COMPSCI 671D HW1

1 Information Gain

1.1. The entropy of X reaches its maximum under uniform distribution.

$$p(X = x_i) = \frac{1}{n}, \quad i = 1, 2, \cdots, n$$

The maximum $H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$

1.2. $I(X, X) = H(X) - H(X|X)$

$$= H(X) - \sum_{i=1}^{n} H(X|X=x_i)\, p(X=x_i)$$

$$= H(X) + \sum_{i=1}^{n}\left(\sum_{i=1}^{n} p(x_i|x_i) \log_2 p(x_i|x_i)\right) p(x_i)$$

Since $p(x_i|x_i) = 1$, $\log_2 p(x_i|x_i) = 0$, $\sum_{i=1}^{n} p(x_i, x_i) \log_2 p(x_i|x_i) = 0$

Therefore, $I(x, x) = H(x)$

1.3. From Jensen's inequality, we have

$$E(-\log_2 x) \geq -\log_2(E(x))$$

$$\sum_{i=1}^{n} p(x_i) \log_2 \frac{p(x_i)}{q(x_i)} = -\sum_{i=1}^{n} p(x_i) \log \frac{q(x_i)}{p(x_i)} = E\left(-\log_2 \frac{q(x)}{p(x)}\right)$$

$$\geq -\log_2\left(E\left(\frac{q(x)}{p(x)}\right)\right) = -\log_2\left(\sum_{i=1}^{n} p(x_i) \frac{q(x_i)}{p(x_i)}\right) = -\log_2 \sum_{i=1}^{n} q(x_i) = 0$$

$$\therefore \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} \geq 0$$

$$I(X, Y) = H(X) - H(X|Y) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) - \sum_{j=1}^{m} H(X|y_j)\, p(y_j)$$

$$= -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) - \sum_{j=1}^{m} p(y_j)\left(-\sum_{i=1}^{n} p(x_i|y_j) \log_2 p(x_i|y_j)\right)$$

$$= -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) + \sum_{j=1}^{m}\sum_{i=1}^{n} p(x_i, y_j) \log_2 p(x_i|y_j)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 \frac{p(x_i)}{p(x_i|y_j)} = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 \frac{p(x_i)p(y_j)}{p(x_i, y_j)}$$

since $p(x_i)p(y_j) \leq p(x_i, y_j)$, $\log_2 \frac{p(x_i)p(y_j)}{p(x_i, y_j)} \leq 0$. Thus, $I(X|Y) \geq 0$

1.4. A_1 : outlook A_2 : Humidity A_3 : Windy

$H(Y) = H(\frac{9}{14}, \frac{5}{14}) \approx 0.94$

$H(Y|A_1) = \frac{9}{14} H(\frac{6}{9}, \frac{3}{9}) + \frac{5}{14} H(\frac{3}{5}, \frac{2}{5}) \approx 0.937$

$H(Y|A_2) = \frac{7}{14} H(\frac{6}{7}, \frac{1}{7}) + \frac{7}{14} H(\frac{3}{7}, \frac{4}{7}) \approx 0.788$

$H(Y|A_3) = \frac{8}{14} H(\frac{6}{8}, \frac{2}{8}) + \frac{6}{14} H(\frac{3}{6}, 3/6) \approx 0.892$

$I(Y, A_1) = H(Y) - H(Y|A_1) \approx 0.003$, $I(Y, A_2) \approx 0.152$, $I(Y|A_3) = 0.048$

Therefore we first split on Humidity (A_2)

2. Decision Trees

2.1 There are two features left, A_1 and A_3

① If Humidity = Normal, continue splitting:

$H(Y) = H(\frac{6}{7}, \frac{1}{7}) \approx 0.59167$

$H(Y|A_1) = \frac{4}{7} H(1,0) + \frac{3}{7} H(\frac{2}{3}, \frac{1}{3}) \approx 0.394$

$H(Y|A_2) = \frac{4}{7} H(1,0) + \frac{3}{7} H(\frac{2}{3}, \frac{1}{3}) \approx 0.394$

$\because I(Y, A_1) = I(Y, A_2)$ $\therefore$ We choose Outlook (A_1)

② If Humidity = High, continue splitting:

