Meet Barot Machine Learning Problem Set 4

1.a. Need to show:

1. $\exists$ set of $\log_2|H|$ points shattered by a hypothesis space H that is finite.
2. There does not exist set of $\log_2|H|+1$ points shattered by the finite hypothesis space H.

Showing 2:

For a set of $\log_2|H|+1$ points, the number of possible 2-class labelings for the set of points is $2^{\log_2|H|+1}$, or $2|H|$. For each labeling pattern i for the set of points, there must be an additional classifier h_i that classifies the points according to that pattern. No classifier can label the same points in more than one way. Therefore, since there are $2|H|$ possible labelings, there must be at least $2|H|$ classifiers to shatter those points. Since $2|H| > |H|$, there does not exist a set of $\log_2|H|+1$ points shattered by H.

Showing 1:

Let there be a set of $\log_2|H|$ points. The number of possible labelings for these points is $2^{\log_2|H|} = |H|$. Since there are $|H|$ classifiers, there exists one classifier for every possible labeling of the points. Assuming there are no constraints to the shape of the classifiers, and that each of the $\log_2|H|$ points are distinct, the classifiers could be such that they map to the labelings in a one-to-one fashion; that is, for every possible labeling, there is a classifier that satisfies that labeling. Therefore, there exists a set of $\log_2|H|$ points that is shattered by finite hypothesis space H. Therefore, the max VC dimension of finite H is $\log_2|H|$.

b. Need to show:

1. There exists a set of 2^d points that is shattered by decision trees of d features that split the tree a maximum of one time into two nodes per path from root to leaf.

Given 1, we can be sure that the VC dimension of the set of such decision trees is at least 2^d.

Showing 1:

Let there be a decision tree of d features that threshold splits once per feature. Each level of the tree has twice the number of nodes as the previous level, and there is one root node. Since the tree splits d times (once per feature), the tree has 2^d leaves. Each leaf is classified in a certain way, so each leaf corresponds to a labeled point. There exists a set of 2^d labeled points that is correctly classified by this tree. Other trees of this type can classify other labelings.

They can classify the same set of points differently and correspond to each different labeling of those points. Therefore, there exists a set of 2^d points that is shattered by the set of decision trees with d features that are threshold split once per path from root to leaf.

2\. a. Entropy: $H(Y) = -\sum_{i=1}^{k} P(Y=y_i)\log_2 P(Y=y_i)$

$$H(\text{Poisonous}) = -\frac{7}{12}\log_2\frac{7}{12} - \frac{5}{12}\log_2\frac{5}{12}$$

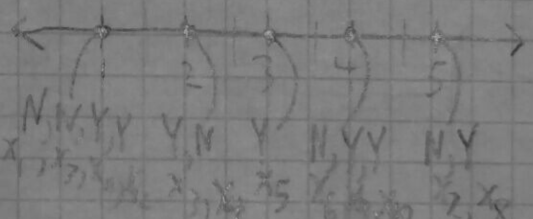
$= .979868$

b. Information gain: Decrease in entropy after splitting.

$$IG(x) = H(Y) - H(Y|x)$$

Calculating information gain for size (real-valued):

Finding threshold value:



Considering 1.5, 2.5, 3.5, and 4.5 as possible splits.

Minimum H(Poisonous | size: {1.5, 2.5, 3.5, 4.5})

$$H(Y|x:t) = p(X<t)H(Y|X<t) + p(X>=t)H(Y|X>=t)$$

$$H(Y|x:t) = p(x<t)\left(-\sum_{j=1}^{v} P(X=x_j)\sum_{i=1}^{k} P(Y=y_i|X=x_j)\log(P(Y=y_i|X=x_j))\right) +$$
$$p(x>=t)\left(-\sum_{j=1}^{v}(X=x_j)\sum_{i=1}^{k} P(Y=y_i|X=x_j)\log(P(Y=y_i|X=x_j))\right)$$

