Aman Patel

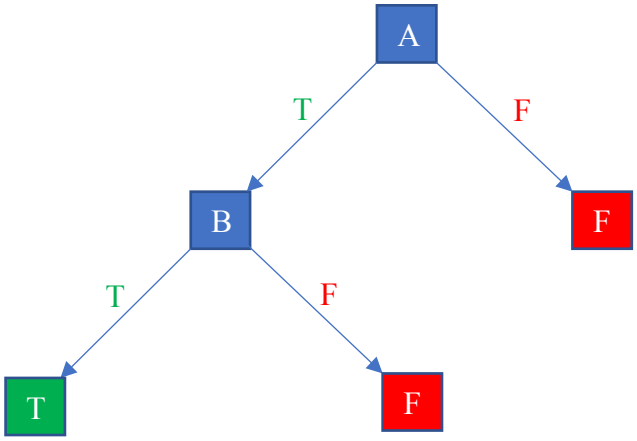
April 18, 2021

CSCI-B 455

Home Assignment 4

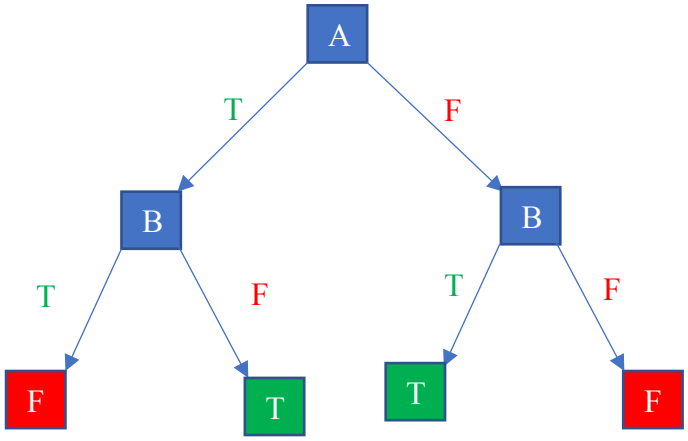
1. Decision trees for AND and XOR

   a. AND



      i. Accuracy: 1.0, this decision tree accounts for all combinations of A and B

   b. XOR



i. Accuracy: 1.0, this decision tree accounts for all combinations of A and B

2. Entropy

   a. Jim: $-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=0.5+0.5=1$

   b. Jane: $-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0.5+\ 0.311=0.811$

   c. Sarah: $-\frac{1}{8}\log_2\frac{1}{8}-\frac{7}{8}\log_2\frac{7}{8}=0.375+0.169=0.544$

   d. Sam: $-\frac{1}{8}\log_2\frac{1}{8}-\frac{7}{8}\log_2\frac{7}{8}=0.375+0.169=0.544$

   e. John: $-1\log_2 1=0$

   f. Average number of questions

      i. Ask Jim. If Y, only 1 question needed. $P(Y)=\frac{1}{2}$

      ii. If N, ask Jane. If Y, 2 questions needed. $P(Y)=\frac{1}{4}$

      iii. If N, ask Sarah. If Y, 3 questions needed. $P(Y)=\frac{1}{8}$

      iv. If N, ask Sam. If Y, conclude Sam. If N, conclude John. $P(Y)=\frac{1}{8}$

      v. $E(\#Q)=1*\frac{1}{2}+2*\frac{1}{4}+3*\frac{1}{8}+4*\frac{1}{8}=1.875$ questions

3. ID3 decision tree

   a. $Entropy(S)=\ -\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0.5+0.311=0.811$

   b. Gender

      i. $\frac{|S_T|}{|S|}\,entropy(S_T)=\frac{1}{2}\left[-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right]=0.5$

      ii. $\frac{|S_F|}{|S|}\,entropy(S_F)=\frac{1}{2}[-1\log_2 1]=0$

      iii. $Gain(S,Gender)=0.811-0.5=\ 0.311$

   c. Student

      i. $\frac{|S_T|}{|S|}\,entropy(S_T)=\frac{1}{2}\left[-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}\right]=0.406$

      ii. $\frac{|S_F|}{|S|}\,entropy(S_F)=\frac{1}{2}\left[-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}\right]=0.406$

      iii. $Gain(S,Student)=0.811-0.406-0.406=0$

d. Pub last night

   i. $\frac{|S_T|}{|S|}\,entropy(S_T) = \frac{5}{8}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] = 0.607$
   ii. $\frac{|S_F|}{|S|}\,entropy(S_F) = \frac{3}{8}[-1\log_2 1] = 0$
   iii. $Gain(S, Pub) = 0.811 - 0.607 = 0.204$

e. Gender has the highest information gain; it will be the root of the tree.

f. All males chose Vodka, $\therefore Gender = F$ will be a leaf with classification Vodka.

g. $Gender = T$ (females)

   i. $Entropy(S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$
   ii. Student

      1. $\frac{|S_T|}{|S|}\,entropy(S_T) = \frac{1}{4}[-1\log_2 1] = 0$
      2. $\frac{|S_F|}{|S|}\,entropy(S_F) = \frac{3}{4}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right] = 0.688$
      3. $Gain(S, Student) = 1 - \frac{3}{4}[0.528 + 0.390] = 0.312$

   iii. Pub last night

      1. $\frac{|S_T|}{|S|}\,entropy(S_T) = \frac{1}{2}[-1\log_2 1] = 0$
      2. $\frac{|S_F|}{|S|}\,entropy(S_F) = \frac{1}{2}[-1\log_2 1] = 0$
      3. $Gain(S, Pub) = 1 - 0 - 0 = 1$

   iv. Pub has the highest information gain; it will be the next child.

   v. All females who went to the pub last night chose Beer $\therefore$ leaf node.

   vi. All females who did not go to the pub chose Vodka $\therefore$ leaf node.

h. Prediction for $\{Gender = T, Student = T, Pub = F\}$

   i. $Gender = T \rightarrow Pub = F \rightarrow$ **Vodka**

4. See attached notebook.

5. Stumping will result in an accuracy greater than 50% in binary classification problems because the predictor chooses the feature that has the highest classification rate when used on its own. Therefore, this feature will have a classification rate greater than or equal to 50%, as at least one of the two classes must have probability greater than or

equal to 50%. For multi-class classification, the classification rate will be greater than or equal to $\frac{1}{n}$, where n is the number of classes.

6. See attached notebook (Discussed this problem with Taral Shah).
7. Dr. Smart is correct because the pruning algorithm is not needed to build a random forest model. Although there is risk of overfitting in individual trees, the model has reduced overfitting due to bootstrapping between samples. Bootstrapping involves creating decision trees using random samples of the data with replacement. This process reduces overfitting and variance for decision trees. As there are multiple trees in the random forest, there is less weight on the individual trees, reducing the impact of overfitting of individual trees on the forest.