$H(Y) = H(\frac{3}{7}, \frac{4}{7}) \approx 0.985$

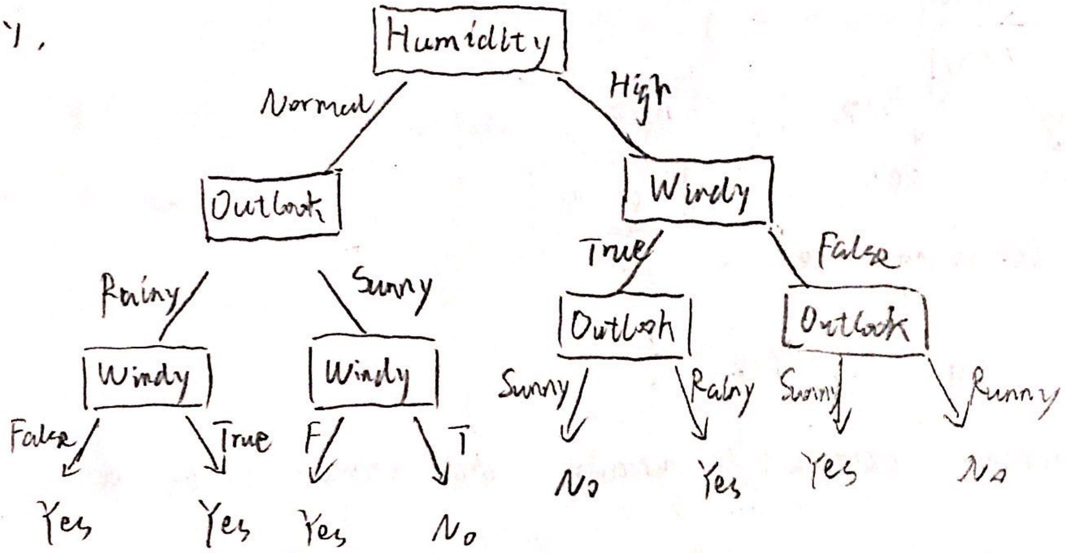
$H(Y|A_1) = \frac{5}{7} H(\frac{2}{5}, \frac{3}{5}) + \frac{2}{7} H(\frac{1}{2}, \frac{1}{2}) \approx 0.97925$

$H(Y|A_2) = \frac{4}{7} H(\frac{2}{4}, \frac{2}{4}) + \frac{3}{7} (\frac{1}{3}, \frac{2}{3}) \approx 0.965$

$I(Y, A_1) = H(Y) - H(Y, A_1) \approx 0.00397$ $I(Y, A_2) \approx 0.02022$

$\therefore$ we choose Windy (A_3) for the second split

Finally,



Error Rate : For training set : $2/14$

For Test set : $2/5$

2.2 (1) From 2-1, when Humidity is High,

$I(Y, A_2) \approx 0.02022 < 0.04$, So we don't do further split

(2) when Humidity is Normal,

$I(Y, A_1) \approx 0.394 > 0.04$

① when Outlook is Rainy, $H(Y) = H(4,0) = 0$

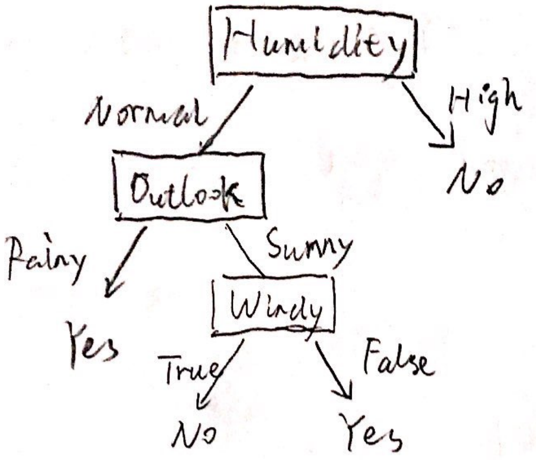
IG must be $0 < 0.04$, so stop splitting

② when Outlook is Sunny

$IG = H(\frac{2}{3}, \frac{1}{3}) - (\frac{2}{3} H(2,0) + \frac{1}{3} H(1,0)) \approx 0.9183 > 0.04$

So continue splitting on A_3

Finally,



Error rate: For training set : $\frac{3}{14}$

For test set : $\frac{1}{5}$

The test performance increases because the former tree seems overfitted. After constraining on IG < 0.04, the model becomes less complicated, which is a pruning process. It makes tree less overfit.