For $t=1.5$: $\frac{4}{12}\left(-\frac{4}{12}\cdot\left(\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right)-\frac{8}{12}\cdot\left(\frac{5}{8}\log\frac{5}{8}+\frac{3}{8}\log\frac{3}{8}\right)\right) +$

$\frac{8}{12}\left(-\frac{8}{12}\cdot\left(\frac{5}{8}\log\frac{5}{8}+\frac{3}{8}\log\frac{3}{8}\right)-\frac{4}{12}\cdot\left(\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right)\right)$

$= .969623$ IG: $.979868 - .969623 = 0.010245$

For $t = 2.5$:

$$\frac{6}{12}\left(-\frac{6}{12}\left(\log\frac{1}{2}\right) - \frac{6}{12}\left(\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6}\right)\right)$$

$$+ \frac{6}{12}\left(-\frac{6}{12}\left(\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6}\right) - \frac{6}{12}\left(\log\frac{1}{2}\right)\right)$$

$$= \boxed{.959148}$$

For $t = 3.5$:

$$\frac{7}{12}\left(-\frac{7}{12}\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right) - \frac{5}{12}\left(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right)\right)$$

$$+ \frac{5}{12}\left(-\frac{5}{12}\left(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right) - \frac{7}{12}\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right)\right)$$

$$= \boxed{0.9792787}$$

For $t = 4.5$:

$$\frac{10}{12}\left(-\frac{10}{12}\left(\frac{6}{10}\log\frac{6}{10} + \frac{4}{10}\log\frac{4}{10}\right) - \frac{2}{12}\left(\log\frac{1}{2}\right)\right)$$

$$+ \frac{2}{12}\left(-\frac{2}{12}\left(\log\frac{1}{2}\right) - \frac{10}{12}\left(\frac{6}{10}\log\frac{6}{10} + \frac{4}{10}\log\frac{4}{10}\right)\right)$$

$$= \boxed{0.9758}$$

$t = 2.5$ has lowest entropy, so highest (of the thresholds)

information gain: $0.979868 - .959148 = \boxed{0.02072}$

Information gain for spots:

Entropy: $H(\text{Poisonous} \mid \text{Spots}) = \frac{-7}{12}\left(\frac{5}{7}\log\frac{5}{7} + \frac{2}{7}\log\frac{2}{7}\right) - \frac{5}{12}\left(\cancel{\log_2 1}\right)$

$= \boxed{0.503487}$ $\quad IG = 0.979868 - 0.503487 = \boxed{0.476381}$

Information gain for color:

Entropy:

$$H(Poisonous|Color) = -\frac{1}{2}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) - \frac{1}{2}\left(\log\frac{1}{2}\right)$$

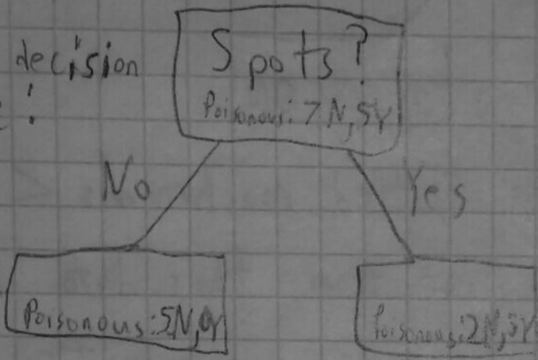
$$= 0.959148$$

$$IG = 0.979868 - 0.959148 = \boxed{0.02072}$$

Highest information gain attribute: spots, with

c. $IG = 0.476381$

d. Partial decision tree:



Next attribute to split on:

Of those with spots:

Poisonous	Size	Color
N	2	White
N	3	Brown
Y	5	White
Y	4	Brown
Y	4	Brown
Y	1	White
Y	1	Brown

