II. Decision Tree

CSCC11 Assignment 3.
Ti Zhang
1000454457

1. What is the entropy of the target variable (i.e. $H(F)$)?
2. What is the entropy of the target variable given the item size being small (i.e. $H(F|S = \text{Small})$)?
3. What is the entropy of the target variable given the item size being not small (i.e. $H(F|S = \text{Medium or Large})$)?
4. What is the information gain if we split based on fruit size being small?
5. At the root node, how many split tests do we need to consider and what is the best split?
6. Draw a picture of a possible final decision tree following the algorithm, and specify the split function in each internal node and the probability of being fruit is specified in each leaf node.

1. Fruit (F) have 7 true & 5 false.

$$H(F) = -\frac{7}{12}\log(\frac{7}{12}) - \frac{5}{12}\log(\frac{5}{12}) = 0.97986 \approx 0.98$$

2. Fruit(F) has 4 ture & 0 false in condition "small"

$$H(F|S=\text{small}) = -P_1 \cdot \log_2 P_1 - P_2 \cdot \log_2 P_2 = -1 \cdot \log_2 1 = 0$$

3. Fruit(F) has 3 ture & 5 false in condition "medium or large"

$$H(F|S=\text{m or L}) = -P_1 \cdot \log_2 P_1 - P_2 \cdot \log_2 P_2 = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.95443\ldots \approx 0.95$$

4. In total 12 objects, we have 4 small & 8 medium or large.

$$\begin{aligned}\text{Information Gain} &= H(F) - H(F|S=\text{small}) \cdot P(\text{small}) - H(F|S=\text{m or L}) \cdot P(\text{m or L}) \\ &= 0.97986 - 0 - 0.95443 \cdot \frac{8}{12} \\ &= 0.34357\ldots \approx 0.3436\end{aligned}$$

5. S has 3 attributes → 2 split ways
   C has 2 attributes → 1 split ways
   T has 2 attributes → 1 split ways
   } 2+1+1 = 4 ways to split.

(1) small / not small : $IG_1 = 0.3436$

(2) large / not large : $IG_2 = H(F) - H(F|S=\text{large}) \cdot P(\text{large}) - H(F|S=\text{s or m}) \cdot P(\text{s or m})$

$$= 0.97986 - [-\frac{1}{4}\cdot\log\frac{1}{4} - \frac{3}{4}\cdot\log\frac{3}{4}]\cdot\frac{4}{12} - [-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\cdot\log\frac{1}{4}]\cdot\frac{8}{12}$$

$$= 0.97986 - 0.81127 = 0.16859\ldots \approx 0.1686$$

(3) red / yellow : $IG_3 = H(F) - H(F|C=\text{red}) \cdot P(\text{red}) - H(F|C=\text{yellow}) \cdot P(\text{yellow})$

$$= 0.97986 - [-\frac{2}{3}\cdot\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}]\cdot\frac{6}{12} - [-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}]\cdot\frac{6}{12}$$

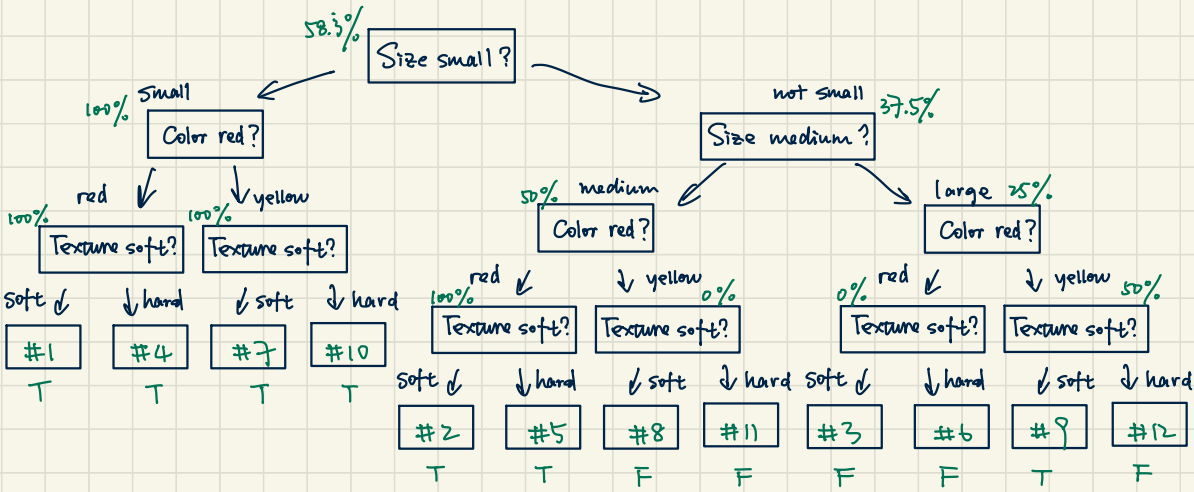
$$= 0.97986 - 0.91829\cdot 0.5 - 1\cdot 0.5 = 0.020715\ldots \approx 0.02$$

