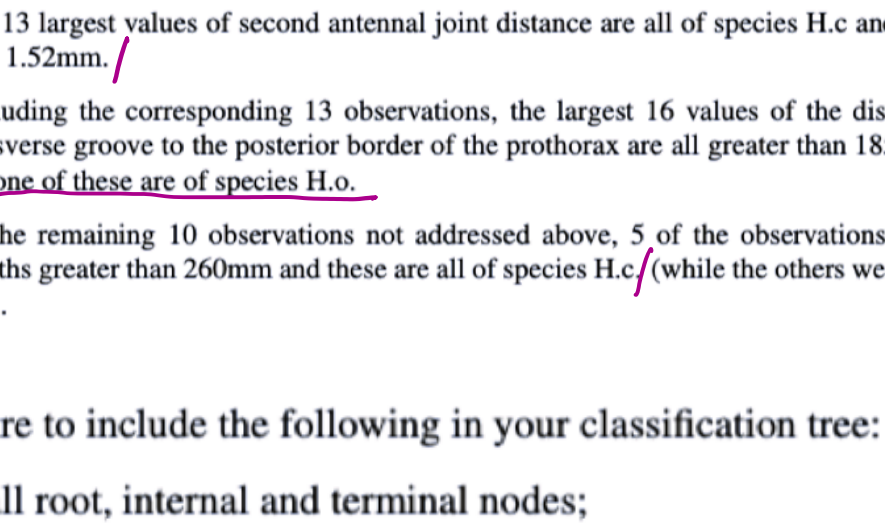


Figure 2: Initial tree

Next, we go from bottom to top, summing up the observations from the terminal nodes and deciding which class the decision nodes should predict. This gives the tree in Figure 3. There is one node marked in red, as the predicted class can be either H.c or H.o.

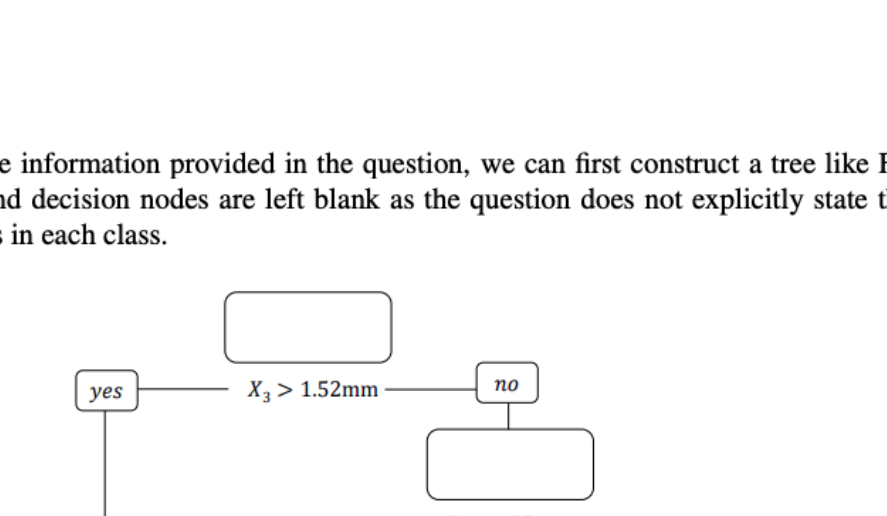


Figure 3: Full tree

There is only 1 misclassifcation, in the second left most branch, where the observation truly of class H.c was predicted to be class H.o. Therefore, the overall misclassification rate is given as $\frac{1}{39}$.