2.3. A_1: $Gini(Sunny) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$

$Gini(Rainy) = 1 - (\frac{6}{9})^2 - (\frac{3}{9})^2 = 0.44$

$Gini(A_1) = (\frac{5}{14}) \times 0.48 + (\frac{9}{14}) \times 0.44 = 0.457$

A_2: $Gini(High) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.489$

$Gini(Normal) = 1 - (\frac{6}{7})^2 - (\frac{1}{7})^2 = 0.244$

$Gini(A_2) = (\frac{7}{14}) \times 0.499 + (\frac{7}{14}) \times 0.244 = 0.367$

A_3: $Gini(False) = 1 - (\frac{6}{8})^2 - (\frac{2}{8})^2 = 0.375$

$Gini(True) = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2 = 0.5$

$Gini(A_3) = (\frac{8}{14}) \times 0.375 + (\frac{6}{14}) \times 0.5 = 0.428$

So we choose Humidity for the first split

When Humidity is High :

$$Gini(A_1) = \frac{2}{7} \times (1-(\frac{1}{2})^2-(\frac{1}{2})^2) + \frac{5}{7}(1-(\frac{2}{5})^2-(\frac{3}{5})^2) = 0.4857$$

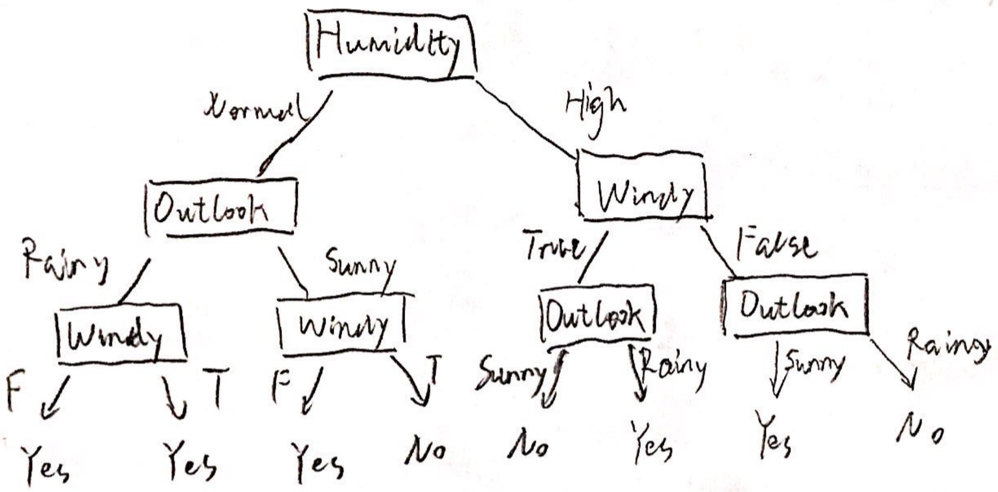
$$Gini(A_3) = \frac{3}{7}(1-(\frac{1}{3})^2-(\frac{2}{3})^2) + (1-(\frac{1}{2})^2-(\frac{1}{2})^2) \times \frac{4}{7} = 0.4762$$

We choose Windy for the second split.

When Humidity is Normal :

$$Gini(A_1) = Gini(A_3) = 0.1905$$

$\therefore$



Error Rate : training set $= 2/14$

test set $= 2/5$

Pruning : (1) When Humidity = Normal :

With Normal as a leaf node,

cost = 1 * Misclassification + 1 * leaf = 2

with further split,

cost = 1 * Misclassification + 2 * leaf = 3

with further split again,

cost = 0 * Misclassification + 4 * leaf = 4

(2) When Humidity = High :

with High as a leaf node ,

cost = 3 * Misclassification + 1 * leaf = 4

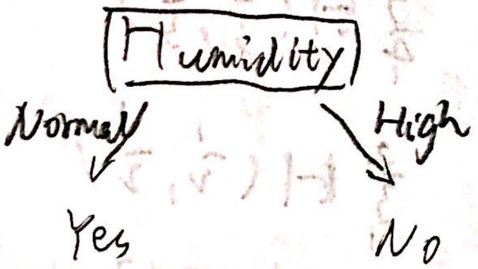
with further split :

cost = 3 * Misclassification + 2 * leaf = 5

with further split :

cost = 2 * Misclassification + 4 * leaf = 6

∴



Error rate = $\frac{4}{14} = \frac{2}{7}$