(4) soft / hard : $IG_4 = H(F) - H(F|T=\text{soft}) \cdot P(\text{soft}) - H(F|T=\text{hard}) \cdot P(\text{hard})$

$$= 0.97986 - [-\frac{2}{3}\cdot\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}]\cdot\frac{6}{12} - [-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}]\cdot\frac{6}{12} \approx 0.02$$

∴ The best split is to split by size "small" and "not small", as its IG largest. ▨

6. 6. Draw a picture of a possible final decision tree following the algorithm, and specify the split function in each internal node and the probability of being fruit is specified in each leaf node.



Percentages in green are the probability of being fruit in specified condition.

2. Random Forest

1. Suppose the dataset samples are not independent. Instead, each pair of sampled datasets are correlated with a positive pairwise correlation ρ (also known as Pearson Correlation), and let $\text{Var}[y_i] = \sigma^2$ for all i. Show that,

$$\text{Var}\left[\frac{1}{M}\sum_{i=1}^{M} y_i\right] = \frac{1-\rho}{M}\sigma^2 + \rho\sigma^2. \tag{5}$$

Hint: The pairwise correlation $\rho_{X,Y}$ between two random variables (X, Y) is

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{6}$$

where Cov is the covariance, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

$$\begin{aligned}
\text{LHS} = \text{Var}\left(\tfrac{1}{M}\textstyle\sum y_i\right) &= \tfrac{1}{M^2}\,\text{Var}\left(\textstyle\sum y_i\right) \\
&= \tfrac{1}{M^2}\left[\textstyle\sum_i^M \text{Var}(y_i) + 2\sum_{i\neq j}\text{cov}(y_i, y_j)\right] \\
&= \tfrac{1}{M^2}\left[\textstyle\sum_i^M \sigma^2 + 2\left(\sum_{i\neq j}\rho\cdot\sigma_x\cdot\sigma_y\right)\right] \\
&= \tfrac{1}{M^2}\left[M\sigma^2 + 2\cdot M(M-1)\cdot\tfrac{1}{2}\cdot\rho\cdot\sigma^2\right] \\
&= \tfrac{1}{M}\sigma^2 + \tfrac{(M-1)}{M}\rho\sigma^2 \\
&= \tfrac{1}{M}\sigma^2 + \rho\sigma^2 - \tfrac{1}{M}\rho\sigma^2 \\
&= \tfrac{(1-\rho)}{M}\sigma^2 + \rho\sigma^2 \\
&= \text{RHS as wanted.}
\end{aligned}$$

2. In few sentences, explain why the independence assumption is violated in a random forest. **Hints:** In addition to the randomness from sampling the dataset, there is also randomness from drawing a pair of decision trees grown to the randomly sampled dataset. What does decision tree evaluate to determine the split?

If the 2 decision trees has picked similar features. or the features are highly correlated.
then the 2 trees are more likely to be similar,
especially when the dominating features are same / similar / colinear / etc.
These are the conditions we cannot avoid,
thus the trees could be correlated.

3. What heuristic does random forest employ in order to decorrelate the trees?

The process in used is called: Bootstrap aggregating. also known as bagging.
This process generates new training datasets by doing sampling to original data set,
And each new set of data is used to construct a tree.
This way, covariance among trees are decreased without underfitting or generate more bias.