Entropy of Poisonous:

$$-\frac{2}{7}\log\frac{2}{7} - \frac{5}{7}\log\frac{5}{7} = 0.86312$$

Conditional Entropy of Color:

$$-\frac{3}{7}\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) - \frac{4}{7}\left(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right)$$

$$= 0.39356 + 0.46359 = 0.85715$$

IG of Color: $0.86312 - 0.85715 = 0.00597$

Entropy of Size: thresholds: 1.5, 2.5, 3.5, 4.5

$t = 1.5$: Size < 1.5, 2Y, 0N; Size ≥ 1.5 3Y, 2N

Entropy: $\cancel{-\frac{2}{7}(\log_2 1)} - \frac{5}{7}\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right) = 0.6935$

$t = 2.5$: Size < 2.5: 2Y, 1N Size ≥ 2.5: 3Y, 1N

Entropy: $-\frac{3}{7}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) - \frac{4}{7}\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right) = 0.3936 + 0.4635$

$= 0.85719$ IG: $0.86312 - 0.85719 = 0.005933$

$t = 3.5$: Size < 3.5: 2Y, 2N Size ≥ 3.5: 3Y, 0N

Entropy: $-\frac{4}{7}\left(\log\frac{1}{2}\right) - \frac{3}{7}(\log_2 1) = -\frac{4}{7}\log\frac{1}{2} = 0.5714$

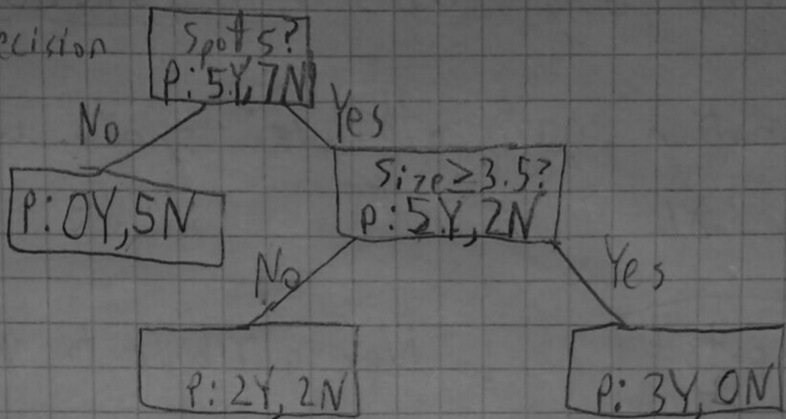
$t = 4.5$ Size < 4.5: 4Y, 2N Size ≥ 4.5: 1Y, 0N

Entropy: $-\frac{6}{7}\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right) - \cancel{\frac{1}{6}(\log 1)} = 0.78711$

Best attribute to split on: Size with threshold of 3.5,

$IG = 0.86312 - 0.5714 = \boxed{0.29172}$

Partial decision tree

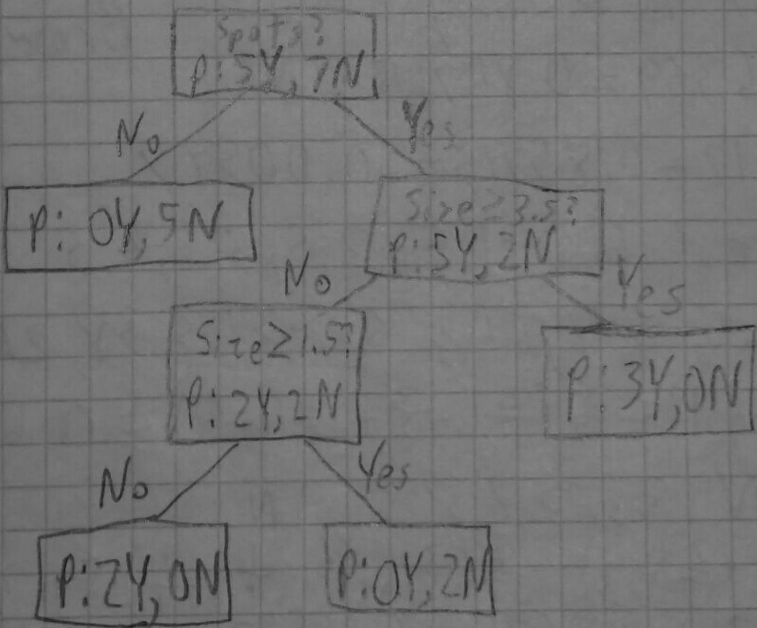


Next attribute to split on? Of Spotted mushrooms < 3.5:

Poisonous	Size	Color
N	2	White
N	3	Brown
Y	1	White
Y	1	Brown

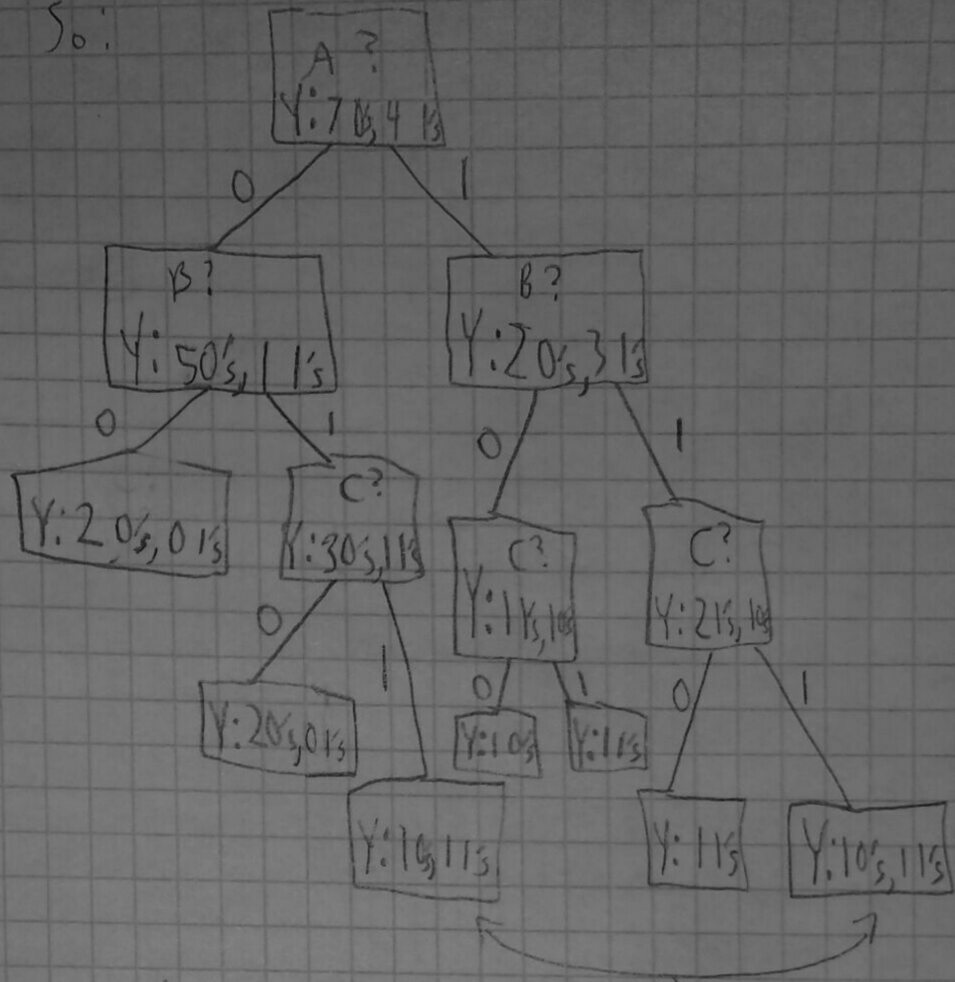
It is clear from this data that color does not give any IG here. It is apparent that a size threshold of 1.5 would accurately classify the remaining mushrooms.

Full Decision Tree



e. The classifier would first split on A, as it has the least conditional entropy, then B, then C.

So:



Training error:

$\frac{2}{11} = \boxed{18.18\%}$

Unable to expand;
so unable to classify correctly.
Predicting majority class of 0,
So 2 errors out of 11 classifications.